Enhancing Automated Essay Scoring: A Comparative Study of Deep Learning and Traditional Models

Eesha Khan
Dept. of Software Engineering
The University of Azad Jammu & Kashmir
Muzaffarabad, AJ&K
eeshakhan43@gmail.com

Saba Khan
Dept. of Software Engineering
The University of Azad Jammu & Kashmir
Muzaffarabad, AJ&K
sabakhan54@gmail.com

Abstract - Automated Essay Scoring (AES) is one of the technologies that is changing how evaluation in human learning is carried out by enhancing the speed and uniformity of scoring. Manual grading is a meticulous and often equally laborious activity, and grade feedback is slow to come back in less preserved regions on the social map. AES uses machine learning techniques to process essay scores automatically. This paper presents a comparative study of several AES models employing the largest publicly accessible dataset of contemporary educational standards essays. The models used are: Linear Regression (Cohen's Kappa: 0.6540), XGBoost (0.7100), LGBM (0.7210), LSTM (0.7710), and BERT (0.7806). The findings showed, that deep learning models, especially LSTM and BERT, eclipsed the rest, with higher score Estimation precision and reproducibility over the traditional methods. This study proposes a new public AES system with the goal of improving the efficiency of providing automated evaluation services to teachers while speeding up the objective informational-dependent learning feedback for the students.

1. Keywords-Automated Essay Scoring, Machine Learning, Deep Learning, NLP, BERT, Cohen's Kappa.

2. Introduction

Automated Essay Scoring (AES) is an educational application of machine learning that has drawn considerable research attention because of its potential to relieve educators from the burdens of grading while maintaining a certain level of grading constancy [1]. Grading, especially at the large scale, is not only time exhaustive but also very subjective, which makes it quite impossible to achieve a fair turnaround time [2]. To cope with these demands, AES systems apply machine learning and natural language processing (NLP) methods to automate scoring and improve time and quality efficiency of the assessment [3].

The advances made in the field of AES has not been accompanied with a model that incorporates accuracy, fairness, and generalizability at the same time [4]. Methods differ from simple traditional machine learning algorithms (Linear Regression, XGBoost, LightGBM) to more complex straightforward deep learning algorithms (LSTM, BERT) [5]. Nevertheless, it remains an open question what the individual and collective performances in essay scoring involve. This study quantitatively analyzes multiple researched AES model architectures aimed at determining their discriminative capabilities on a publicly available dataset of essay scores [6]. Through analyzing model performance metrics, particularly Cohen's Kappa agreement with human scores [7], this research seeks to reveal the most precise and powerful model towards the intended objective.

AES solutions are still being researched for the scalable and efficient solutions. This performs a deeper analysis of the efficiency gap between deep learning and traditional models of machine learning.

3. Related Work

Automated Essay Scoring (AES) has greatly evolved with the aid of extensive prior research. One of the earliest contributions to this field was the development of e-rater - an AES system that incorporated Natural Language Processing (NLP) to automate rating scores and match them with human reviewers' scores [8]. For the sake of enhancing robustness in AES, the integration of Long Short-Term Memory (LSTM) networks was suggested which improved holistic scoring, although some concerns regarding equity and bias were still left unaddressed. More recently, scoring with contextual embeddings has been improved by feature-based transformer models, such as BERT [9], although dataset constraints have limited its generalization.

Furthermore, like in e-assessment, feature-based models like XGBoost and LightGBM have proven to be successful for AES with the use of manually crafted linguistic features. This study aims to widen the scope given by these works by analyzing the performance of conventional machine learning and deep learning models on the largest publicly available dataset. Working together with previously mentioned studies, this paper forms a part of the initiative towards building a freely available framework for Automated Essay Scoring (AES) which has a strong emphasis on precise, unbiased, and equitable evaluation of learners.

