



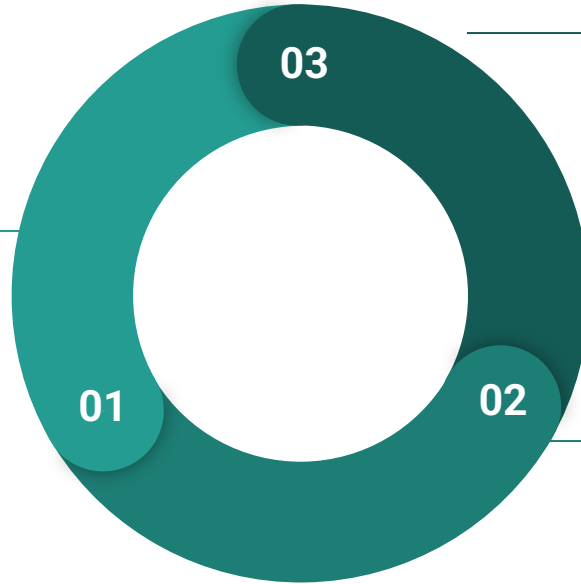
# Automated Essay Scoring Using NLP

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.

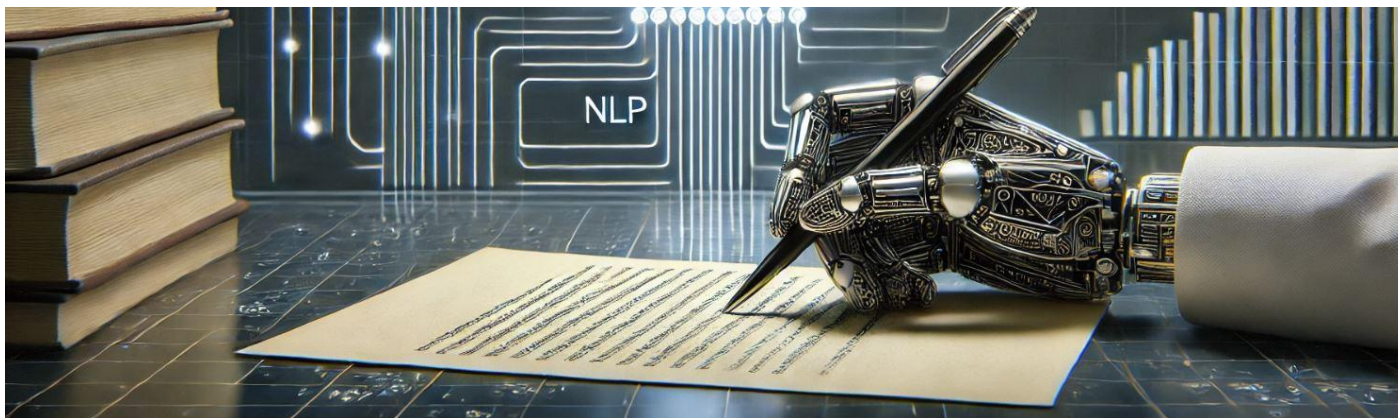
## EVALUATION IS LABOR-INTENSIVE

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

We need something *fast, consistent and scalable*.

