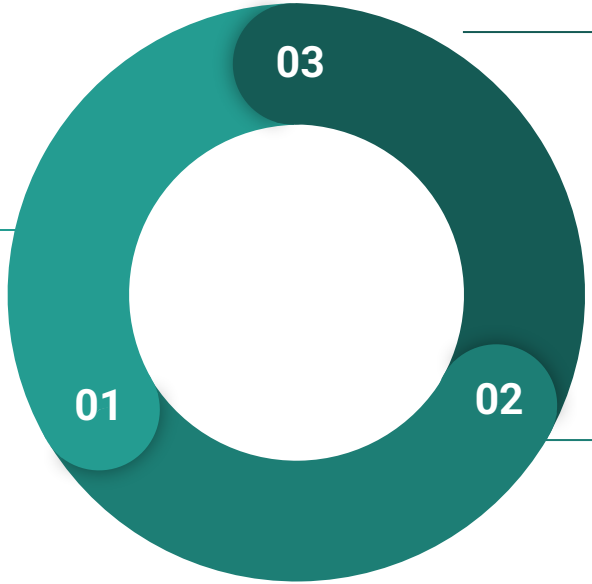
Automated Essay Scoring (AES) Using NLP

Date: 3/11/2025

Motivation

ESSAYS ARE ESSENTIAL

They enable students convey their ideas in a persuasive manner.



SCHOOLS RESORT TO LESS EFFECTIVE TOOLS

In the absence of good automated evaluation tools, schools have to compromise.

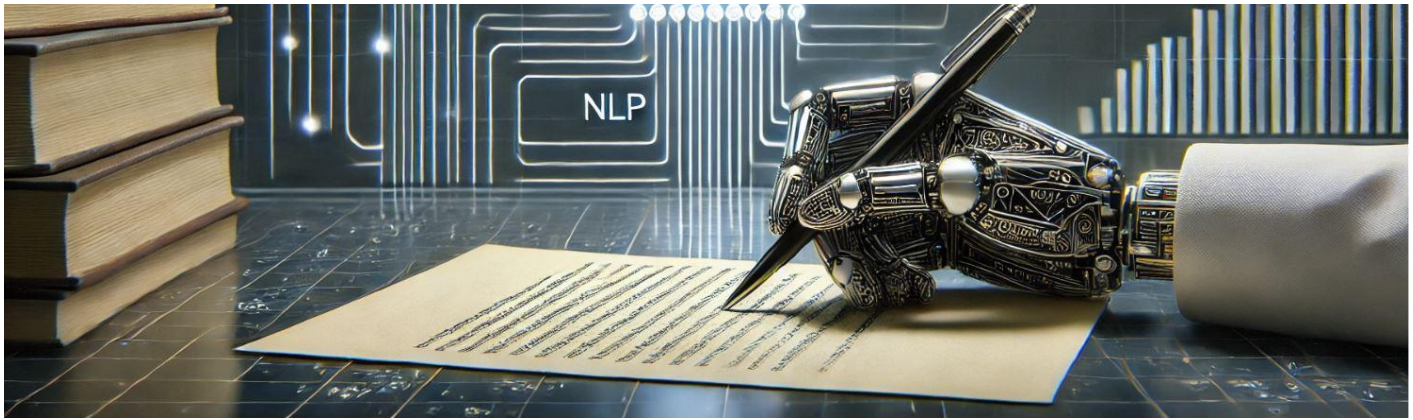
EVALUATION IS LABOR-INTENSIVE

The evaluation of essays is time-intensive and burdens already under-resourced state schools.

We need something ***fast, consistent and scalable.***

The Goal

Build a Machine Learning model that can automatically predict essay scores in a reliable manner.



