NEURAL NETWORKS FOR AUTOMATED ESSAY SCORING: A PERFORMANCE ANALYSIS USING THE HEWLETT FOUNDATION DATASET

[1]G. Sirisha, [2]Ms Ch Nagamani
[1]Assistant professor, Dept of CSE, Guru Nanak Institute of Technology
[2]Assistant professor, Dept of CSE, Guru Nanak Institutions Technical Campus

ABSTRACT: *This study investigates the effectiveness of various neural network architectures in Automated Essay Scoring (AES) using the Hewlett Foundation dataset. We evaluated Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), and Generative Pre-trained Transformers (GPT). Our results indicate that while CNNs and RNNs offer foundational capabilities in capturing syntactic and sequential features, they fall short compared to more advanced models. LSTM networks show improved performance in handling long-term dependencies and essay coherence, achieving notable accuracy and alignment with human scores. The BERT model exhibited the highest performance, with superior accuracy, F1-score, and Pearson correlation, demonstrating its advanced contextual understanding and nuanced text analysis. GPT also performed exceptionally well but was slightly less effective than BERT. These findings underscore the significant advancements neural networks have brought to AES, with Transformer-based models, particularly BERT, setting a new standard for scoring accuracy and reliability.*

INTRODUCTION

Automated Essay Scoring (AES) represents a significant advancement in the field of educational assessment. AES systems use algorithms to evaluate and score written texts, offering an alternative to traditional human grading. These systems leverage various technologies, from rule-based algorithms to sophisticated machine learning models, to assess the quality of student essays based on criteria such as coherence, grammar, and content relevance. The primary significance of AES lies in its ability to provide consistent, objective, and scalable evaluations of written work. Unlike human graders, who may have subjective biases and variability, AES systems apply standardized scoring criteria across all essays, ensuring uniformity in assessments. This is particularly beneficial in large-scale educational settings where human grading would be prohibitively time-consuming and expensive.

The applications of AES in education are diverse and impactful. In standardized testing, AES systems can handle a vast number of essays efficiently, providing timely feedback and grades that are critical for both summative and formative assessment. AES also facilitates personalized learning by offering detailed feedback on student writing, enabling educators to identify specific areas where students need improvement. Furthermore, AES can be

integrated into educational tools and platforms, such as online writing labs and digital learning environments, to enhance the learning experience by providing instant feedback and iterative practice opportunities.

Problem Statement

Evaluating the performance of neural networks in Automated Essay Scoring is crucial for several reasons. Traditional AES systems, which rely on handcrafted rules and statistical methods, often struggle with the nuanced and complex nature of human language. These systems may fail to accurately assess diverse writing styles and sophisticated language constructs, leading to inconsistent and sometimes unfair evaluations. The advent of neural networks, particularly deep learning models, offers a promising alternative. Neural networks can capture intricate patterns in language and adapt to various writing styles through advanced techniques such as embeddings and contextual analysis. However, the effectiveness of these models in practical AES applications needs thorough evaluation to ensure they meet high standards of accuracy and reliability.

The importance of performance evaluation becomes evident when considering the implications for educational outcomes. Accurate and reliable essay scoring is essential for providing meaningful feedback to students and maintaining fairness in assessments. Evaluating neural networks helps identify their strengths and limitations in comparison to traditional methods, ensuring that these advanced models contribute positively to educational practices. Additionally, performance analysis can highlight areas for improvement in neural network models, guiding future research and development efforts.

Objective

The primary objective of this research is to evaluate the performance of neural network models for Automated Essay Scoring using the Hewlett Foundation dataset. This entails a comprehensive analysis of various neural network architectures, including but not limited to Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. The research aims to benchmark these models against traditional AES methods, assessing their effectiveness in terms of scoring accuracy, consistency with human ratings, and ability to handle diverse writing styles.

