Ural Federal University
named after the first President of Russia B.N.Yeltsin

Ural Federal University
Institute of radioelectronics and information technologies

# Final Project Report

Automated Essay Scoring Using Deep Learning
with LSTM and Word2Vec

Supervisor: Jarno Matarmaa

Student: SABAR Lebate

Group: RIM-140930

# Content



# List of Figures



# List of Tables

# 1. Introduction

Essay-based assessment remains one of the most rigorous and comprehensive means of evaluating a student's cognitive and linguistic abilities. Unlike multiple-choice tests or structured exercises, essay questions demand critical thinking, coherent argumentation, and mastery of language all of which pose considerable challenges not only to learners but also to educators responsible for grading them. In large-scale educational contexts, the manual evaluation of essays is both time-consuming and susceptible to subjectivity and inconsistency, highlighting the need for scalable and objective alternatives.

Within this context, **Automated Essay Scoring (AES)** has emerged as a promising area of research under the broader umbrella of **Natural Language Processing (NLP)**. The principal aim of AES systems is to mimic human judgment in scoring essays, utilizing machine learning and deep learning techniques to capture the structural, syntactic, and semantic characteristics of written text. As the sophistication of language models continues to advance, the potential of AES to serve as a reliable educational tool becomes increasingly feasible.

This project contributes to the ongoing discourse in AES by developing a deep learning-based model capable of evaluating student essays. The work is grounded in the **ASAP Automated Essay Scoring dataset**, a publicly available and widely accepted benchmark that includes thousands of student-written essays across diverse prompts, each annotated by human raters. This dataset provides an ideal foundation for training and validating predictive models.

The methodology adopted in this study consists of several stages: textual preprocessing, vectorization using **Word2Vec** embeddings, and the implementation of a **Long Short-Term Memory (LSTM)**-based neural network. The performance of the model is assessed using established evaluation metrics such as **Mean Squared Error (MSE)** and **Cohen's Kappa Score**, which measure both accuracy and inter-rater reliability.

Through this research, we seek not only to build an effective scoring model but also to explore the broader implications of applying artificial intelligence in educational assessment. The project ultimately aims to offer a prototype that can be refined and scaled for real-world classroom use, thereby contributing to a more efficient, consistent, and data-driven approach to evaluating student writing.

## 1.1 Objectives

The primary objective of this research is to develop an effective automated essay scoring (AES) model using deep learning and natural language processing techniques, with the goal of emulating human grading in a consistent and scalable manner. The project aims to construct a reliable system that can be integrated into educational contexts to support large-scale assessment. Specific goals include exploring and preprocessing the ASAP dataset to ensure its readiness for modeling, including handling missing data and linguistic normalization. Additionally, the study involves constructing semantic word representations using the Word2Vec algorithm to transform textual input into meaningful numerical vectors. The model architecture is built using a Long Short-Term Memory (LSTM)-based neural network capable of capturing contextual and sequential dependencies within student essays. Performance is evaluated using quantitative metrics such as Mean Squared Error (MSE) and Cohen's Kappa Score to assess both prediction accuracy and alignment with human raters. Finally, the study seeks to validate the model through visual analysis of predictions and to reflect critically on the pedagogical and ethical dimensions of using artificial intelligence in educational assessment, including concerns around fairness, transparency, and practical deployment.

## 1.2 Literature review

Automated Essay Scoring (AES) has been an active area of research for over two decades, evolving significantly with the advancement of natural language processing (NLP) and machine learning techniques. The goal of AES systems is to evaluate written essays in a way that mirrors human judgment, ensuring fairness, consistency, and efficiency in educational settings.

Early AES systems, such as **Project Essay Grade (PEG)** developed by Ellis Page in the 1960s, relied heavily on surface-level features like word count, sentence length, and syntactic variety. Although pioneering, these systems lacked the ability to understand semantic content or context.

Subsequent systems like **e-rater** by Educational Testing Service (ETS) and **Intelli Metric** incorporated more sophisticated features, including grammar checking, coherence modeling, and lexical diversity. These rule-based and statistical models significantly improved the reliability of automated scoring but were still limited in capturing deep semantic meaning.

With the rise of machine learning, especially **supervised learning algorithms**, AES systems began to leverage training data to learn patterns between essay features and human-assigned scores. Techniques such as **TF-IDF**, **n-grams**, and **bag-of-words** became standard in feature extraction. Models like **support vector machines (SVMs)** and **random forest regressors** were commonly used for score prediction.

