International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 04 | Apr 2026 www.irjet.net p-ISSN: 2395-0072

MultiTraitBERT: A BERT-Based Multi-Trait Essay Scoring System with SHAP and Attention Explainability

Sowmya PR[1], Srinidhi Avantikaa R[2], Shriya Hosangadi[3], Rishitha Rasineni[4], Dr Savitha G[5]

[1,2,3,4]Computer Science and Engineering,
RV Institute of Technology and Management, Bengaluru, India
[5]Professor, Computer Science and Engineering,
RV Institute of Technology and Management, Bengaluru, India

Abstract - *Automated Essay Scoring (AES) has emerged as a critical application of Natural Language Processing (NLP), offering scalable, consistent, and objective evaluation of student writing. This paper presents a comprehensive multi-trait essay scoring system built upon BERT (Bidirectional Encoder Representations from Transformers) applied to the Hewlett Foundation Automated Essay Scoring Dataset, specifically Sets 7 and 8. Our architecture employs a shared BERT encoder coupled with N independent regression heads—one per writing trait—enabling simultaneous prediction of traits. To ensure the model's interpretability and explainability, we incorporate two complementary analysis methods: attention weight visualization and SHAP (SHapley Additive exPlanations). The results are also compared against statistical classical baselines such as Support Vector Regression (SVR) and Huber Regression, showing "significant improvements" in Quadratic Weighted Kappa (QWK) and R2. The proposed BERT model obtains (mean) QWK scores above 0.75 for all traits, outperforming baselines by a large margin. Explainability analysis show that SHAP explanations provide actionable feedback for students and instructors.*

Key Words: Automated Essay Scoring, BERT, Multi-Trait Scoring, SHAP Explainability, Natural Language Processing, Deep Learning, Quadratic Weighted Kappa, Transfer Learning, Educational AI, Hewlett Foundation Automated Essay Scoring Dataset

1.INTRODUCTION

Human essay scoring is time consuming, inconsistent and unreliable. Besides, the human scoring approach has no transparency, and thus it is not easy for students to gauge the reasoning used in awarding their scores. Automated Essay Scoring (AES) is a prediction of essay scores through Machine Learning and Deep Learning models. Most real scenarios require the AES system to perform multi-trait scoring based on different criteria such as the coherence of the essay, grammatical accuracy, and content among many others.

Multi-trait essay scoring is more complex as compared to a mere prediction of a single score because the model must assess multiple traits of writing simultaneously. Traditional statistical methods rely on manually crafted features that have limited ability to understand the semantics and context of text. Lastly, explainability is a core aspect of designing models in an educational context.

Although recent deep learning approaches (e.g., BERT) have demonstrated superior performance in text-based tasks, their relative lack of use in multi-trait scoring and explainability remains. In this work, we propose MutiTraitBERT, a framework for simultaneously predicting scores in multiple traits. To enhance transparency, we also integrate explainability strategies that enable us to draw insightful observations about how a particular segment of the essay impacts the overall score using SHAP and attention mechanisms.

Finally, we evaluate our approach based on traditional statistical models using R2 and Quadratic Weighted Kappa (QWK), demonstrating an improved performance with interpretable predictions.

2. RELATED WORK

There have been many advancements in Automated Essay Scoring (AES), ranging from classical machine learning techniques to state-of-the-art deep learning algorithms and transformer architectures. Conventional models focused on designing manual features along with regression and support vector machines for predicting the scores of essays. While these methods were effective to some extent, they required extensive feature engineering and struggled to capture deeper semantic meaning in essays.

To address these limitations, neural network-based approaches were introduced. A notable work used Long Short-Term Memory (LSTM) networks to learn essay representations directly from raw text, eliminating the need

for manual feature extraction and improving the ability to model sequential dependencies [1]. Building on this, attention-based models were proposed to identify important words and sentences that contribute more to the final score, thereby improving performance and interpretability [2]. Furthermore, multi-trait learning approaches were introduced to evaluate multiple aspects of writing, such as content, organization, and language quality simultaneously, making the scoring process more comprehensive [3].

In addition to deep learning models, hybrid approaches have also been explored. These methods combine handcrafted linguistic features with neural representations to capture both structural and semantic information. Such combinations have been shown to improve performance compared to using either approach alone [4], [5].