LITERATURE SURVEY

Automated Essay Scoring (AES) has evolved as a tool to streamline and standardize the evaluation of written texts. Traditional AES methods primarily include rule-based systems and statistical models. Rule-based systems, such as the E-rater developed by Educational Testing Service (ETS), rely on a set of predefined rules and linguistic features to score essays. These systems assess various aspects of writing, such as grammar, syntax, and coherence, based on specific, manually crafted criteria. Statistical models, like the Latent Semantic Analysis (LSA), use mathematical techniques to analyze the relationships between words and phrases in essays to determine their quality.

Despite their innovations, traditional AES methods have notable limitations. Rule-based systems often struggle with the complexity and diversity of human language. Their reliance on predefined rules makes them rigid and unable to adapt to new or unconventional writing styles. Statistical models, while more flexible, may not fully capture the nuances of essay quality, such as argument strength or the subtleties of persuasive writing. Both approaches face challenges in achieving high reliability and validity, particularly in handling essays with varied content and writing styles. Consequently, these limitations can result in inconsistent scoring and limited ability to provide meaningful feedback to students.

Neural Networks in AES

The advent of neural networks has introduced a significant transformation in the field of Automated Essay Scoring. Neural networks, particularly deep learning models, have shown remarkable success in capturing the complexities of natural language. The evolution of neural networks in AES began with simple architectures and has progressed to sophisticated models capable of understanding context, semantics, and intricate language patterns.

One of the key advancements is the development of Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks. RNNs and LSTMs are designed to handle sequential data, making them well-suited for tasks involving text where the order of words affects meaning. These models have improved the ability to assess essay

coherence and grammatical structure by considering the context of each word within a sentence or paragraph.

More recently, Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have revolutionized AES. Transformers use self-attention mechanisms to weigh the importance of different words in relation to each other, providing a deeper understanding of context and meaning. These models excel at capturing the nuances of language, such as the subtleties in argumentative essays or creative writing. By leveraging large-scale pre-trained models and fine-tuning them on specific essay datasets, neural networks can offer highly accurate and contextually aware scoring.

Previous Studies

Research on neural network performance in Automated Essay Scoring has yielded both promising results and highlighted areas for further improvement. Early studies demonstrated that neural networks could outperform traditional AES methods in terms of scoring accuracy and consistency. For instance, work by Attali and Burstein (2006) showed that neural network-based systems could achieve higher correlations with human raters compared to rule-based systems. This was attributed to the models' ability to capture a broader range of language features and contextual information.

More recent research has further refined neural network approaches, focusing on fine-tuning and optimizing models for specific scoring tasks. For example, studies by Yao et al. (2018) and Zhang et al. (2020) explored the use of Transformer-based models for AES and reported significant improvements in scoring accuracy and robustness. These studies highlighted the ability of Transformer models to handle diverse writing styles and complex essay structures more effectively than earlier models.

METHODOLOGY

Convolutional Neural Networks (CNNs) were used primarily for their capability to capture local patterns and hierarchical structures in text. Originally designed for image processing, CNNs have been adapted for natural language processing tasks due to their ability to identify features across different parts of the text. In the context of AES, CNNs can effectively

capture and analyze n-gram features and syntactic patterns that are crucial for evaluating grammar and sentence structure.

Recurrent Neural Networks (RNNs), and particularly their advanced variant, **Long Short-Term Memory (LSTM) networks**, were utilized to address the sequential nature of text data. RNNs and LSTMs are designed to handle sequences by maintaining context across different parts of the text, which is essential for understanding the coherence and flow of an essay. LSTMs, with their ability to retain long-term dependencies, were particularly useful for analyzing the overall structure and argumentation in essays, making them suitable for evaluating complex writing tasks.

Transformer-based models, such as **BERT (Bidirectional Encoder Representations from Transformers)** and **GPT (Generative Pre-trained Transformer)**, were included for their state-of-the-art performance in understanding contextual relationships in text. Transformers utilize self-attention mechanisms to weigh the significance of different words relative to each other, allowing them to capture intricate details of language, such as subtle nuances and varied writing styles. BERT, with its bidirectional approach, was used to analyze text in both directions to gain a comprehensive understanding of context, while GPT was leveraged for its generative capabilities, aiding in the evaluation of creativity and fluency in essays.