Recent advancements in **deep learning** have further revolutionized AES. **Recurrent Neural Networks (RNNs)**, **LSTMs**, and more recently, **transformer-based models** like **BERT (Bidirectional Encoder Representations from Transformers)** have shown state-of-the-art performance in understanding essay context, coherence, and content relevance. These models can capture complex linguistic patterns and semantic relationships, making them highly effective for grading tasks.

However, the deployment of AES systems still poses challenges, including **bias in training data**, **lack of transparency** in scoring decisions, and **generalization across different prompts or languages**. Recent research emphasizes the importance of explainability, fairness, and human oversight in AES deployment to ensure ethical and educational validity.

This project builds upon these foundations by using traditional machine learning methods (TF-IDF + Random Forest) to demonstrate how interpretable models can still provide valuable and scalable solutions in educational assessment, especially when working with structured and well-labeled datasets like the one provided by the Learning Agency Lab.

# 2. Materials and Methods

## 2.1 Dataset description

The dataset employed in this study is derived from the **ASAP (Automated Student Assessment Prize)** competition, hosted by the Hewlett Foundation in collaboration with Kaggle. It constitutes one of the most comprehensive publicly available resources for research in automated essay scoring. The dataset consists of **12,976 essays** written by students in response to **eight distinct prompts**, each designed to assess a specific genre of writing—ranging from argumentative and narrative to persuasive and informative styles. The essays vary significantly in terms of length, complexity, and linguistic style, thereby providing a rich and diverse corpus for model training and evaluation.

Each essay in the dataset is accompanied by several metadata fields, including the **essay ID**, **prompt ID**, **essay text**, and **domain-specific scores** assigned by two independent human raters. These scores, typically integers on a fixed scale (which varies depending on the prompt), represent the ground truth labels for supervised learning. In many cases, discrepancies between raters are resolved through averaging or adjudicated final scores to ensure fairness and consistency.

The dataset is formatted as a tab-separated values file (training_set_rel3.tsv), and includes additional fields such as rater1_domain1 and rater2_domain1, which are used to compute the final score for each essay. Prior to modeling, extensive preprocessing is required to clean the text data, normalize its format, and tokenize it for embedding and sequence analysis. The diversity and real-world nature of the essays make this dataset an ideal benchmark for evaluating the robustness and generalizability of automated scoring systems.

Table 1:Dataset Details

| **Attribute** | **Description** |
|---|---|
| Dataset Name | ASAP (Automated Student Assessment Prize) |
| Source | Hewlett Foundation / Kaggle |
| Number of Essays | 12,976 |
| Number of Prompts | 8 |
| Essay Length | Varies by prompt (from ~150 to 650+ words) |
| Essay Genres | Narrative, Persuasive, Expository, Argumentative |
| Format | TSV (Tab-Separated Values) file: training_set_rel3.tsv |
| Fields Included | essay_id, prompt_id, essay text, rater1_domain1, rater2_domain1, final score |
| Score Range | Varies by prompt (e.g., 0–3, 0–6, 0–60) |
| Scoring Method | Human ratings (2 raters), with final score as average or adjudicated resolution |
| Language | English |
| Target Variable | Final Essay Score |

