# Machine Learning for Automated Essay Scoring : Classification or Regression

Sameek Bhattacharya
*sameekbhattacharya@my.unt.edu*
Department of Computer Science and Engineering
University of North Texas

Tapajit Chandra Paul
*tapajitchandrapaul@my.unt.edu*
Department of Computer Science and Engineering
University of North Texa

*Abstract* – **Automated Essay Scoring (AES) is critical for addressing the inconsistency and resource demands of manual grading, yet many existing systems suffer from poor interpretability due to their reliance on large datasets and complex black-box models. This study proposes and validates a lightweight, interpretable AES approach suitable for resource-limited settings. We first determined an appropriate sample size using hypothesis testing, selecting 100 essays for analysis. The core methodology involved extracting common words from the highest-scoring essays to create a frequency-based feature set. This high-dimensional feature space was then reduced using Principal Component Analysis (PCA) to yield a compact set of meaningful components. This reduced dataset was used to train and compare simple machine learning models (both classification and regression) for score prediction. Our work justifies this approach through statistical rigor and feature analysis, demonstrating that simple models can effectively approximate human scoring and identifying the key linguistic patterns that contribute most significantly to essay quality. The findings underscore the potential of low-resource AES strategies for future development.**

## I. Introduction

Assessing student essays is a crucial part of education because written responses allow instructors to evaluate critical thinking, organization and communication skills of students. However, manual grading often requires significant time and effort. Especially when dealing with large numbers of essays. Human scoring can also be inconsistent as graders may differ in their interpretations or become fatigued over time. These problems make essay evaluation costly, slow, and difficult to standardize. As a result, Automated Essay Scoring (AES) has become an important research area which offers the possibility of faster, fairer, and more reliable assessment methods. Previous studies show that machine learning models can closely approximate human grades when trained on large datasets. However, many existing systems depend on either black box neural models or require thousands of essays which is not practical in small or resource limited settings.

In this project, we investigate whether a simple, interpretable, and low resource machine learning approach can produce meaningful essay scores. Instead of using thousands of essays, we rely on a carefully selected set of 100 essays. This number is determined using hypothesis testing to ensure statistical validity. The core idea is to identify which words and linguistic patterns appear most frequently in high-scoring essays. To do this, we first extract the common words from essays that received the highest score. Using these words, we then construct a dataset where each essay is represented by the frequency of those selected words. As this initial feature set may still contain redundant or weak predictors, we apply Principal Component Analysis (PCA) to reduce the number of variables and retain only the most important or informative components. This reduced dataset becomes the input to a machine learning model, which we use to predict the final essay score.

This study is motivated by two main issues. First, traditional AES systems often rely on proprietary or hard-to-interpret models which makes it unclear why certain scores are assigned. Second, many systems assume the availability of very large, labeled datasets, which is unrealistic for many educational institutions. Our approach attempts to overcome these limitations by only focusing on a feature-based model that is simple to explain and does not rely on massive training sets. In addition, our hypothesis states that essays containing fewer stop words relative to total words tend to receive higher scores. It is based on the assumption that strong writing uses more meaningful words and fewer filler words. Our experimental design directly tests this hypothesis by examining word-frequency patterns, extracting principal components, and evaluating the performance of different machine learning models.

This paper is organized as follows. Section II presents a detailed review of the literature, which covers five relevant studies on Automated Essay Scoring and discusses their strengths and limitations. Section III describes our proposal, including the problem statement, research questions, and hypothesis that guide this work. Section IV presents the research and experimental methodology, such as the dataset selection process, preprocessing steps, statistical analysis, and machine learning models used. Section V presents the results obtained from the experiments. Section VI provides conclusions and outlines directions for future work. The paper ends with an acknowledgement section recognizing the contributions that supported this research.