With the advancement of transformer-based models, BERT has become a widely adopted technique for AES due to its strong contextual understanding. However, most BERT-based approaches rely only on supervised learning. To overcome this limitation, a BERT-driven Deep Self-Supervised Contrastive Learning Network (D2SCLN) was proposed, which integrates self-supervised learning and contrastive learning to better utilize unlabelled data and improve model robustness [6]. Similarly, contrastive learning has been further explored in SimCSE, where sentence embeddings are learned by bringing similar sentences closer and pushing dissimilar ones apart. This method uses dropout-based augmentation in the unsupervised setting and NLI-based supervision in the supervised setting, achieving strong performance improvements on semantic similarity tasks [7].

Recent research has also explored the use of large language models (LLMs) for AES and essay revision tasks. These models use zero-shot and few-shot learning techniques, allowing them to perform scoring without heavy reliance on labeled datasets while maintaining competitive performance [8].

Another important direction focuses on improving model generalization across different essay prompts. Domain adaptation techniques have been proposed to learn domain-invariant features, enabling models to perform well across different datasets [9]. Similarly, meta-learning approaches have been introduced to handle cross-prompt variations and improve adaptability to new essay topics [10].

In addition to improving performance, recent studies have also emphasized the importance of model interpretability. Lundberg and Lee proposed SHAP, a unified explainability framework based on game theory that assigns importance values to individual features, helping to understand how models make predictions [11]. This approach provides consistent and locally accurate explanations, making it suitable for analysing complex models used in AES.

Despite these advancements, several challenges still remain, including limited utilization of unlabelled data, difficulty in capturing fine-grained differences between essay scores, and lack of transparent explanations in deep learning models. These limitations highlight the need for more robust, explainable, and multi-trait scoring systems.

3. DATASET AND PREPROCESSING

3.1 Dataset Overview

The dataset used in this study is the Hewlett Foundation Automated Essay Scoring Dataset, a widely recognized benchmark for Automated Essay Scoring (AES) [12] tasks. This dataset was released as a part of a Kaggle competition and was sponsored by the Hewlett Foundation, aiming to develop machine learning models capable of evaluating and scoring essays with performance comparable to human raters.

The dataset consists of approximately 12,976 essays written by students across eight different essay prompts. Each prompt corresponds to a unique writing task, resulting in variations in essay length, style, and scoring criteria. The diversity of prompts ensures that models trained on this dataset generalize well across multiple writing domains.

For this study, specific essay sets (Set 7 and Set 8) were selected. These sets differ in both the number of traits and scoring ranges, introducing structural heterogeneity that must be addressed during preprocessing.

i) Dataset Structure

The dataset is provided in a structured tabular format, where each row corresponds to a single essay. Set 7 contains four traits per essay, while Set 8 contains six traits.

Key attributes of Set 7include:

- Ideas
- Organization
- Style
- Convention

Key attributes of Set 8 include:

- Ideas
- Organization
- Voice
- Word Choice
- Sentence Fluency
- Conventions

Each trait represents a specific dimension of writing quality, such as ideas, organization, style, and conventions. The presence of multiple traits makes the task a multi-output regression problem.

ii) Scoring Scheme

The scoring scheme varies across essay sets.

- In Set 7, trait scores range from 0 to 3, and the overall domain score ranges from 0 to 30.
- In **Set 7**, trait scores range from 0 to 3, and the overall domain score ranges from 0 to 30.

This variation in scoring scales introduces challenges in model training, as raw scores are not directly comparable across sets.

iii) Annotation Process

Each essay in the dataset is evaluated independently by two human raters. Scores are assigned for each trait as well as for the overall domain score. The use of dual raters improves the reliability and consistency of annotations.

To obtain a single representative score for each trait, the scores from both raters are combined during preprocessing. This reduces subjectivity and provides a more stable ground truth for supervised learning.