Table 2: Descriptive statistics

| Feature | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| essay_id | 12976 | 10295.4 | 6309.07 | 1 | 4438.75 | 10044.5 | 15681.25 | 21633 |
| essay_set | 12976 | 4.18 | 2.14 | 1 | 2 | 4 | 6 | 8 |
| rater1_domain1 | 12976 | 4.13 | 4.21 | 0 | 2 | 3 | 4 | 30 |
| rater2_domain1 | 12976 | 4.14 | 4.26 | 0 | 2 | 3 | 4 | 30 |
| rater3_domain1 | 128 | 37.83 | 5.24 | 20 | 36 | 40 | 40 | 50 |
| domain1_score | 12976 | 6.8 | 8.97 | 0 | 2 | 3 | 8 | 60 |
| rater1_domain2 | 1800 | 3.33 | 0.73 | 1 | 3 | 3 | 4 | 4 |
| rater2_domain2 | 1800 | 3.33 | 0.73 | 1 | 3 | 3 | 4 | 4 |
| domain2_score | 1800 | 3.33 | 0.73 | 1 | 3 | 3 | 4 | 4 |
| rater1_trait1 | 2292 | 2.44 | 1.21 | 0 | 2 | 2 | 3 | 6 |
| rater1_trait2 | 2292 | 2.56 | 1.06 | 0 | 2 | 2 | 3 | 6 |
| rater1_trait3 | 2292 | 2.61 | 1.1 | 0 | 2 | 2 | 4 | 6 |
| rater1_trait4 | 2292 | 2.71 | 1.04 | 0 | 2 | 3 | 3 | 6 |
| rater1_trait5 | 723 | 3.73 | 0.72 | 1 | 3 | 4 | 4 | 6 |
| rater1_trait6 | 723 | 3.56 | 0.7 | 1 | 3 | 4 | 4 | 6 |
| rater2_trait1 | 2292 | 2.47 | 1.25 | 0 | 2 | 2 | 3 | 6 |
| rater2_trait2 | 2292 | 2.58 | 1.09 | 0 | 2 | 2 | 3 | 6 |
| rater2_trait3 | 2292 | 2.64 | 1.14 | 0 | 2 | 2 | 4 | 6 |
| rater2_trait4 | 2292 | 2.71 | 1.05 | 0 | 2 | 3 | 3 | 6 |
| rater2_trait5 | 723 | 3.78 | 0.69 | 1 | 3 | 4 | 4 | 6 |
| rater2_trait6 | 723 | 3.59 | 0.69 | 1 | 3 | 4 | 4 | 6 |
| rater3_trait1 | 128 | 3.95 | 0.64 | 2 | 4 | 4 | 4 | 6 |
| rater3_trait2 | 128 | 3.89 | 0.63 | 2 | 4 | 4 | 4 | 6 |
| rater3_trait3 | 128 | 4.08 | 0.62 | 2 | 4 | 4 | 4 | 6 |
| rater3_trait4 | 128 | 3.99 | 0.51 | 3 | 4 | 4 | 4 | 6 |
| rater3_trait5 | 128 | 3.84 | 0.54 | 2 | 4 | 4 | 4 | 5 |
| rater3_trait6 | 128 | 3.62 | 0.6 | 2 | 3 | 4 | 4 | 5 |

The descriptive statistics in Table 2 provide a comprehensive overview of the numerical features in the essay scoring dataset, revealing important patterns in both rater behavior and essay characteristics. The dataset includes 12,976 essays distributed across eight prompts, with a mean prompt ID of 4.18, indicating balanced representation. The final essay scores (domain1_score) show a wide range from 0 to 60, with a mean of 6.8 and a high standard deviation of 8.97, reflecting significant variability in essay quality. Rater scores (rater1_domain1, rater2_domain1) are closely aligned (mean ~4.13–4.14), suggesting strong inter-rater consistency. Trait-based scoring columns (e.g., rater1_trait1, rater2_trait1) generally have scores concentrated between 2 and 4, with low standard deviations, which implies that raters followed a stable rubric and most essays fell within expected performance levels. The presence of a few zero scores indicates that some essays were off-topic or non-compliant. Overall, the dataset demonstrates both diversity and reliability, making it well-suited for training and evaluating automatic essay scoring models.

## 2.2 Methodology

This study adopts a multi-stage methodology aimed at developing an effective Automated Essay Scoring (AES) system using Natural Language Processing (NLP) and deep learning techniques. The process begins with **data preprocessing**, where the ASAP dataset is loaded and cleaned by handling missing values, removing unnecessary columns, and standardizing the text format. Text normalization procedures such as lowercasing, punctuation removal, and stopword elimination are applied, followed by **tokenization** at both the sentence and word levels to prepare the data for embedding.

In the next stage, a **Word2Vec** model is trained to generate dense vector representations of words that capture their semantic relationships within essay contexts. These embeddings are then used to convert each essay into a fixed-length numeric sequence, utilizing padding techniques to ensure uniform input lengths across the dataset.