## II. Related Study

This section focuses on previous work in Automated Essay Scoring (AES). Most of these studies try to replace or provide support to human graders using either machine learning or deep learning. For each paper, we first briefly describe what they did, what worked well or what the strengths of the paper were, and what the limitations were. Focus is particularly on features, models, data size and transparency. We then explain how our approach is different from them and which gap it tries to fill.

### A. *A Neural Approach to Automated Essay Scoring [1]*

In this paper, Taghipour and Ng proposed one of the earliest neural network approaches for Automated Essay Scoring. Their model used word embeddings with a combination of convolutional and recurrent layers which instead of relying on hand-crafted linguistic features learns representations of essays directly from raw text. They evaluated their method on the widely used ASAP dataset and showed that neural architectures outperform traditional regression and feature-engineering approaches. The strength of this work is that it demonstrates how deep models can automatically extract meaningful semantic and syntactic information from essays. However, the main limitation is that neural models require large amounts of training data, substantial computational resources, and they offer limited interpretability which makes it difficult to deploy in low-resource educational environments.

### B. *Automated Essay Scoring Using Siamese BiLSTM Networks [2]*

In this paper, Liang and colleagues introduced a Siamese Bidirectional LSTM architecture for essay scoring. Their model compares student essays with reference high-quality essays. For that they encode both texts and measure their similarity. This approach allows the scoring system to account for content relevance and also writing quality. The model achieved good results on benchmark datasets and showed that comparison-based architectures can improve content sensitivity in AES systems. The strength of the study lies in its ability to incorporate semantic similarity into scoring, which many earlier neural models did not explicitly address. A limitation is that providing high quality reference essays for each prompt can be difficult.

### C. *Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring [3]*

Dong, Zhang, and Yang proposed an attention based recurrent convolutional neural network for scoring essays automatically. They used hand-crafted features and their model learns a representation of the essay directly from the raw text using a combination of recurrent and convolutional layers. They also use an attention mechanism which assigns higher weights to more important words and sentences. Their model was able to outperform earlier baselines on the ASAP dataset. They showed that attention helps the model focus on the most relevant parts of the essay. The strength of this work is that it reduces manual feature engineering, also captures both local and global information in the essay. However, the model is complex, requires both substantial training data and computation. It is also harder to interpret why a particular score was assigned compared to simpler, feature-based models.

### D. *Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input [4]*

Farag et al. showed that many strong AES systems can be fooled by essays that are grammatically correct however incoherent. For example, when random sentences are stitched together in an essay. They proposed a neural model that explicitly learns local coherence between sentences. Then they combined this coherence model with a standard AES model. Their joint system improved essay scoring performance and the ability to detect adversarial and incoherent essays. The strength of this paper includes highlighting a serious weakness of many AES systems. It also offers a concrete way to handle it by modeling sentence connectedness. A limitation is that the approach is again based on deep neural networks that are data hungry and less transparent. It does not explore simpler feature-based methods which may be easier to deploy and also explain in small or resource limited settings.

### E. *Structural Explanation of Automated Essay Scoring [5]*

In this paper, Doewes and Pechenizkiy discussed the problem of transparency in automated essay scoring. They said that many AES models, particularly deep neural models, behave like "black boxes" which means they do not explain why a particular score was given. Their work investigated a framework which provides structural explanations of essay scores and tries to connect model decisions to understandable features or structures in the essay. The main strength of this study is its focus on interpretability and fairness. It is important in educational settings where students and teachers need to trust the system. However, the paper is more about explanation frameworks than about proposing a simple and end-to-end scoring model. It does not directly address how to design lightweight models that still perform well with limited data.

### F. *Summary and Research Gap*

It is evident from the prior discussion that past research has shown that:

1. Traditional machine learning models with hand-crafted features can reach high agreement with human graders on large essay datasets.

2. Deep neural models with attention, coherence modeling, or transformer-based representations can improve accuracy and robustness. However, they are complex, data intensive and harder to interpret.

3. There is also growing concern about the transparency of AES systems. The need for explanations that teachers and students can understand.