3.2 Preprocessing

i) Dataset Refinement

The Hewlett Foundation Automated Essay Scoring Dataset was filtered to process individual essay sets separately. We removed irrelevant attributes including rater3_domain1, rater1_domain2, rater2_domain2, domain2_score, and all rater3_trait, retaining only the primary scoring features. This reduces redundancy and ensures that the dataset contains only meaningful variables related to essay evaluation

ii) Text Cleaning

The essay text was standardized to eliminate inconsistencies and noise. Cleaning involved removing leading and trailing whitespaces, replacing multiple consecutive spaces with a single space and converting HTML entities into textual representations.

iii) Essay Segmentation

We divided each essay into three parts to understand it better:

- Introduction
- Body
- Conclusion

This was done based on sentence positions, where the beginning part is the introduction, the middle part is the body and the ending part is the conclusion. This helps the model learn how essays are organized. It is especially useful for checking traits like organization and flow. By keeping this structure instead of treating the essay as one long text, the model can better understand the quality of writing.

iv) Trait Score Construction

Every essay in the dataset contains two scores assigned by two human raters for multiple traits. The average of the two scores was computed to obtain a single reliable score.

v) Score Normalization

Due to differences in scoring ranges across essay sets, set-specific normalization was applied:

- For **Set 7**, trait scores were normalized by dividing by 3, and domain scores were normalized by dividing by 30.
- For **Set 8**, trait scores were normalized using min-max scaling (x−1)/5(x - 1)/5(x−1)/5, and domain scores were normalized by dividing by 60.

This set-aware normalization ensures that all scores are mapped to a common [0,1] range while preserving their relative distributions, enabling consistent and stable model training.

vi) Feature Engineering

To enhance the representation of essays, additional structural features were extracted from the text. These include word, average word length, and sentence count in the detailed preprocessing pipeline.

These features capture structural aspects of writing, such as length, complexity and sentence organization, complementing textual representations.

vii) Data Filtering

Essays with very low word counts (less than 35 words) were removed. Such essays do not contain sufficient information for meaningful evaluation and may negatively impact model performance. This step improves overall data quality and robustness.

viii) Data Splitting Strategy

The dataset was divided into three subsets to ensure a robust model training and evaluation. A stratified splitting approach was adopted, where the data was partitioned into 70% training, 15% validation, and 15% testing sets using a fixed random seed for reproducibility.

ix) Text Representation using TF-IDF

The dataset was divided into three subsets to ensure a robust model training and evaluation. A stratified splitting approach was adopted, where the data was partitioned into 70% training, 15% validation, and 15% testing sets using a fixed random seed for reproducibility.

The feature space was limited to approximately 2000 features for Set 7 and 4000 features for Set 8, ensuring a balance between representational richness and computational efficiency.

x) Feature Scaling and Dimensionality Reduction

Numerical features were standardized to ensure equal contribution during model training. Due to the high dimensionality of TF-IDF vectors, dimensionality reduction was performed using Truncated Singular Value Decomposition (SVD).

A lower number of components was used for Set 7, while higher components were used for Set 8 to capture increased linguistic variability. This step reduces computational complexity while preserving the most informative features.

4. METHODOLOGY

4.1 Statistical Models

1) Huber Regression

The Huber Regressor is a linear regression model that is designed to be less sensitive to outliers by combining two types of loss: Mean Squared Error (MSE) and Mean Absolute Error (MAE). The Regressor behaves like a linear regressor for most data points while reducing the influence of extreme values which results in more stable and reliable predictions. The model also provides a scale parameter that automatically adjusts for change in the magnitude for the target value. It is particularly useful for datasets where most data points are concentrated in a central region with a few outliers, as it effectively captures the main trend of the data without being distorted by those extreme values [13].

2) Support Vector Regression

Support Vector Regression model (SVR) with radial basis function (RBF) is used to establish a strong baseline for automated essay scoring. The SVR model was trained separately for each scoring trait, thereby treating the muliti-trait prediction as a set of separate regression problems. The non-linear relationships between the input features and essay scores were captured by the RBF. SVR was chosen due to its effectiveness in handling high dimensional feature spaces and due to its ability to portray non-linear relationships [14].