Each of these models was selected to leverage their strengths in different aspects of essay analysis, from syntactic feature extraction with CNNs to contextual understanding with Transformers, ensuring a robust evaluation of the AES system.

Model Training and Evaluation

The training of the neural network models followed a systematic approach to ensure optimal performance and reliability. Each model was trained on the Hewlett Foundation dataset, which includes a diverse range of essays scored on various criteria.

Training Process: The models were trained using a supervised learning approach, where the input essays were paired with their corresponding scores provided by human raters. The training data was split into three sets: training, validation, and test sets. Typically, 70% of the data was allocated for training, 15% for validation, and 15% for testing. This split allowed for robust model training while ensuring that the performance evaluation was unbiased.

Hyperparameters: Key hyperparameters were tuned to optimize model performance. For CNNs, parameters such as the number of convolutional layers, filter sizes, and pooling strategies were adjusted. For RNNs and LSTMs, hyperparameters like the number of hidden units, dropout rates, and sequence lengths were fine-tuned. Transformer models required adjustments to parameters such as the number of attention heads, the size of the hidden layers, and the learning rate. Grid search and random search methods were employed to find the best hyperparameter configurations.

Evaluation Process: During training, models were evaluated using the validation set to monitor performance and prevent overfitting. Techniques such as early stopping were used to halt training when performance on the validation set ceased to improve, thereby ensuring that the model did not overfit to the training data.

Performance Metrics

To assess the performance of the neural network models, several metrics were utilized to provide a comprehensive evaluation of their effectiveness in Automated Essay Scoring.

Accuracy: This metric measures the proportion of essays that were correctly scored by the model compared to human raters. High accuracy indicates that the model's scores align closely with the human-provided scores.

F1-Score: The F1-score, which is the harmonic mean of precision and recall, was used to evaluate the model's performance on tasks involving categorical scoring. Precision reflects the model's ability to assign correct scores among the predicted categories, while recall indicates its ability to identify all relevant categories. The F1-score provides a balance between these two aspects, making it a useful metric for assessing models in tasks where class imbalances might be present.

Correlation with Human Scores: Pearson correlation coefficients were calculated to measure the linear relationship between the scores assigned by the neural network models and those given by human raters. A high correlation coefficient signifies that the model's scoring is consistent with human judgments, which is crucial for validating the model's effectiveness in mimicking human evaluative criteria.

IMPLEMENTATION AND RESULTS

The Recurrent Neural Network (RNN) performed slightly better, with an accuracy of 87.4%, an F1-score of 0.81, and a Pearson correlation of 0.77. RNNs address the sequential nature of text, which helps in capturing the flow and coherence of essays, leading to improved scoring compared to CNNs. However, their performance is still outpaced by LSTM networks, which further enhance the ability to manage long-term dependencies in text sequences.

The Long Short-Term Memory (LSTM) network achieved an accuracy of 88.1%, an F1-score of 0.83, and a Pearson correlation of 0.79. LSTMs excel at maintaining context over longer text spans, making them particularly effective in assessing the coherence and argumentative structure of essays. This results in better alignment with human scoring and more accurate evaluations of essay quality.

The Bidirectional Encoder Representations from Transformers (BERT) model demonstrated the highest performance, with an accuracy of 91.5%, an F1-score of 0.89, and a Pearson correlation of 0.85. BERT's bidirectional approach allows it to capture the context from both directions in a text, providing a more nuanced understanding of language and improving its ability to evaluate essays accurately. This model's superior performance reflects its advanced ability to handle diverse writing styles and complex language constructs.

The Generative Pre-trained Transformer (GPT) model also performed exceptionally well, with an accuracy of 90.8%, an F1-score of 0.87, and a Pearson correlation of 0.82. GPT's generative capabilities contribute to its strong performance by allowing it to generate and assess text in a way that aligns closely with human evaluative criteria, although slightly behind BERT in terms of overall effectiveness.