However, most of these works use very large essay collections (thousands of essays). They also rely on either many hand-crafted features or heavy neural architectures. They rarely study how far a simpler model can go when we: (1) carefully choose a smaller sample size using hypothesis testing, (2) focus on a small set of high-impact words that distinguish high-scoring essays, and (3) apply dimensionality reduction such as PCA to keep only the most informative word-count features before training a machine learning model. Our project is designed to explore this specific gap: we take 100 essays selected through hypothesis testing, extract common words from the highest scoring essays, reduce the feature space using PCA, and then train a ML model on this compact feature set. By doing this, we aim to see whether a lightweight, easily interpretable and word-count-based system can still give useful scores. Also aim to better understand the trade-off between model simplicity, data size, and scoring performance.

## III. Proposal

### A. *Problem Statement*

The primary objective of this research is to test how implementation of machine learning models can improve the scoring of essays in terms of automating them and saving time and resources. Traditional Automated Essay Scoring (AES) systems often rely on proprietary black-box models or require vast, professionally annotated datasets. This research addresses the challenge of building an interpretable and effective AES system using limited, domain-specific data (141 essays) and low-resource feature engineering. The research looks at which are the most contributing features to the scoring using statistics and also which type of machine learning, classification or regression, is most suitable for this task.

### B. Research Questions

In this is experimental research, we have tried to address a few research questions. First, what should be a good distribution of dataset among the different classes. Second, what should be a suitable dataset size, assuming the standard deviation and length of the population. Third, which are the most contribution features in the processed dataset. Fourth, which machine learning task, whether classification or regression, should be the most suitable for this project to form a good Automated Essay Scoring algorithm.

### C. Hypothesis

In our hypothesis, we have assumed the expected outcomes for each of our research questions and proceeded with the experimentations. Firstly, we have considered that a normally distributed dataset across all the classes would be a correct choice. Meaning, that the scores of 12 and 2, being highest and lowest respectively, should have the lowest number of samples and classes like 4, 6, 8 and 10 should have a greater number of samples. Secondly, we have performed a statistical calculation assuming the standard deviation and length and found that dataset size of above 113 is suitable for our task. Thirdly, we have performed Principal Component Analysis to find the top 50 contributing features from the processed dataset. And fourthly, after trying out machine learning models, treating our task as classification and regression, we expect that regression would be a more suitable choice.

## IV. Research And Experiments

### A. Dataset Description

The originally provided corpus consists of essays collected for 8 different prompts and manually scored by two different human graders and a third in specific cases and the sum of the two graders. In our research, we have only selected the essays related to prompt 1. We have manually selected 141 essays for the dataset in a normal distribution among the classes. Figure 1 shows how our selected essays are distributed among the score classes.

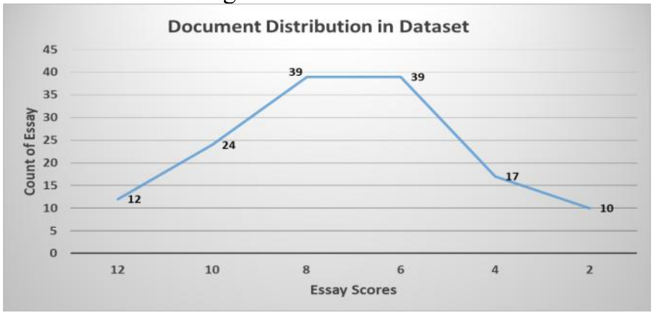

Figure 1. Distribution of the essay dataset according to scores.