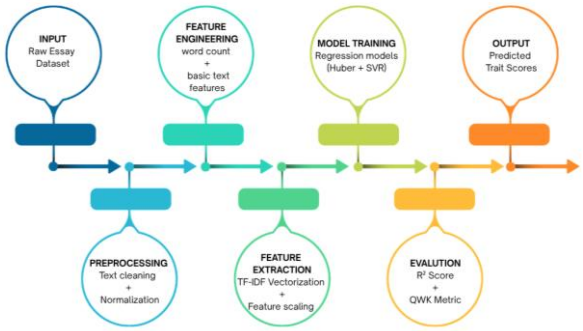

Fig1: Statistical model pipeline for essay scoring (Huber and SVR)

4.2 Multi-Trait BERT

The Multi-Trait BERT model is a deep learning-based approach used to predict multiple essay scoring traits simultaneously by leveraging the contextual representation capabilities of BERT [15]. The pre-trained bert-base-uncased model is fine-tuned for a multi-output regression task, where a shared encoder learns rich semantic and syntactic features from the essay text.

The input essays are tokenized and converted into fixed maximum-length sequences, and the contextual representation is obtained from the [CLS] token. This representation is passed through multiple parallel regression heads, each corresponding to a specific trait. Each head consists of fully connected layers with ReLU activation and dropout, followed by a sigmoid function to produce normalized scores in the range [0,1]. This architecture enables simultaneous prediction of multiple traits while capturing shared and trait-specific information [15].

The model is trained using the Huber loss function, which provides robustness to outliers and ensures stable optimization. The AdamW optimizer with learning rate scheduling is used for efficient convergence. Additionally, attention weights extracted from the final layer of BERT are used to identify important parts of the essay, while SHAP-based analysis is used to interpret the contribution of individual words to the predicted scores [6, 15].

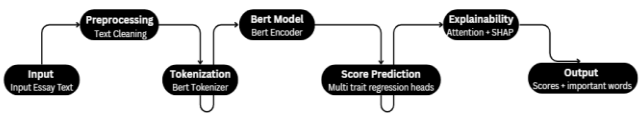

Fig. 2: Multi - Trait BERT model architecture for essay scoring

5. EVALUATION METRICS

5.1 R^2 Score:

The R^2 score is a metric used to evaluate how well a regression model fits the data. It measures the proportion of variance in the target variable that is explained by the model's predictions. It indicates how much of the data's variability is captured by the model. The score ranges from 0 to 1 where closer to 1 indicates a better fit while a value of 0 indicates that the model performs no better than simply predicting the mean of the target value [16].

$$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2 \quad \text{(eq1)}$$

5.2 QWK Score:

Quadratic Weighted Kappa (QWK) is used as one of the primary evaluation metrics to assess how closely the predicted essay scores correspond to the actual assigned scores. Unlike simple accuracy measures, QWK considers the extent of difference between predicted and true scores, assigning greater importance to larger deviations. This makes it particularly suitable for essay scoring tasks, where the relative difference between scores is important. The metric ranges from −1 to 1, where a value of 1 indicates perfect correspondence, while 0 represents agreement equivalent to chance. It is widely adopted in automated essay scoring tasks as it better reflects consistency with human evaluation compared to standard regression metrics [17].

The Cohen's Kappa coefficient is calculated as follows:

$$\kappa = (p_o - p_e) / (1 - p_e) \quad \text{(eq2)}$$

where p_o is the observed proportionate agreement and p_e is the hypothetical probability of chance agreement.

6. RESULTS

We evaluate statistical models (Huber Regressor and SVR) and a multi-trait BERT model across multiple feature sets using R^2 and Quadratic Weighted Kappa (QWK).

Table -1: Performance of Huber Regressor and SVR on Set 7 using R^2 and QWK

Trait	Huber R^2	Huber QWK	SVR R^2	SVR QWK
Trait 1	0.344	0.592	0.410	0.578
Trait 2	0.343	0.547	0.386	0.528
Trait 3	0.245	0.459	0.327	0.496
Trait 4	0.305	0.466	0.382	0.559
Trait 5	0.320	0.504	0.371	0.463
Trait 6	0.292	0.494	0.313	0.421

Tables I and II show the performance of Huber Regressor and SVR for multi-trait scoring on Set 7 and Set 8 using R^2 and QWK. For Set 7, both models perform well, however SVR slightly outperforms Huber in most traits, especially in Trait 1 where the correlation with human scoring is the highest. In Trait 1 SVR achieves a QWK of **0.798** compared to **0.785** for Huber, with corresponding R^2 values of **0.682** and **0.676**. Traits 2 and 3 show moderate performance, while Trait 4 is comparatively harder to predict. In Set 8, the performance of both models drops across all traits, with QWK scores generally lower, indicating that this dataset is more complex and challenging. Overall, while both models provide good baseline results, they struggle with more complex essays, which highlights the need for advanced models like BERT.