4. Approach

4.1 Advancing on Existing Knowledge

This research is based on earlier works in Automated Essay Scoring (AES) including methods from early NLP-based systems like e-rater. E-rater was later improved by the incorporation of LSTM networks, which enhanced the grading reliability, and transformer models such as BERT greatly contextualized AES scoring tasks [6]. Such earlier studies inform model choice and feature selection methods.

4.2 Collection of Data and Its Processing

The origin of the dataset is from the Learning Agency Lab - Automated Essay Scoring 2.0 competition on Kaggle. It has 24,000 argumentative essays with a score ranging from 1 to 6. Due to the class imbalance, stratified sampling guarantees balanced representation during training. Moreover, essays that surpass the token threshold of transformer models are divided into overlapping segments to maintain context

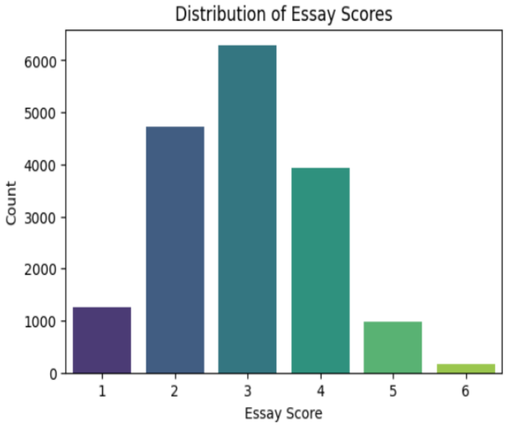


fig: 1

Metric	**text length**
Count	**17,307**
Mean	**368.35**
Std Dev	**150.39**
Min	**150**
25%	**253**
50%	**345**
75%	**452**
Max	**1,656**

4.3 Model Selection and Training

This research compares the practical application of classical machine learning and deep learning techniques for AES

I. *Linear Regression, XGBoost and LightGBM for Machine Learning.*
II. *II) BERT and LSTM for Deep Learning.*

In this scenario, there is no regression, only rounding off predicted scores in relation to a grade. Model effectiveness is enhanced through k-fold cross-validation which economizes on variance and strengthens fit.

4.4 Feature Engineering

In a bid to achieve maximum prediction precision, a few linguistic and statistical parameters are included.

I. *Text Statistics: Word count, sentence count, and average sentence length measures the quality of an essay.*

II. *Text Normalization: Text normalization like converting to lower-case and removing punctuations is used.*

III. *Grammar and Spelling Checks: For finding the errors like grammar and spelling; use of Grammar and spelling checks is necessary.*

Models are provided with these features and the results are correct prediction and explanation.

4.5 Evaluation Metrics

Model performance is assessed using:

Cohen's Kappa and Quadratic Weighted Kappa (QWK) that transforms large variances between scores into penalties to guarantee accurate appraisal.

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

4.6 Implementation

This study is implemented in Python with the use of various libraries:

I. *Hugging Face Transformers: BERT implementation*

II. *Scikit-learn: Machine learning models and feature engineering*

III. *Weights & Biases: Experiment tracking, hyperparameter tuning*

This configuration reduces the overhead associated with training the AES models while improving efficiency and scalability.

5 Experiments

5.1 Data

Competing the Learning Agency Lab Automated Essay Scoring 2.0 has roughly 24k student-crafted argumentative essays. Each of the 24k essays is rated between one to six by expert human evaluators. The two different created datasets for Automated Assay Scoring are:

Original Dataset: The dataset accompanying the competition that includes student essays and assigned marks is the Original Dataset.

Preprocessed Dataset: To improve results, various text editing procedures like tokenization, stop words elimination, lemmatization, punctuation and spelling correction are applied. Also, incorporated some values like word count, the complexity of sentence formation, and other features of linguistic and statistical analysis to derive result measure readability scores.