### B. Data Preprocessing

The initial feature set required comprehensive preprocessing to prepare the raw essay text for modelling. This involved several crucial steps: punctuation removal, selective case handling (non-special words converted to lowercase, i.e., words not starting with “@”), custom token grouping (e.g., “CAPS1” to “CAPS”), lemmatization and lastly stopword removal. The cleaned corpus was then converted into features by creating count columns for all unique processed words found in the dataset, resulting in exactly 1,221 initial features. To address this high dimensionality and focus on predictive power, Univariate Feature Selection (using the F-test for regression) was applied, selecting the top 50 ranked word-count features based on their strong correlation with the target variable “domain1_score”. These 50 features, target variable, along with word counts, unique word count and stopword counts, formed the final reduced dataset for model training. (Having 53 features and 1 target).

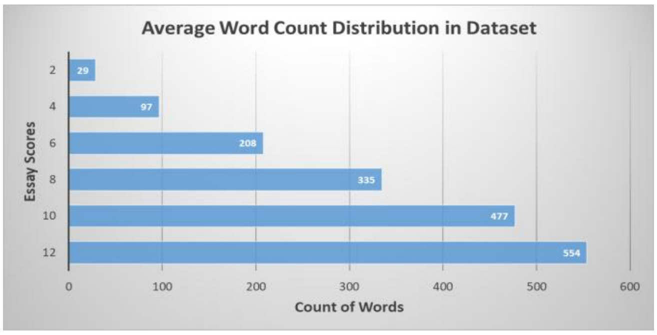

Figure 2. Average total word count as per score classes

As shown in Figure 2, Figure 3 and Figure 4, we have performed descriptive statistical analysis by calculating the mean count of words, stopwords, and unique words for the essays, grouped according to their respective score classes. This establishes the average lexical characteristics for each scoring level.

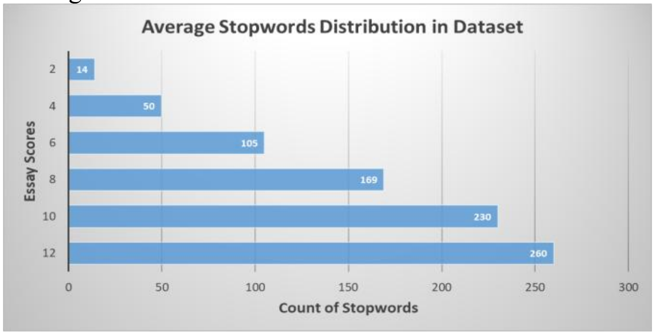

Figure 3. Average stopword count as per score classes

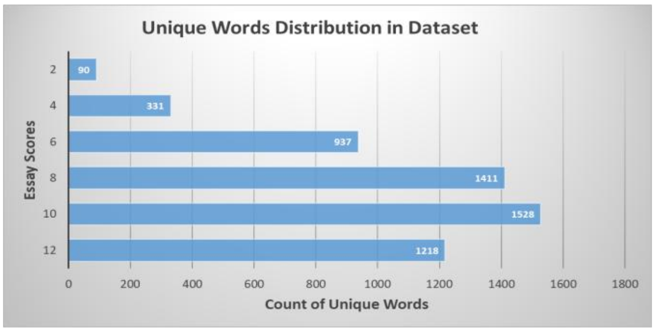

Figure 4. Average unique word count as per score classes

### C. Statistical Approach

We have performed basic mathematical operations like averaging, to find the average count of words, average count of stopwords, average count of unique words in the essays grouped by their score classes. We have plotted in horizontal bar graphs to get a better understanding through the visualization.

In a separate test, we have tried to find what would be a suitable size of the dataset. For that, we have started the test by using the score classes ($x_i$) and their average word count per class ($f_i$) to get the mean using the following formula :

$$\boxed{\mu = \frac{\sum(x_i \cdot f_i)}{\sum f_i}}$$

We have obtained a mean of 7.219858. Then we have computed and found the variance, using the following the formula :

$$\boxed{\sigma^2 = \frac{\sum f_i(x_i - \mu)^2}{\sum f_i}}$$

We have obtained a variance of 7.022584. Then have found the standard deviation, using this following formula :

$$\boxed{\sigma = \sqrt{\sigma^2}}$$

Getting a value of 2.65002, using this value as our standard deviation, assuming an interval length of 1 and t-value as 2, we have computed the suitable size of our dataset. This is in accordance with a 95% confidence two-sided t-interval. The formula we used is :

$$\boxed{n = \left(\frac{2 \cdot t \cdot s}{L}\right)^2}$$

We have obtained a value of n = 112.36. So, the minimum number of samples in the dataset should be 113. In our case, we have taken 141 samples in the dataset.