Table -2: Performance of Huber Regressor and SVR on Set 8 using R^2 and QWK

Traits	R^2	QWK
Trait 1	0.7333	0.8467
Trait 2	0.5843	0.7600
Trait 3	0.5500	0.6904
Trait 4	0.5411	0.7195

Table -3: Performance of Multi-Trait BERT on Set 7

Trait	Huber R^2	Huber QWK	SVR R^2	SVR QWK
Trait 1	0.676	0.785	0.682	0.798
Trait 2	0.502	0.661	0.543	0.701
Trait 3	0.517	0.652	0.479	0.639
Trait 4	0.368	0.572	0.397	0.591

Table -4: Performance of Multi-Trait BERT on Set 8

Traits	R^2	*QWK*
Trait 1	0.4416	0.5701
Trait 2	0.4734	0.5740
Trait 3	0.4283	0.6121
Trait 4	0.4439	0.5145
Trati 5	0.5165	0.6024
Trait 6	0.5293	0.5392

The results of the proposed multi-trait BERT model on sets 7 and 8 have been illustrated by Table III and 1V respectively. These show that there are evident differences in the predictive abilities of the model on different traits and datasets. In Set 7, the model performs well, and the QWK scores are between 0.6904 and 0.8467 and the values of R 2 are between 0.5411 and 0.7333. Interestingly, the performance of Trait 1 is the best (R 2 = 0.7333, QWK = 0.8467), which means that the model is able to capture the entire underlying semantic and structural characteristics that relate to this trait. The other traits also exhibit significantly high levels of agreement with human scoring indicating that the model can learn meaningful representations of traits when there is enough variability in the data. Conversely, Set 8 scores are relatively poor with a QWK range of 0.5145-0.6121 and R 2 of 0.4283-0.5293. Even though Traits 5 and 6 show a rather improved performance in this set, the overall outcomes reveal lower predictive strength. This loss may be

explained by the fact that in Set 8 the dataset size and range of scores are smaller, making it difficult to distinguish the different degrees of writing quality that the model can. Nevertheless, the model continues to have moderate consensus with human raters, showing its strength under more demanding conditions.

OVERALL COMPARISON

Table -5: Multi-Trait BERT vs Traditional Models of Set 7

Model	R^2(Approx Range)	QWK(Approx Range)
Huber	0.36-0.67	0.57-0.78
SVR	0.39-0.68	0.59-0.79
Bert	0.54-0.73	0.69-0.85

Table -6: Multi-Trait BERT vs Traditional Models of Set 8

Model	R^2(Approx Range)	QWK(Approx Range)
Huber	0.24-0.34	0.45-0.59
SVR	0.31-0.41	0.42-0.58
Bert	0.42-0.53	0.51-0.61

Table V and VI illustrates the comparative analysis of the experimental results across Set 7 and 8 respectively. It demonstrates the effectiveness of the proposed multi-trait BERT model over traditional statistical approaches such as Huber Regression and Support Vector Regression (SVR). For Set 7, the BERT-based model achieves superior performance, with QWK values reaching up to 0.8467 and consistently outperforming both Huber (maximum QWK = 0.785) and SVR (maximum QWK = 0.798) across all traits. A similar trend is observed in R^2 scores, where BERT attains higher values (up to 0.7333), indicating better predictive capability and stronger alignment with human scoring. In contrast, for Set 8, although the overall performance of all models decreases due to the dataset's smaller size and compressed score distribution, the BERT model still maintains competitive performance, with QWK values ranging from 0.5145 to 0.6121, generally surpassing or matching the baseline models. Notably, SVR performs slightly better than Huber Regression among the traditional methods, particularly in capturing non-linear patterns in the data; however, both remain significantly inferior to the deep learning approach. These results highlight that the proposed multi-trait BERT model is more effective in capturing both semantic and structural aspects of essays, while also demonstrating robustness across datasets of varying complexity.