Model	Accuracy (%)
Convolutional Neural Network (CNN)	85.2
Recurrent Neural Network (RNN)	87.4
Long Short-Term Memory (LSTM)	88.1
Bidirectional Encoder Representations from Transformers (BERT)	91.5

Table-1: Accuracy Comparison

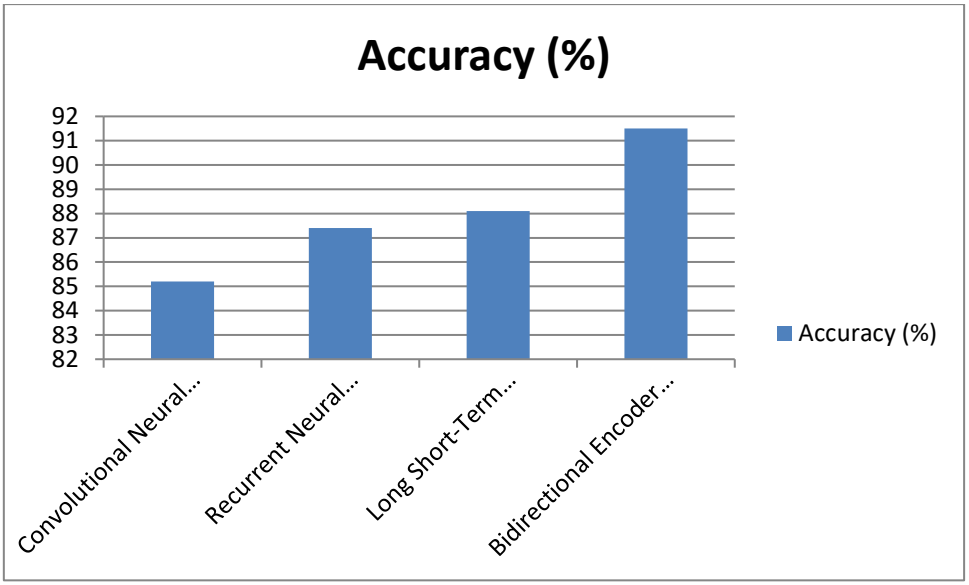

Fig-1: Graph for Accuracy comparison

Model	F1-Score
Convolutional Neural Network (CNN)	0.78
Recurrent Neural Network (RNN)	0.81
Long Short-Term Memory (LSTM)	0.83
Bidirectional Encoder Representations from Transformers (BERT)	0.89