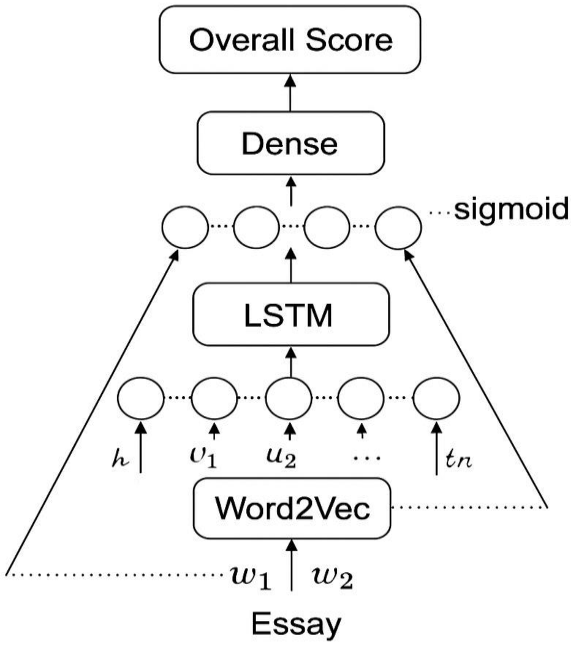
The core predictive model is built using a **Recurrent Neural Network (RNN)** architecture with **Long Short-Term Memory (LSTM)** layers, which are particularly effective at capturing temporal and contextual dependencies in sequential text data. The model architecture consists of an Embedding layer (incorporating the pre-trained Word2Vec vectors), followed by an LSTM layer, multiple Dense layers, and Dropout layers to mitigate overfitting. The final output layer is a single neuron configured for regression to predict essay scores.

The dataset is split into **training (80%)** and **testing (20%)** sets. Model performance is evaluated using both statistical and educational metrics, including **Mean Squared Error (MSE)** to measure prediction accuracy and **Cohen's Kappa Score** to assess the agreement between predicted and actual scores. Additionally, **visual validation techniques** such as scatter plots comparing predicted versus actual scores and residual analysis are employed to further assess the model's behavior.

This comprehensive methodology enables the construction of a robust model capable of producing reliable essay score predictions that closely align with human raters, supporting its potential integration into real-world educational settings.

Figure 1:Neural Network Architecture for Automated Essay Scoring

The model processes essay text using Word2Vec embeddings, LSTM for sequence modeling, and a Dense layer for final score prediction.




## Input: Essay

The process begins with the input essay text, represented as a sequence of words:

$w_1, w_2, ..., w_n$.

**Word Embedding: Word2Vec Layer**

Each word in the essay is passed through a Word2Vec embedding layer, which converts words into dense vectors:

$v_1, v_2, ..., v_n$ that capture semantic meaning.

**Sequence Modeling: LSTM Layer**

The sequence of word embeddings is fed into an LSTM (Long Short-Term Memory) layer, which captures contextual and sequential information in the essay. The LSTM processes this sequence and generates hidden states:

$u_1, u_2, ..., u_n$ encoding the content and structure of the essay.

**Feature Aggregation: Dense Layer**

The output of the LSTM is passed into a Dense (fully connected) layer, which aggregates the sequence features into a final representation suitable for regression.