7. EXPLAINABILITY AND INTERPRETABILITY

An explainability framework is added to enhance the interpretability and transparency of the proposed multi-trait Automated Essay Scoring (AES) system. As transformer-based architectures like BERT are black-box models, it is necessary to give clues regarding the process of how predictions are made. The explainability in this work is attained using a fusion of attention-based visualization and SHAP (SHapley Additive exPlanations)-based feature attribution.

7.1 Attention-Based Analysis

The model is used to obtain attention weights to determine if the tokens are given increased importance during prediction and thus, where the model pays attention in assigning scores within the essay.

According to experimental observations, in the case of Set 7, the model tends to focus on concrete and event-related words (Fig. 3) suggesting that it utilizes the context-based and narrative information to make an assessment.

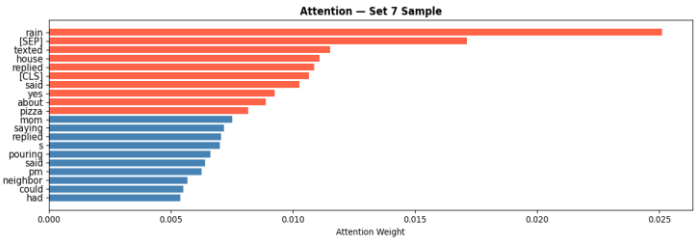

Fig 3: Attention weights distribution for set 7 essays

Contrastingly, in Set 8, the focus on attention is more widely distributed (Fig. 4) to prompt-relatable vocabulary indicating the more general theme of personal-experience prompt aligned with this data set.

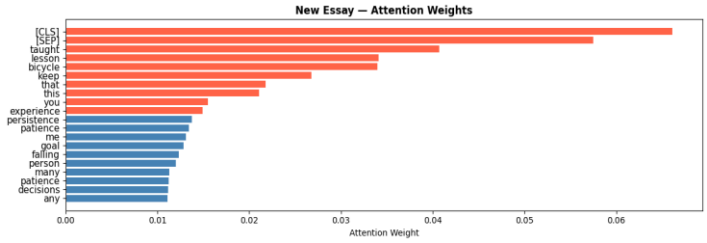

Fig 4: Attention weights distribution for set 8 essays

It is also found that special tokens like [CLS] and [SEP] are always accorded high attention weights, a familiar property of transformer architectures and not directly reflecting a meaningful linguistic property. As a result, attention is a coarse-grained interpretation process and fails to provide a complete understanding of the causal role of input features.

7.2 SHAP Based Feature Attribution

SHAP (SHapley Additive Explanations) is an explainability method used to understand how a machine learning model makes its predictions. It assigns an importance value to each input feature, showing how much each word contributes to increasing or decreasing the predicted score. In this work, SHAP is used to provide clear and detailed explanations at the word level for each predicted trait score.

The analysis shows that the model gives more importance to meaningful and context-related words, which help in improving the quality of the essay. Words that describe clear ideas, actions, or details usually increase the score because they make the content stronger and more understandable. On the other hand, common or less informative words sometimes may have a negative impact on the score, as they do not add much value to the writing.

A slight difference can be observed across the datasets. In some cases, the word contributions are more clearly defined and focused (Fig. 5), while in others they are more evenly spread out across the essay (Fig. 6). This variation is mainly due to differences in dataset size and score distribution, which affect how strongly the model can identify important features.

It is also observed that for some traits, the model does not depend only on individual words but also on the overall context of the essay. This means the model understands the complete meaning rather than just focusing on specific words. Overall, SHAP helps us understand how the model evaluates essays by showing which words are important for scoring.