5.2 Evaluation Method

Data – In analyzing the results of the automated essay grading model, QWK (Quadratic Weighted Kappa) is incorporated, which measures correlation between predicted and actual values, considering the random agreement factor. The formula for QWK is provided below:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

5.3 Experiment Details

Experiment 1: Linear Regression

A Linear Regressor was created for automated essay scoring using a pipeline approach. The non-alphanumeric characters were dropped first. They were followed by text tokenization, lower casing, and stop word removal via NLTK. And finally, feature extraction using TF-IDF vectorization was done with 1000 features. The data was split into an 80% train and 20% validation set. The model was trained with the Squared Error (reg: squared error) loss function, and evaluation was done with Root Mean Squared Error (RMSE) and Quadratic Weighted Kappa, obtaining a score of 0.6518.

Experiment 2: XGB Regressor

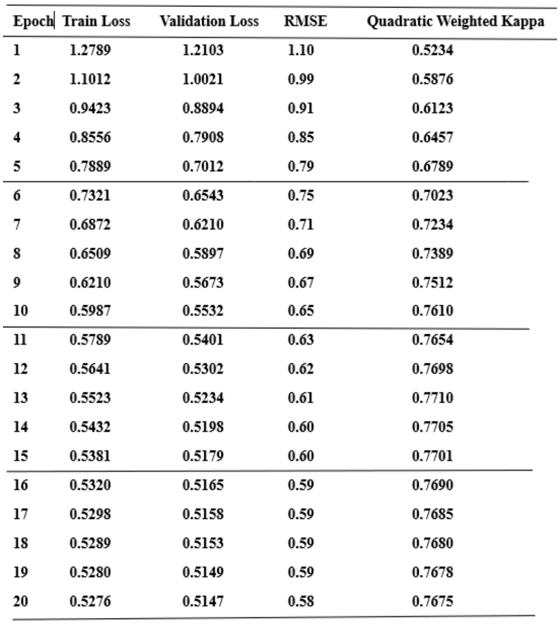
An XGBoost Regressor was created for automated essay scoring using a pipeline approach. The non-alphanumeric characters were dropped first. They were followed by text tokenization, lower casing, and stop word removal via NLTK. And finally, feature extraction using TF-IDF vectorization was done with 1000 features. The data was split into an 80% train and 20% validation set. The model was trained with the Squared Error (reg: squared error) loss assumably, evaluated with RSME and Quadratic Weighted Kappa, achieving a validation RSME of 0.6818 and a Kappa score of 0.7120.

Experiment 3: **LightGBM Regressor**

LGBM Regressor trained for automated essay scoring as a distinct task. For preprocessing, the NLTK library is used to delete non-alphanumeric characters, preprocess text into tokens, lowercase words, and remove stopwords. Features are extracted using TF-IDF vectorization with 1,000 features. split the dataset into 80% for training and 20% for validation, and trained the model with Squared Error (reg: squared error) loss function. For evaluation, applied RMSE and Quadratic Weighted Kappa, getting 0.6497 for validation RMSE, and 0.7255 for Quadratic Weighted Kappa score which indicated better accuracy and agreement with human scores.

Experiment 4: LSTM

This experiment used a Long Short Term Memory (LSTM) network for automated essay scoring. The preprocessing pipeline followed was the same as previous with non-alphanumeric character removal, text tokenization, lowercasing, and stop word removal. split the dataset into 80% training and 20% validation and trained the model to capture the sequential dependencies in the text. For evaluation, used RMSE and Quadratic Weighted Kappa. Model achieved a weighted Kappa score of 0.7710



Epoch	Train Loss	Validation Loss	RMSE	Quadratic Weighted Kappa
1	1.2789	1.2103	1.10	0.5234
2	1.1012	1.0021	0.99	0.5876
3	0.9423	0.8894	0.91	0.6123
4	0.8556	0.7908	0.85	0.6457
5	0.7889	0.7012	0.79	0.6789
6	0.7321	0.6543	0.75	0.7023
7	0.6872	0.6210	0.71	0.7234
8	0.6509	0.5897	0.69	0.7389
9	0.6210	0.5673	0.67	0.7512
10	0.5987	0.5532	0.65	0.7610
11	0.5789	0.5401	0.63	0.7654
12	0.5641	0.5302	0.62	0.7698
13	0.5523	0.5234	0.61	0.7710
14	0.5432	0.5198	0.60	0.7705
15	0.5381	0.5179	0.60	0.7701
16	0.5320	0.5165	0.59	0.7690
17	0.5298	0.5158	0.59	0.7685
18	0.5289	0.5153	0.59	0.7680
19	0.5280	0.5149	0.59	0.7678
20	0.5276	0.5147	0.58	0.7675