Data

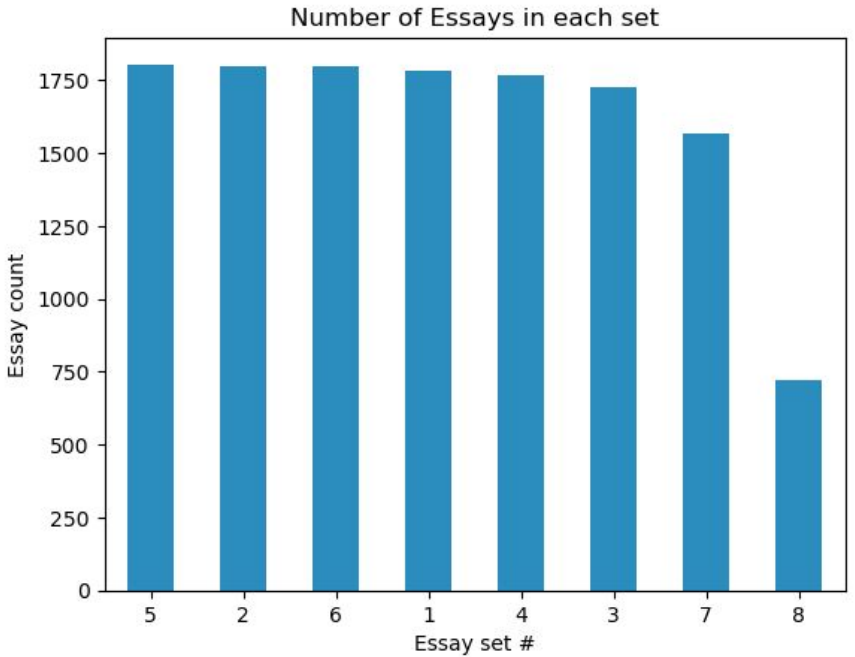
- The Hewlett Foundation: Automated Essay Scoring Dataset - Link
- 8 essay sets (~13000 essays).
- Grades 7 to 10.
- Varying prompts and rubric ranges.

Table 1 Prompts in ASAP Dataset

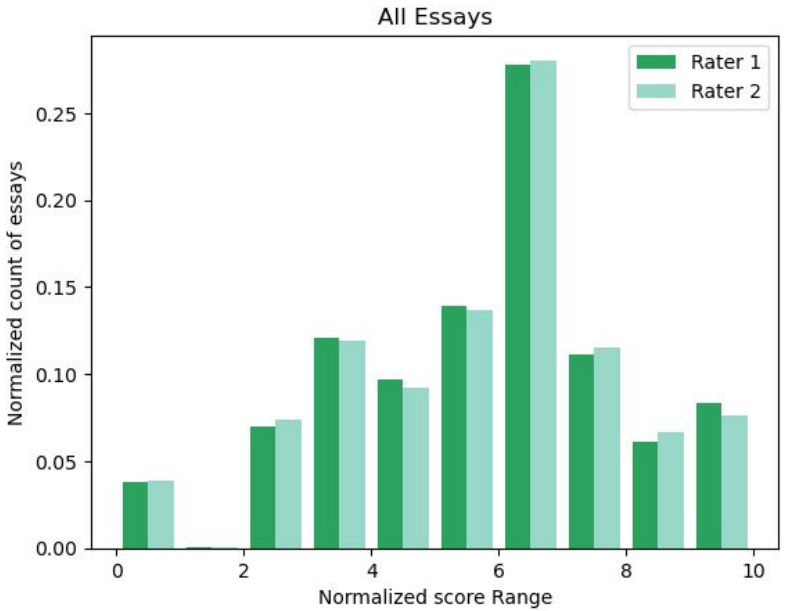
ASAP Dataset	Topics
Prompt 1	The effects computers have on people
Prompt 2	Censorship in the libraries
Prompt 3	Respond to an extract about how the features of a setting affected a cyclist
Prompt 4	Explain why an extract from *Winter Hibiscus* by Minfong Ho was concluded in the way the author did.
Prompt 5	Describe the mood created by the author in an extract from *Narciso Rodriguez* by Narciso Rodriguez
Prompt 6	The difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there
Prompt 7	Write a story about patience
Prompt 8	The benefits of laughter

Data

- The Hewlett Foundation: Automated Essay Scoring Dataset - Link
- 8 essay sets (~13000 essays).
- Grades 7 to 10.
- Varying prompts and rubric ranges.