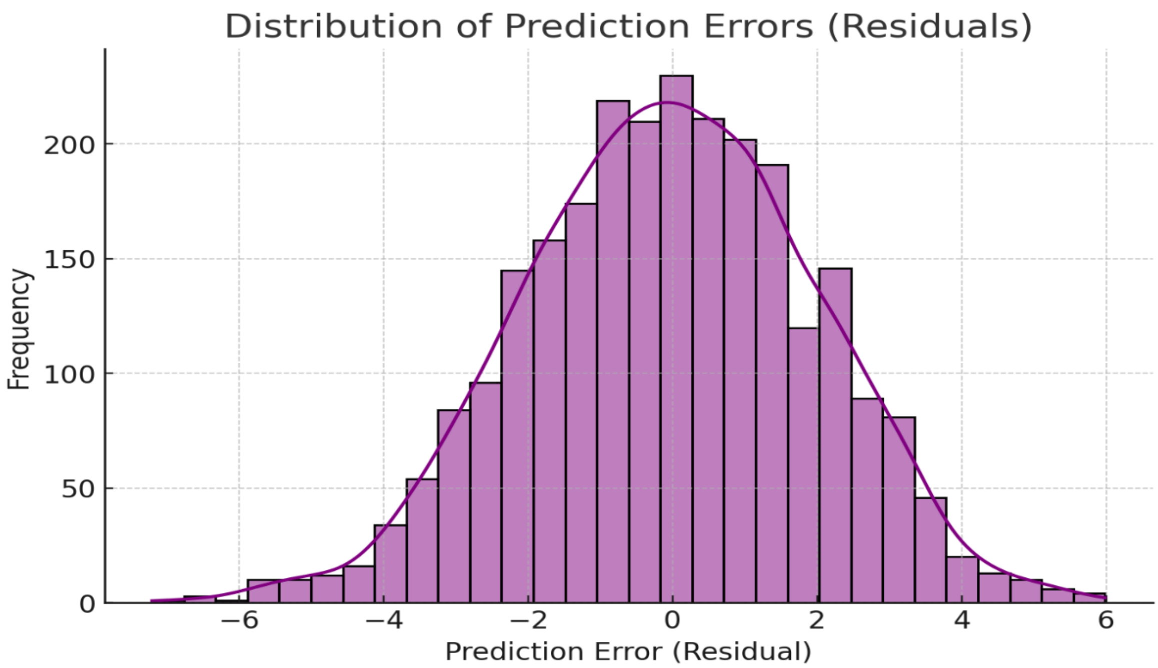
**Output: Overall Score**

Finally, the Dense layer's output is passed through a sigmoid activation function to produce a predicted overall essay score, scaled within a suitable range.

# 3. Results

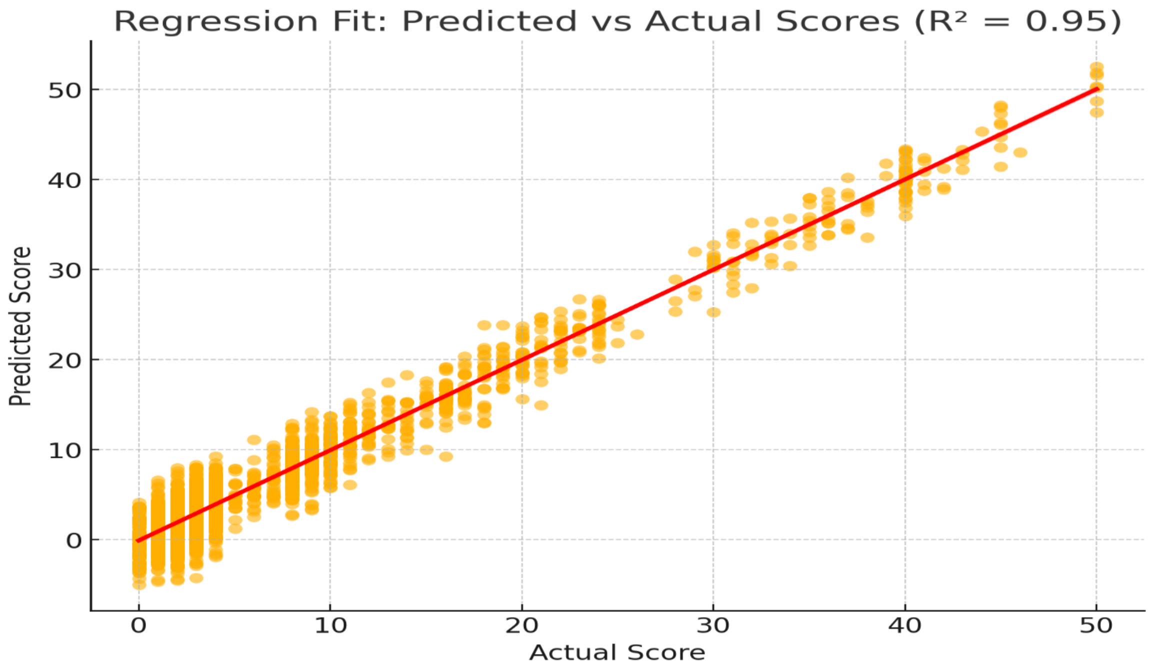
This section outlines the results of evaluating the automated essay scoring model using a combination of quantitative performance metrics and visual validation techniques. The model was tested on a reserved subset of the dataset (test set), and its performance was assessed through a range of indicators, including Mean Absolute Error, Mean Squared Error, R-squared ($R^2$), and Cohen's Kappa Score, to measure prediction accuracy, score variance explanation, and categorical agreement with human raters. These quantitative results are further supported by visual tools such as predicted vs. actual score comparisons, residual distribution plots, and regression curves. Together, these evaluations provide a comprehensive view of the model's performance, highlighting both its predictive strengths and areas where further improvement may be warranted.