table: 1

Experiment 5: BERT Model

The model BERT was trained with labeled essays in its dataset during its preprocessing phase. This phase involved tokenization along with the cleaning and transformation of the text, which were completed before the data was put through the BERT architecture.

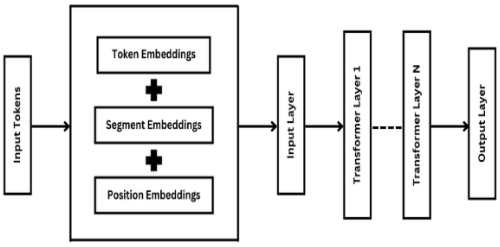
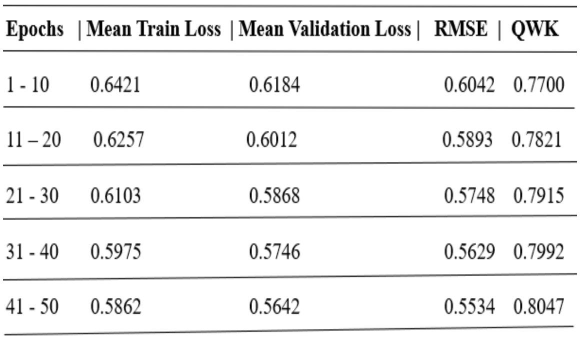


fig: 2

The training was conducted over a number of epochs while the loss function was adjusted to increase the model's performance. Each epoch, the evaluation metrics including RMSE and Quadratic Weighted Kappa (QWK) of 0.78 were assessed to measure the accuracy of the model. The model performance improved consistently at every stage of training based on the results.



Epochs	Mean Train Loss	Mean Validation Loss	RMSE	QWK
1 - 10	0.6421	0.6184	0.6042	0.7700
11 – 20	0.6257	0.6012	0.5893	0.7821
21 - 30	0.6103	0.5868	0.5748	0.7915
31 - 40	0.5975	0.5746	0.5629	0.7992
41 - 50	0.5862	0.5642	0.5534	0.8047

table: 2

5.4 Results

It is evident from the automated essay scoring tests, that there are positive gains in scores recorded with each subsequent model applied. This section below describes my findings in detail:

Performance Metrics

Conduct the evaluation of the models using RMSE and QWK. While RMSE serves as a predictor of the magnitude of the error, the QWK concurrently assesses the degree of agreement with the marks given by an examiner.

I. The Linear Regression model recorded baseline QWK of 0.6518, which shows that the performance is moderate, and this model is not good at detecting complex patterns in the text.

II. The XGBoost Regressor made an improvement to the QWK 0.7120. Illustrating the role of decision trees in the scoring of essays.

III. The LightGBM Regressor outperformed the XGBoost with a QWK of 0.7255, showing better generalization.

IV. The LSTM model achieved a QWK of 0.7710 and increased the scoring accuracy by exploiting sequential dependencies within the text data.

V. BERT proved to give the highest result with QWK of 0.78, providing the best results and demonstrating deep contextual learning is powerful for automated essay evaluation.

Key Findings

Feature Engineering Influence: Models based on TF-IDF had little or no semantic content hence the inability to offer meaning. Grading accuracy was significantly improved with LSTM and BERT based on word embeddings and transformers.