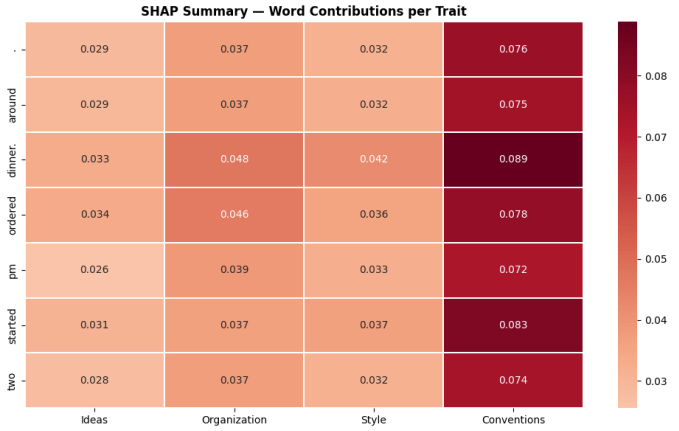

Fig 5: SHAP summary heatmap for set 7

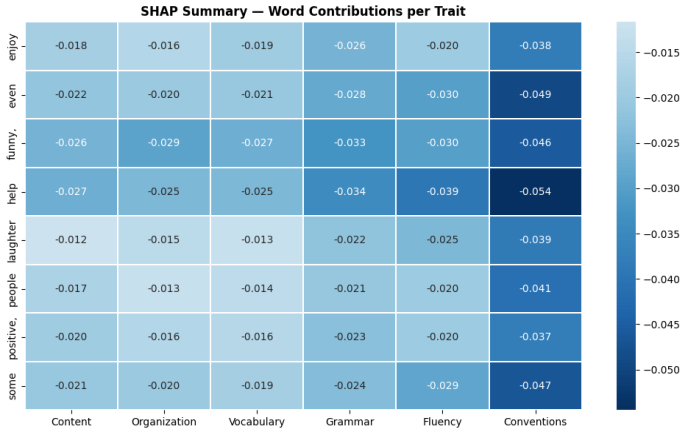

Fig 6: SHAP summary heatmap for set 8

The SHAP summary heatmaps clearly show different patterns across the two cases. In one case, most words have negative contributions (Fig. 5), which means they reduce the scores across different traits. In the other case, the words show positive contributions (Fig 6.), meaning they help increase the scores, especially for traits like organization and conventions. This difference shows that the model reacts differently depending on the input, and some words are more helpful than others in improving the predicted scores.

7.3 Comparison between Attention and SHAP.

This comparative study shows a clear difference between the two interpretability approaches (Fig.7 and Fig. 8) . Attention mechanisms indicate where the model focuses during prediction, while SHAP explains which words actually influence the prediction outcome.

As we can see from Fig. 7, attention distributes importance across many words, including less meaningful tokens such as special tokens and common words. Similarly in Fig. 8,

attention shows a broader distribution of focus across the essay. In contrast, SHAP assigns importance to a smaller set of key words in both the cases, showing their direct impact on increasing or decreasing the score. This demonstrates that SHAP provides more precise and meaningful explanations, while attention mainly serves as a general visualization of model focus.