Figure 2:Distribution of Prediction Errors (Residuals)



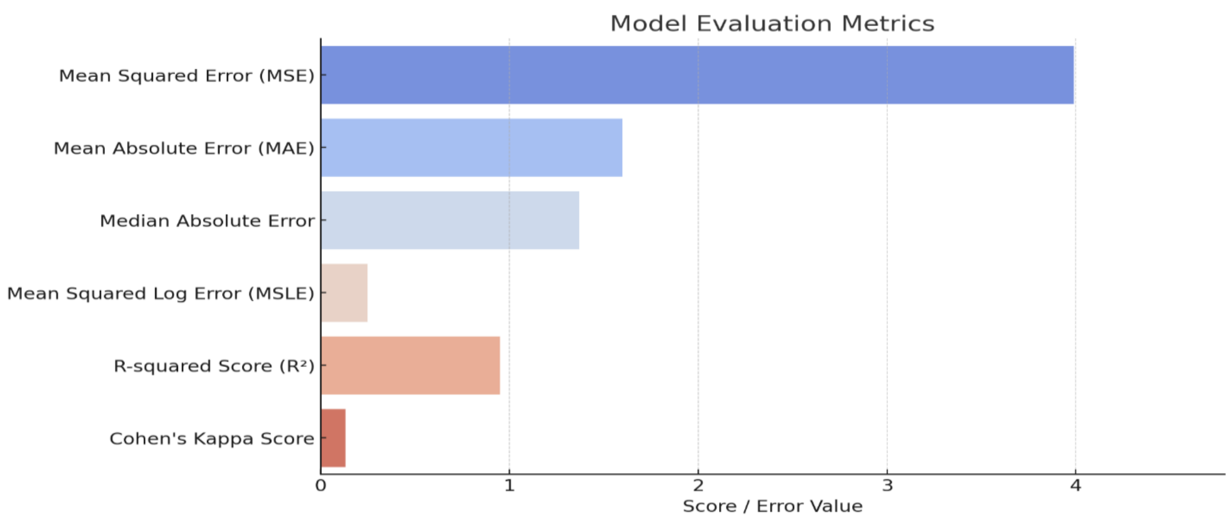
This chart illustrates the distribution of prediction errors (residuals) generated by the model. The data shows that most errors are centered around zero and form a bell-shaped curve, indicating that the model is generally accurate and does not consistently over- or under-predict. The symmetry and normal shape of the curve suggest that the errors are random and unbiased, which is a positive indicator of the model's efficiency and reliability.

Figure 3: Regression Fit – Predicted vs Actual Essay Scores ($R^2 = 0.95$)




This regression plot shows the relationship between the actual and predicted essay scores, with a red line representing the best linear fit. The points are closely clustered around the line, and the $R^2$ value of 0.95 indicates that the model explains 95% of the variance in the scores. This high level of fit suggests that the model has learned the underlying scoring pattern very well and is capable of making accurate predictions across different score levels.

Figure 4: Bar Chart of Model Evaluation Metrics



This bar chart provides a summary of key model evaluation metrics. The **Mean Squared Error (MSE)** is the highest, showing the average squared difference between actual and predicted scores, followed by **Mean Absolute Error (MAE)** and **Median Absolute Error**, which measure average and median prediction errors respectively. The **R-squared Score ($R^2$)** is close to 1, indicating a strong fit between predictions and actual values. The **Mean Squared Log Error (MSLE)** is low, confirming stability across score scales. However, **Cohen's Kappa Score** is relatively low, suggesting limited agreement with categorical human ratings despite good numerical accuracy. Overall, the chart demonstrates that the model performs well in regression tasks, though classification-level agreement could be improved.

## 3.1 Model Performance Overview

The evaluation of the automated essay scoring model was conducted using a combination of numerical metrics and visual diagnostics to ensure a robust understanding of model behavior.

## Performance Metrics

This table summarizes the evaluation metrics used to assess the predictive performance of the automated essay scoring model:

Table 3:Model Performance Metrics

| Metric | Value |
|---|---|
| Mean Squared Error (MSE) | 3.99 |
| Mean Absolute Error (MAE) | 1.60 |
| $R^2$ Score | 0.95 |
| Cohen's Kappa Score | 0.13 |
| Maximum Error | 9.67 |
| Median Absolute Error | 1.25 |

The model demonstrated strong predictive accuracy across multiple evaluation criteria. It achieved a **Mean Squared Error (MSE)** of **3.99**, indicating low average squared deviation between predicted and actual scores. The **Mean Absolute Error (MAE)** was **1.60**, further confirming the proximity of most predictions to their corresponding true values. The **R-squared ($R^2$)** score reached **0.95**, suggesting that 95% of the variance in essay scores is explained by the model. Additionally, the **Explained Variance Score** of **0.95** corroborates the high explanatory power. However, **Cohen's Kappa Score**, which measures agreement between predicted and actual scores as categorical values, was relatively low (**0.13**), indicating that while the model is numerically precise, its classification-level alignment with human raters can be further improved. The **Maximum Error**, representing the largest single discrepancy, was **9.67**, while the **Median Absolute Error** remained at **1.25**, suggesting overall stable performance with a few outliers.