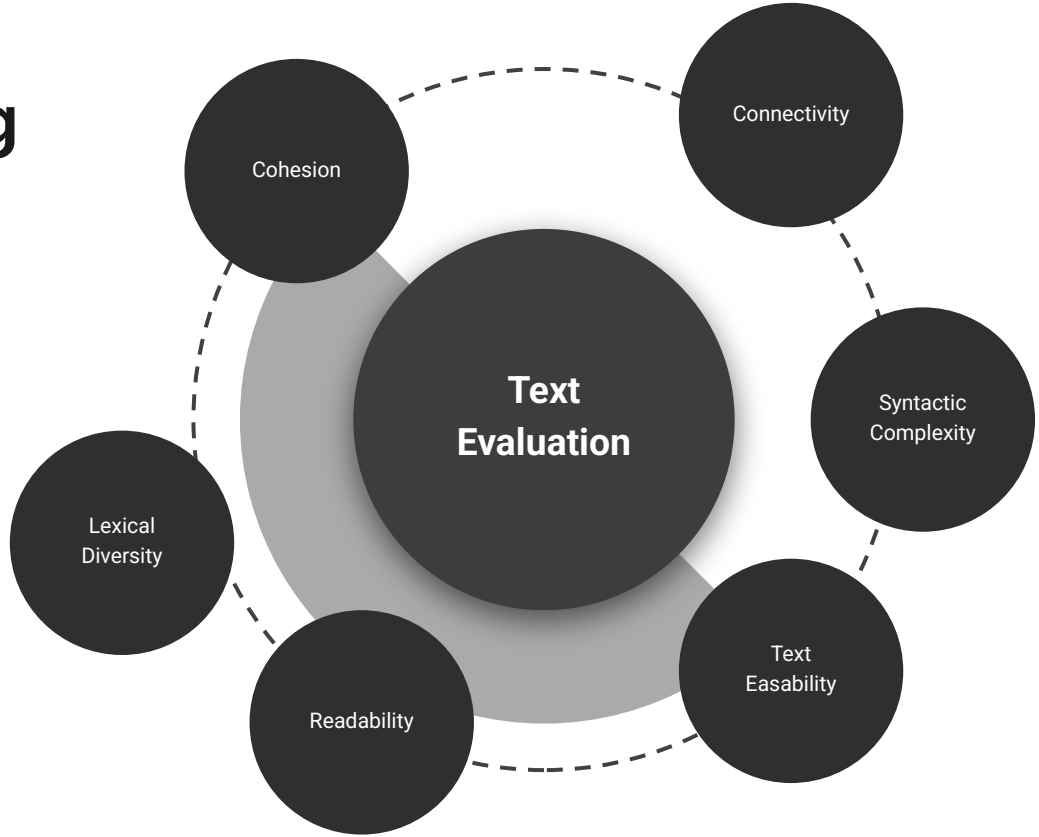
How well do human raters agree anyway?