### D. Machine Learning Approach

In our research work, we have decided to take 2 models each of regression and classification type. We have considered Support Vector Regressor (SVR) and Gradient Boosting Regressor under the regression task and Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) under the classification task. We have considered the value of k=5 for our KNN classifier. In the preprocessing step, we have also tried to perform tf-idf of the features and then perform Principal Component Analysis (PCA). Here tf is term frequency which is frequency of each word in each of the documents. And idf is inverse document frequency which is a a measure of how important a word is to a document within a collection, by determining its rarity across that collection. And then we have multiplied the value of tf with idf to get the tf-idf values. But in our final dataset, we have considered the PCA over the raw word features and considered the top 50 contributing words. For our measure of model performance, we have considered accuracy, F1 score and quadratic weighted kappa for the classification models and root mean squared error, mean absolute error and quadratic weighted kappa for the regression models.

## V. Results

### A. Statistical Findings

Through statistical approaches, we have found that the number of samples in the dataset should be above 113. So, we have considered 141 essay samples. The way we have chosen the dataset is to maintain a normal distribution across all the classes. 12 essays for the score class of 12, 24 essays for the score class of 10, 39 essay samples for the score class of 8 and 6, 17 essays for the score class of 4 and 10 essays for the score class of 2.

### B. Machine Learing Findings

Through our machine learning approach, we were successfully able to form an algorithm capable of grading an essay. We have achieved a highest quadratic weighted kappa score of 0.9595 for the Gradient Boosting Regressor under the regression task. We have seen that, the regression models, in general, perform better than the classification models. Here, quadratic weighted kappa (QWK) is a statistical metric used to measure the agreement between two raters on an ordinal scale, where errors farther away from the correct classification are penalized more heavily. The metric's value ranges from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating agreement no better than chance.

The tables 1 and 2 below shows the performance of the models on their specific metrics.

Table 1. Performance of Regression models

| Model | Root Mean Squared Error (RMSE) | Mean Absolute Error (MAE) |
|---|---|---|
| Support Vector Regressor | 1.1008 | 0.889 |
| Gradient Boosting Regressor | 0.9264 | 0.6134 |

Table 2. Performance of Classification models

| Model | Accuracy | F1 Score |
|---|---|---|
| Support Vector Classifier | 0.5517 | 0.5123 |
| K-Nearest Neighbors Classifier | 0.2414 | 0.2007 |

The table 3 below shows the performance of all the regression as well as classification models using the common metric of quadratic weighted kappa (QWK).

Table 3. QWK score across all models

| Model | Quadratic Weighted Kappa (QWK) |
|---|---|
| Support Vector Regressor | 0.9327 |
| Gradient Boosting Regressor | 0.9595 |
| Support Vector Classifier | 0.8537 |
| K-Nearest Neighbors Classifier | 0.6266 |

The graph in Figure 5 shows the QWK score of all the models.