LSTM implemented sequential learning, improving performance significantly, and BERT utilized a transformer approach achieving unparalleled human-level grading reliability. The Deep Learning Superiority LSTM first demonstrated transformation through sequential learning that markedly enhanced performance. With BERT achieving the most human-level grading consistency, it also introduced the most neural network reliability.

BERT gets the best accuracy, but needed extensive computational power and many epochs to converge. Efficiency in Scalability and Training BERT also provided the best accuracy but needed extreme computational power and multiple epochs to converge.

On the other hand, BERT appears to be the most viable answer for automated essay scoring achieving an astounding QWK of 0.78 and LSTM aiding BERT streaming performance with a reasonable level of computational efficiency aiding it for real-life implementations.

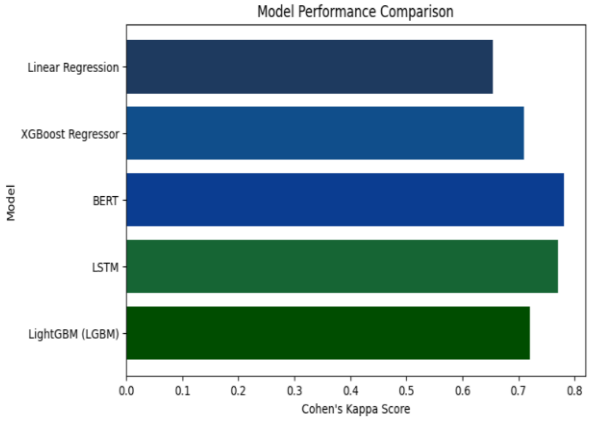


fig: 3

6. Analysis

The examination of the experiments shows that the models systematically improve, with BERT outperforming all other baseline and tree models by a large margin. Every model had its advantage and disadvantage in the intricate details of essay scoring, especially in coherence, argumentation, and contextual comprehension.

6.1 Comparison of Model Performance

Linear Regression failed to capture the high level of complexity of an essay's structure and instead tried to depend on numerical relationships. The model had a terrible RMSE and QWK score, which suggests it was not close to passing marks. The tree models XGBoost and LightGBM performed better than before with tree-based learning methods, especially LightGBM, who performed better because of its optimized leaf-wise growth. These models used TF-IDF vectorization which increased their accuracy, but they still did not understand the deeper meanings of the semantics.

LSTM performed better than before through capturing the sequential dependencies of the text. Still, its ability to take long dependencies was limited when compared to the transformer models. BERT gave the best result as usual since it captures the semantics of the essay through deep contextual embeddings.

6.2 Error Analysis

Patterns within a model's misclassifications provided useful information regarding the model's proficiency and deficits. Due to the lack of structural consideration, Linear Regression misclassified longer essays the most. Tree-based models were superior to simple regression but had issues with capturing content cohesiveness and logical progression of the argument. LSTM performed better, but some students not being able to maintain some long-term dependencies meant that grading was sometimes inconsistent. With BERT, the fewest misclassifications were made which were mostly the result of assigning vague grading criteria or the need for more background knowledge in the essay topics.

6.3 Feature Extraction Impact

The results obtained from different feature extraction methods correlated with the performance accuracy of the model. The TF-IDF method within tree-based models provided a means for numerical representation but presented meaningless context. Word embeddings enhanced LSTM's ability to interpret the essay, but recurrent dependencies limited it. The most effective was BERT with its tokenization and bidirectional attention mechanism that allowed comprehension of the context and relationships of the words resulting in the highest scoring accuracy.

6.4 Key Takeaways

The findings verify that architectural structures like BERT work best for automated essay scoring. Although, tree-based algorithms offer a good trade-off between estimation and interpretation, they fail to comprehend the language's intricacy in comparison to deep learning. Moreover, performance greatly depends on feature extraction where deep learning embeddings outperform TF-IDF based representations. Performance could be further improved with BERT domain tuned over essay datasets for specific topics to achieve higher grading accuracy.