Quadratic Weighted Kappa: 0.75±0.07

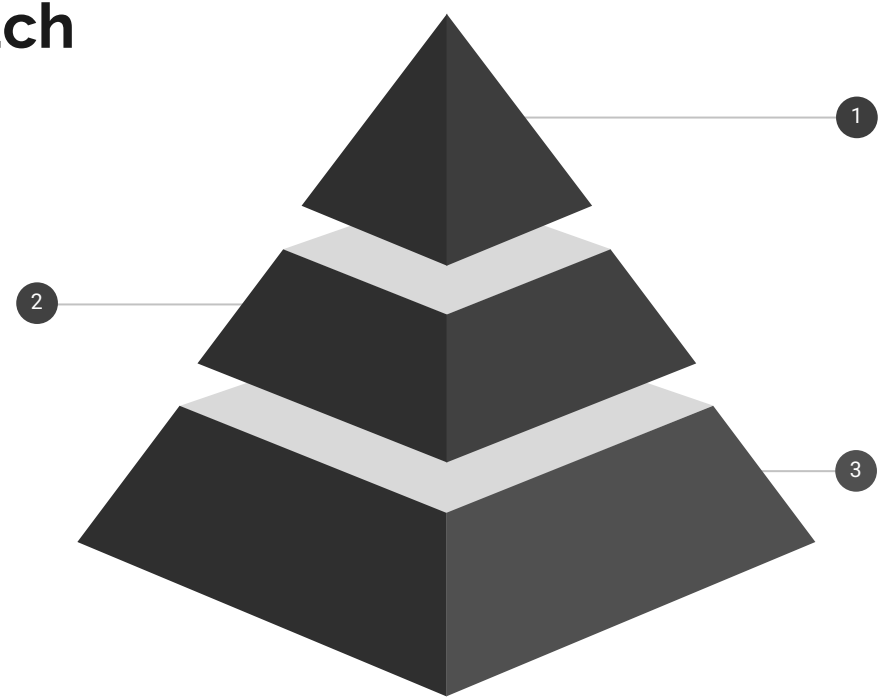
Feature-Engineering

Inspired by metrics developed in the Coh-Metrix tool designed to analyze discourse.



- MSE:~ 2 (units of normalized score)
- R2 Score: ~0.6
- Quadratic Weighted Kappa - Mean: 0.69

The Approach



Extract Features From Essays

Features were extracted and grades normalized on a consistent scale.

Train a Random Forest Regressor

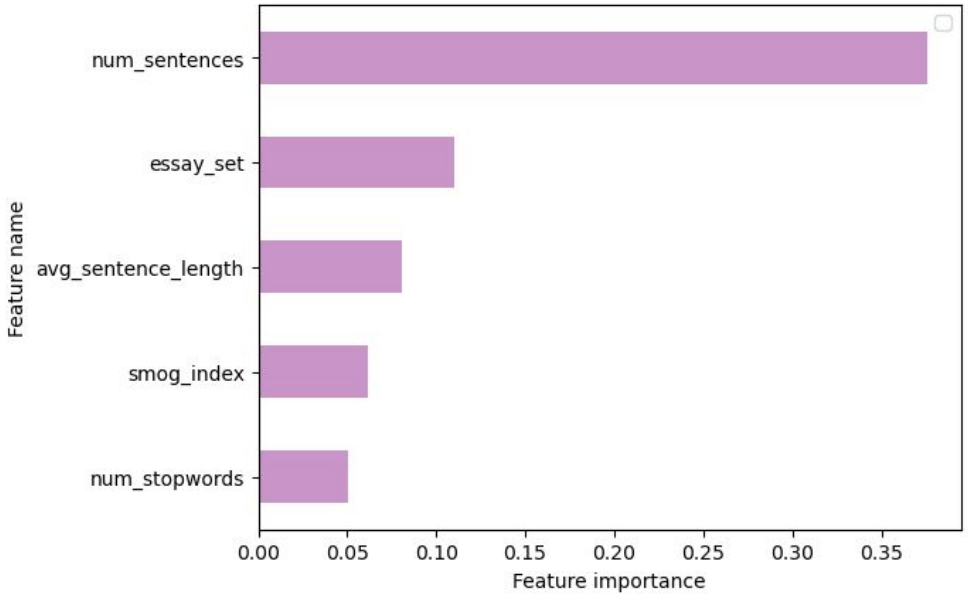
3-fold cross-validation and hyperparameter tuning.

Evaluate Performance

Mean Squared Error (MSE); R2 Score; Quadratic Weighted Kappa (QWK)

Challenges

- Moderate bias towards length-based features → inability to focus on content.
- Evaluation Metrics: MSE is not sufficient.



What's Next?

Enrich Features with Embeddings

Use embeddings from transformer-based models.

Set-Specific Models

Train one model per essay set to improve performance.

Hierarchical Modeling

Explore hierarchical models to mimic how humans grade better.