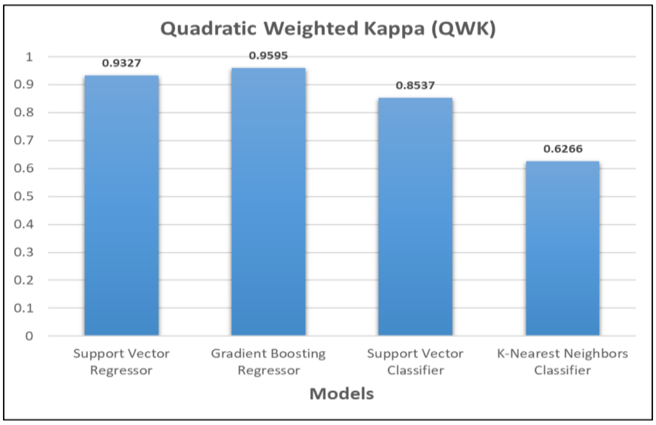

Figure 5. QWK score of all the models

## VI. Conclusions And Future Work

This empirical study successfully addressed the challenge of developing an effective and interpretable Automated Essay Scoring (AES) system using a limited, curated dataset. We established the statistical reliability of our corpus by confirming the sample size of 141 essays was sufficient for reliable inference, exceeding the minimum requirement calculated via power analysis. The core success lay in the rigorous feature engineering and reduction strategy. Raw essay text underwent meticulous cleaning—including selective casing, punctuation removal, lemmatization, and the critical step of custom token grouping (e.g., CAPS1 to CAPS) to normalize specialized features. This process generated an initial matrix of 1,218 features, which was then intelligently distilled using Univariate Feature Selection (F-test for regression). By retaining only, the top 50 word-count features highly correlated with the domain1_score, and augmenting these with the essential lexical counts (word count, unique word count, and stopword count), we created a highly optimized dataset of 53 features, successfully mitigating the high-dimensionality risks associated with small training sets.

The Machine Learning phase definitively validated this approach, demonstrating that simple models operating on statistically optimized features can achieve human-level agreement. The regression framework proved superior for this ordinal scoring task, with the Gradient Boosting Regressor achieving an outstanding Quadratic Weighted Kappa (QWK) score of 0.9595. This result is highly significant, indicating near-perfect agreement with human graders and establishing a strong benchmark for low-resource AES. The success of both the Gradient Boosting and Support Vector Regressor (SVR) models, compared to the slightly lower performance of classification models, underscores the importance of treating essay scores as an ordinal scale. Our findings support the thesis that high predictive performance in AES can be achieved without relying on resource-intensive methods or massive datasets, provided the feature selection is guided by strong statistical principles.

Future research should focus on further enhancing model interpretability and robustness. We recommend exploring Lasso Regression to enforce sparsity and confirm the predictive stability of the top 50 features. Additionally, incorporating more sophisticated feature engineering derived from linguistic theories, such as features related to essay coherence, structural complexity (syntax), and grammatical correctness (part-of-speech counts), could potentially push the QWK metric even closer to a perfect 1.0. Scaling this optimized methodology to other essay prompts in the original corpus would be essential to confirm the external validity and generalizability of the developed AES system.

## Acknowledgement

We, Sameek Bhattacharya and Tapajit Chandra Paul, would like to thank everyone who made this research project possible. We are deeply grateful to Dr. Jesus U. Quevedo, our professor for the CSCE5310 – Methods in Empirical Analysis course at the University of North Texas. The guidance of the professor and the constructive feedback from the fellow classmates and professor, helped in shaping the research and successful completion of this experimental Automated Essay Scoring task.

## References

[1] K. Taghipour and H. T. Ng, “A neural approach to automated essay scoring,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 1882–1891.

[2] G. Liang, B.-W. On, D. Jeong, H.-C. Kim, and G. S. Choi, “Automated essay scoring: A Siamese bidirectional LSTM neural network architecture,” Symmetry, vol. 10, no. 12, p. 682, 2018.

[3] F. Dong, Y. Zhang, and H. Yang, “Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring,” in Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL), 2017, pp. 153–162.

[4] Y. Farag, H. Yannakoudakis, and T. Briscoe, “Neural automated essay scoring and coherence modeling for adversarially crafted input,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), 2018, pp. 2632–2642.

[5] T. Doewes and M. Pechenizkiy, “Structural Explanation of Automated Essay Scoring,” in Proceedings of the 13th International Conference on Educational Data Mining (EDM), 2020, pp. 339–34