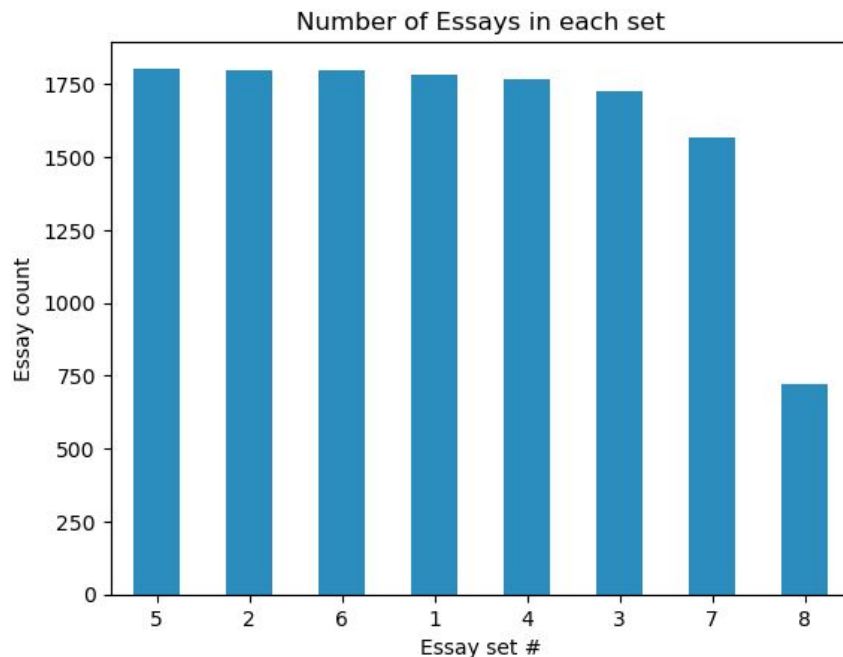
# The Goal

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

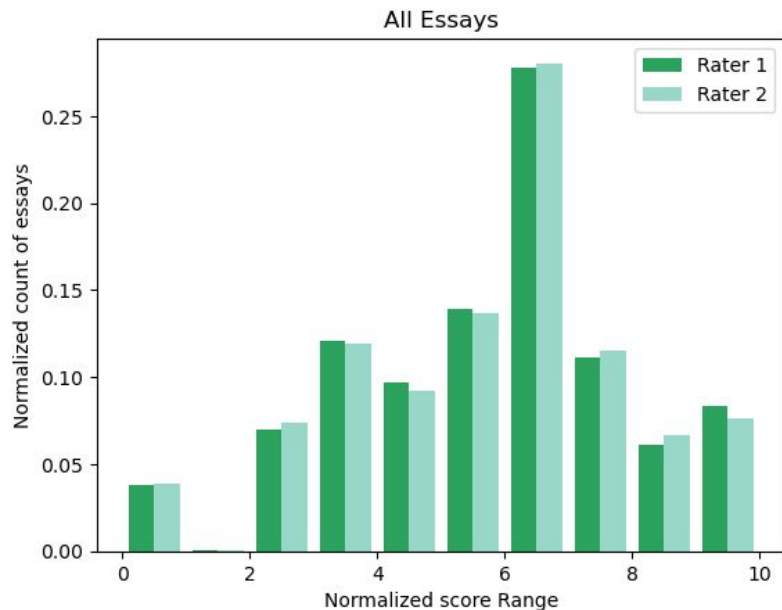


# Data

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



## How well do human raters agree anyway?

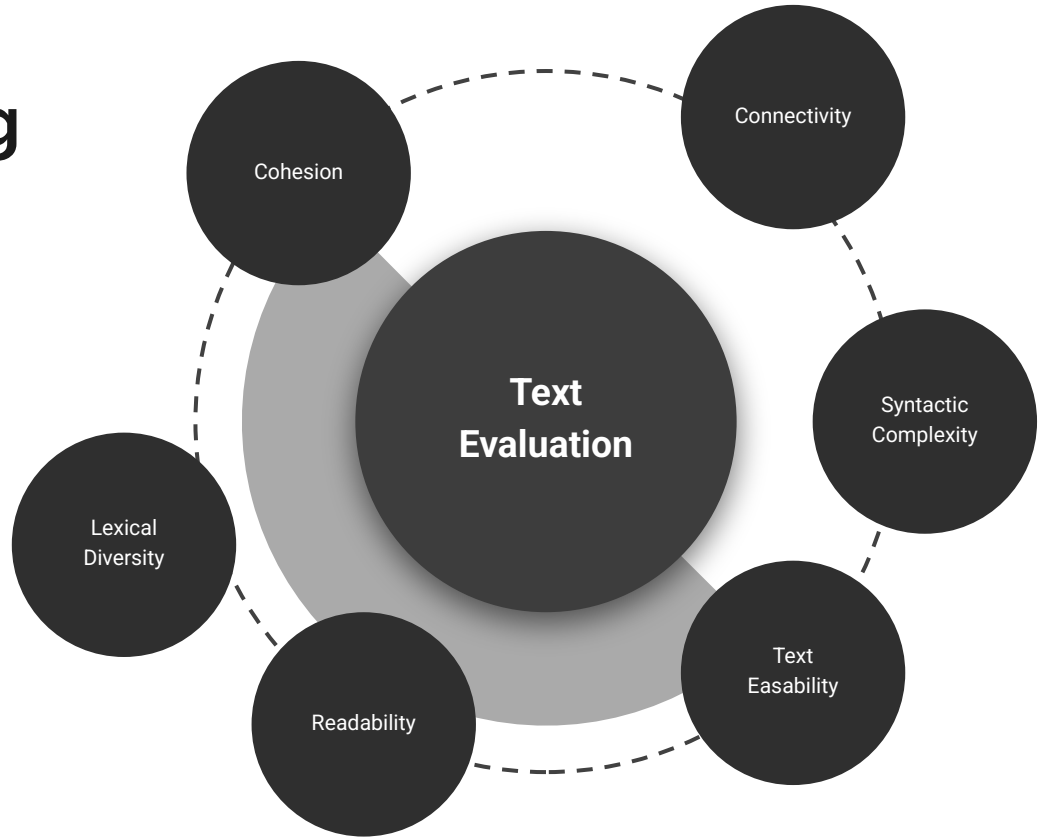


Quadratic Weighted Kappa:  $0.75 \pm 0.07$



# Feature-Engineering

Inspired by metrics developed in the Coh-Metrix tool.