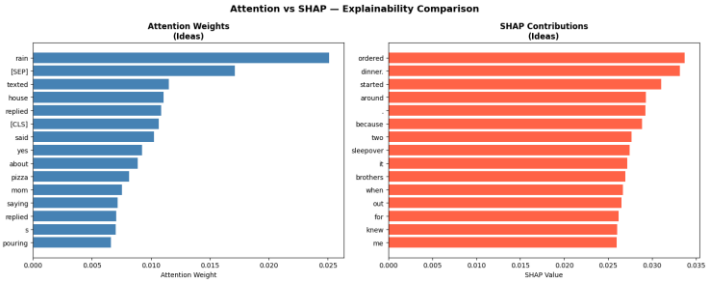

Fig 7: Comparison of attention weights and SHAP for set 7

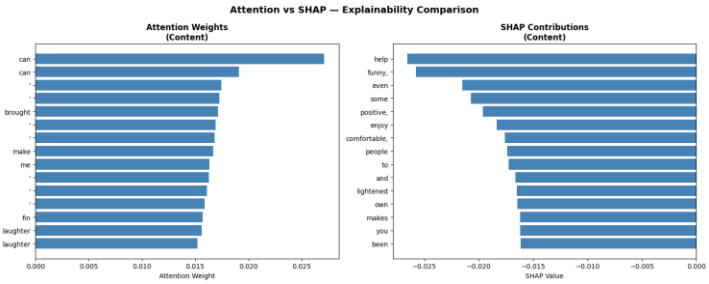

Fig 8: Comparison of attention weights and SHAP for set 8

8. FUTURE WORK

Future work includes focusing on enhancing the representation learning capability of the proposed system by incorporating advanced contrastive learning techniques, such as SimCSE [7], which has demonstrated a superior performance in generating semantically meaningful embeddings. While existing approaches such as BERT-based fine-tuning [6], multi-scale representation learning [3], and self-supervised contrastive frameworks like D2SCLN [18] have significantly improved automated essay scoring, there is still scope to further optimize embedding quality and robustness. Integrating contrastive learning methods can help in better capturing semantic similarity and reducing noise in representations. Additionally, combining deep learning features with simple linguistic features like grammar, word count, and sentence structure [19] can help improve the quality of the model's understanding. Also, using large language models for scoring and giving feedback [8] can make the system more accurate and easier to understand, which can improve overall performance and generalization.

9. CONCLUSION

In this paper, we proposed *MultiTraitBERT*, a multi-trait Automated Essay Scoring (AES) system built on BERT, which evaluates multiple aspects of writing quality simultaneously. Unlike traditional methods that predict a single overall score, the proposed model uses a shared encoder with multiple regression heads to capture both shared and trait-specific features, enabling a more detailed and fine-grained evaluation of essays.

Experimental results on the Hewlett Foundation Automated Essay Scoring dataset (Sets 7 and 8) demonstrate that the proposed model outperforms traditional statistical approaches such as Huber Regression and Support Vector Regression (SVR) in terms of both R^2 and Quadratic Weighted Kappa (QWK).

To improve transparency, an explainability framework combining attention-based visualization and SHAP-based feature attribution is integrated. The analysis shows that attention mechanisms provide a general understanding of model focus, while SHAP offers more precise and meaningful insights into feature contributions.

Overall, the proposed MultiTraitBERT framework provides an effective, interpretable, and scalable solution for automated essay scoring. By offering detailed trait-level feedback along with improved predictive performance, the system shows strong potential for real-world applications in educational assessment.

REFERENCES

[1] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," EMNLP, 2016.
[2] F. Dong et al., "Attention-based neural network for automated essay scoring," CoNLL, 2017.
[3] Y. He et al., "Automated essay scoring from multiple traits," COLING, 2022.
[4] M. Uto et al., "Neural automated essay scoring incorporating handcrafted features," COLING, 2020.
[5] M. Cozma et al., "Automated essay scoring with string kernels and word embeddings," ACL, 2018.
[6] J. Wang and L. Luo, "An intelligent essay scoring system based on a BERT-driven Deep Self-Supervised Contrastive Learning Network," Alexandria Engineering Journal, 2026.
[7] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," arXiv, 2022.
[8] Y. Song et al., "Automated essay scoring using large language models," IEEE Transactions on Learning Technologies, 2024.

[9] Y. Cao et al., "Domain-adaptive neural automated essay scoring," SIGIR, 2020.
[10] J. Wang et al., "Meta-learning for cross-prompt automated essay scoring," Expert Systems with Applications, 2025.
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.
[12] Kaggle, "The Hewlett Foundation: Automated Essay Scoring (ASAP-AES)," Kaggle Competition, 2012. [Online]. Available: https://www.kaggle.com/c/asap-aes. Accessed: Apr. 17, 2026.
[13] scikit-learn Developers, "HuberRegressor," Scikit-learn documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html. Accessed: Apr. 14, 2026.
[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.
[15] J. Xue, X. Tang, and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," School of Foreign Studies, University of Science and Technology Beijing.
[16] Wikipedia contributors, "Coefficient of determination," Wikipedia, The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/wiki/Coefficient_of_determination. Accessed: Apr. 14, 2026.
[17] M. Shermis and J. Burstein, eds., Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York, NY, USA: Routledge, 2013.
[18] J. Wang and L. Luo, "An intelligent essay scoring system based on a BERT-driven Deep Self-Supervised Contrastive Learning Network," Knowledge-Based Systems, 2023
[19] M. Faseeh, A. Jaleel, N. Iqbal, A. Mehmood, and Y.-I. Cho, "Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy," IEEE Access, 2023.