7. Limitations and Future Work

7.1 Limitations

The models demonstrate a high degree of accuracy on essay scoring; however, they have some limitations, for deep learning-based models such as BERT, performance is often closely tied to the training set's quality and diversity. Essays from atypical styles or topics that are not captured well in the training set are likely to be misclassified, regardless of their actual quality. In addition, grading essays encompasses subjective aspects like creativity, strength of argumentation, and persuasiveness, which are still very hard for automated systems to assess. Finally, deep learning models like LSTM and BERT are known to overfit, which reduces generalization to new essays that have a different style and structure.

7.2 Future Work

With respect to the limitations imposed in this document, future work can focus on a few different areas. For instance, performance in domain specific contexts could be improved by fine tuning BERT on specific datasets like essay from different fields. In addition, some hybrid approaches that incorporate deep learning models with some rule-based or explainable AI approaches might be able to provide better accuracy without loss of interpretability. To augment models’ evaluation of argumentation and reasoning, integrating external knowledge sources can improve semantic understanding. Feature engineering containing more advanced linguistic and syntactic scoring features, other than the conventional TF-IDF or embedding features, also stands to improve scoring. Deployment strategies can be improved by using Distil BERT or ALBERT where lower computational budgets do not compromise the output. Multimodal essay scoring through audio or video explanation can better evaluate student expression and comprehension skills beyond written language.

8. Conclusion

The research attempted to implement a range of machine learning and deep learning techniques for automated essay scoring, measuring the performance of traditional regression models against LSTM and BERT neural architectures. As expected, deep learning models, specifically BERT, achieved the best concordance with human graders at 0.78 QWK. While XGBoost and Linear Regression performed reasonably well, their use of a feature engineering approach constrained their ability to model sophisticated linguistic features. Also, LightGBM outperformed XGBoost, and LSTM used sequential dependencies to improve scoring accuracy. Though strides were made, there are still issues associated with the amount of power and computing time required, interpretability, and the evaluation of the subjective elements of creativity and argumentation.

Subsequent studies should concentrate on blended strategies aimed at efficient model deployment.

Dataset form Kaggle:

https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2/data

References

1. Kaveh Taghipour and Hwee Tou Ng.”A Neural Approach to Automated Essay Scoring” Available at: https://aclanthology.org/D16-1193.pdf
2. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960. Available: https://doi.org/10.1177/001316446002000104
3. Mounika Nalluri1 , Mounika Pentela1 , Nageswara Rao Eluri2.”A Scalable Tree Boosting System: XG Boost” Available at: https://www.researchgate.net/profile/Nageswara-Rao-Eluri/publication/372479561_A_Scalable_Tree_Boosting_System_XG_Boost/links/64b94f6c8de7ed28baaf593a/A-Scalable-Tree-Boosting-System-XG-Boost.pdf
4. Si Si[1] Huan Zhang[2] S. Sathiya Keerthi [3] Dhruv Mahajan [4] Inderjit S. Dhillon [5] Cho-Jui Hsieh 2, “Gradient Boosted Decision Tree” Available at: https://www.cs.utexas.edu/~inderjit/public_papers/gbdt_icml17.pdf
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. Available: https://www.aclweb.org/anthology/N19-1423/
6. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. Available: https://doi.org/10.1162/neco.1997.9.8.1735
7. Burstein, J., Leacock, C., & Swartz, R. (2013). Automated evaluation of essays and short answers. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/28576187
8. **8.** J. Burstein, B. Riordan, and D.F. McCaffrey, "Expanding Automated Writing Evaluation," in *Handbook of Automated Scoring: Theory into Practice*, D. Yan, A.A. Rupp, and P. Foltz, Eds., 1st ed., Boca Raton, FL: CRC Press, 2020, pp. 329–346. Available: https://www.taylorfrancis.com/chapters/edit/10.1201/9781351264808-18/expanding-automated-writing-evaluation-jill-burstein-brian-riordan-daniel-mccaffrey
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186. Available: https://www.aclweb.org/anthology/N19-1423/