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

# The Approach

## Extract Features From Essays

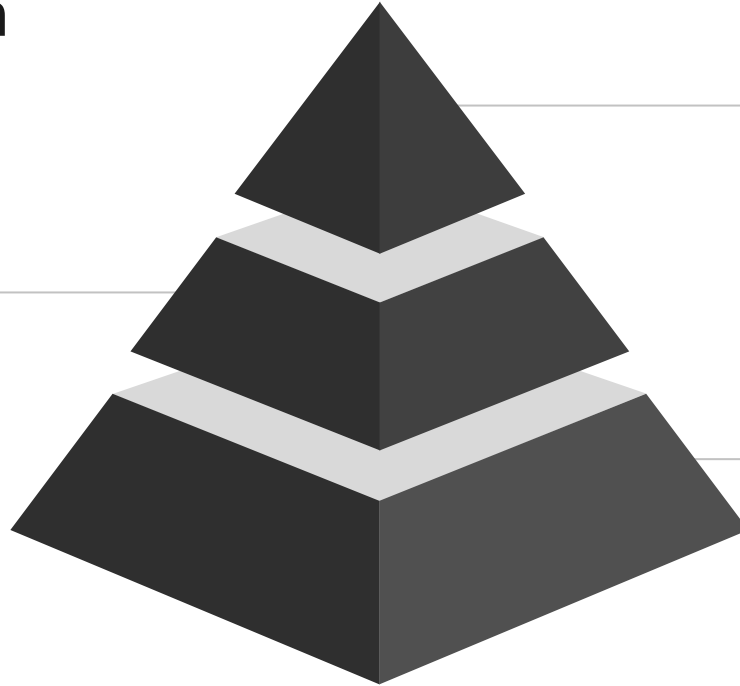
Features were extracted and grades normalized on a consistent scale.

## Evaluate Performance

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

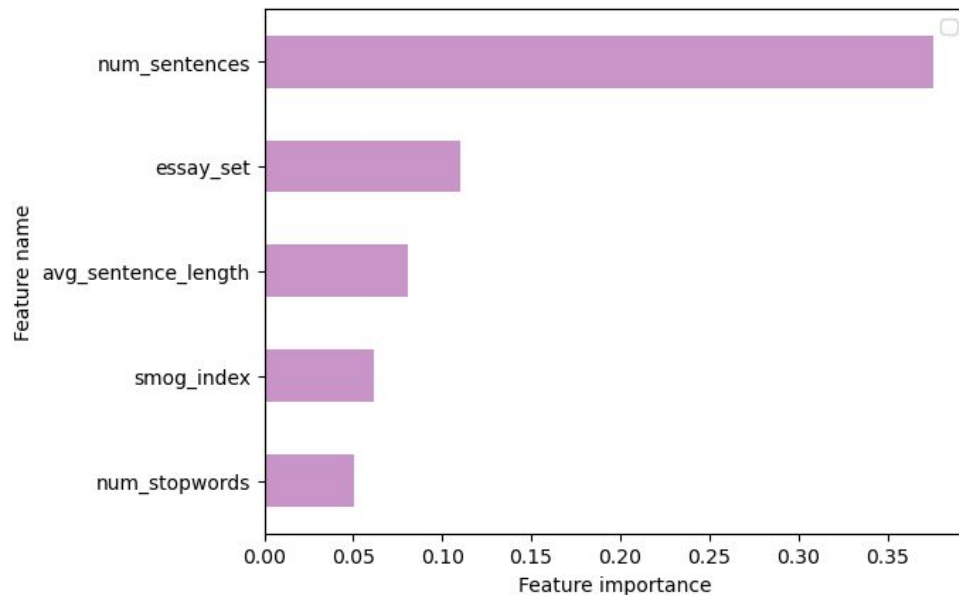
## Train a Random Forest Regressor

3-fold cross-validation and hyperparameter tuning.



# Challenges

- Moderate bias towards length-based features → Inability to focus on content.
- Evaluation Metrics: Accuracy is not enough; rater agreement matters.







# 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.