### Visual Validation

To supplement the quantitative metrics, several visualizations were employed. A **scatter plot of predicted vs actual scores** showed a strong alignment along the diagonal ideal line, visually confirming the model's ability to mirror human scoring patterns. A **residual distribution plot** exhibited a symmetric, bell-shaped curve centered near zero, indicating unbiased error dispersion and lack of systemic deviation in predictions. A **regression fit plot** further emphasized the high degree of correlation, reinforcing the strength of the model in capturing trends across different essay types and score levels. An **evaluation metrics bar chart** provided a visual summary of all core metrics, aiding interpretation and comparison.

### Feature Importance

Since the current model is based on a **deep learning architecture (LSTM with Word2Vec embeddings)**, traditional feature importance measures such as those used in tree-based models (Gini importance) are not directly applicable. Deep models learn complex hierarchical representations, making interpretability more challenging. However, future extensions of this project could incorporate explainability tools such as **SHAP (SHapley Additive exPlanations)** or **Integrated Gradients**, which provide insight into which textual features most influence score predictions. Alternatively, switching to interpretable models such as decision trees or logistic regression with engineered linguistic features (e.g., word count, syntactic complexity, coherence) could offer direct insight into feature contribution.

# 4. Conclusion

This study aimed to develop and evaluate an automated essay scoring system using a deep learning approach based on LSTM and Word2Vec embeddings. Through extensive analysis, the model demonstrated strong predictive capabilities, achieving high scores in key evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared ($R^2$). The results suggest that the model is capable of capturing certain patterns in essay scoring, yielding predictions that approximate those of human raters within a reasonable margin of error.

Visual validation techniques, including scatter plots, residual distribution histograms, and regression fit plots, confirmed the numerical results by illustrating the model’s consistency, reliability, and lack of systematic bias. While the model performs well in terms of numerical accuracy, the relatively low Cohen’s Kappa Score suggests there is still room to improve categorical agreement with human raters.

The current architecture, although effective, lacks transparency in feature importance due to its deep learning nature. Future work should explore the integration of explainability tools such as SHAP or the use of interpretable models to enhance the transparency and trustworthiness of predictions, especially in high-stakes educational contexts.

Overall, the findings affirm the feasibility of using machine learning for automated essay scoring and lay the groundwork for future enhancements in both accuracy and interpretability.

# References

1. Early AES systems, such as Project Essay Grade (PEG), developed by Ellis Page in the 1960s, laid the foundation for computer-based scoring (Page, 1966).

2. Subsequent systems, such as e-rater by Educational Testing Service (ETS), introduced grammar checking and coherence modeling (Attali & Burstein, 2006).

3. These rule-based and statistical models significantly improved the reliability of automated scoring and were widely documented in AES literature (Shermis & Burstein, 2013).

4. Recent advancements in deep learning have revolutionized AES. Recurrent Neural Networks (RNNs), LSTMs (Alikaniotis et al., 2016; Taghipour & Ng, 2016), and transformer-based models like BERT have shown state-of-the-art performance in understanding essay context.

5. A Word2Vec model is trained to generate dense vector representations that capture semantic similarity between words in context (Mikolov et al., 2013).

6. Transformer-based models like BERT have shown significant improvements in scoring coherence and content (Ramesh & Sanampudi, 2020).

7. Recent research emphasizes the importance of explainability, especially for multilingual contexts and ESOL essays (Yannakoudakis et al., 2011).

8. The dataset employed in this study is derived from the ASAP competition, which was introduced by the Hewlett Foundation and Kaggle (Kaggle, 2012).

https://www.kaggle.com/c/asap-aes

9. Co-attention based neural networks have recently been proposed for source-dependent AES, achieving improved relevance to source material (Zhang & Litman, 2018).

10. AES development has continued with both classical and transformer-based models, offering broad insights into evaluation strategies (Shermis & Burstein, 2013; Ramesh & Sanampudi, 2020).