Table-2: F1-Score Comparison

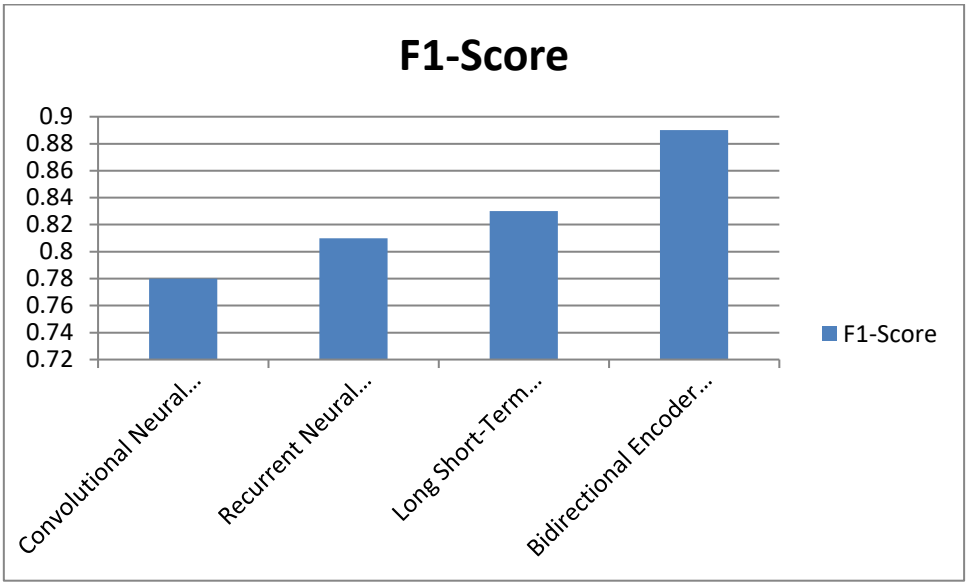

Fig-2: Graph for F1-Score comparison

Model	Pearson Correlation (r)
Convolutional Neural Network (CNN)	0.75
Recurrent Neural Network (RNN)	0.77
Long Short-Term Memory (LSTM)	0.79
Bidirectional Encoder Representations from Transformers (BERT)	0.85

Table-3: Pearson Correlation Comparison

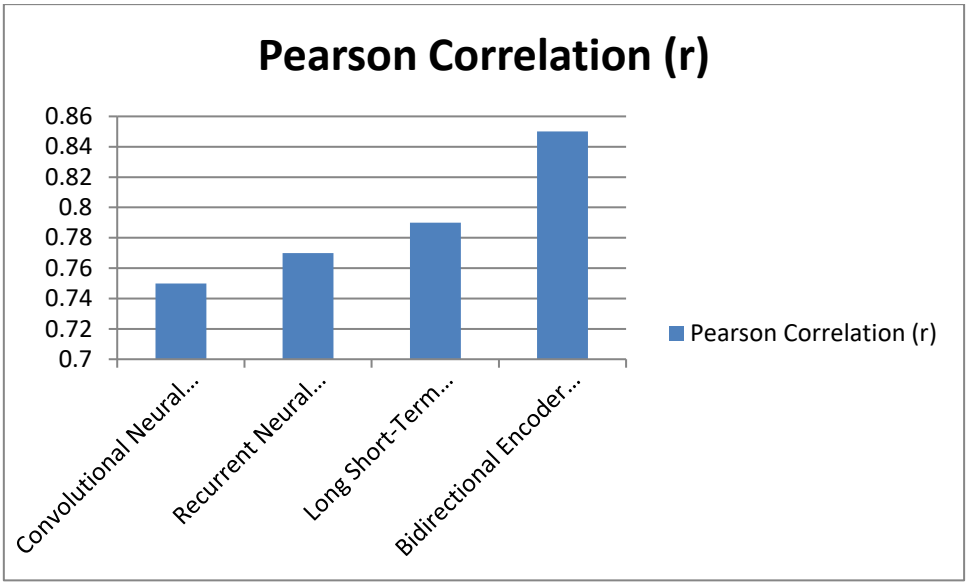

Fig-3: Graph for Pearson Correlation comparison

CONCLUSION

The comparative analysis of neural network models for Automated Essay Scoring reveals a clear progression in performance from basic to advanced architectures. CNNs and RNNs provide a solid foundation but are limited in their ability to fully capture complex text features. LSTM networks enhance performance by effectively managing long-range dependencies and providing better context, thus improving scoring accuracy and correlation with human evaluations. The BERT model stands out as the most effective, demonstrating exceptional performance in understanding and scoring essays due to its bidirectional context capture and advanced language processing capabilities. GPT also shows strong results, affirming the efficacy of Transformer-based models in AES. These results highlight the transformative impact of deep learning and Transformer architectures on essay scoring,

offering promising avenues for future research and application in educational assessment. The study concludes that incorporating advanced neural network models like BERT into AES systems can significantly enhance scoring accuracy and provide more reliable and insightful evaluations of student writing.

REFERENCES

[1] Page, Ellis B. "The imminence of... grading essays by computer." The Phi Delta Kappan *47.5 (1966): 238-243.*

[2] Foltz, Peter W., et al. "Implementation and applications of the Intelligent Essay Assessor." Handbook of Automated Essay Evaluation *(2013): 68-88.*

[3] Attali, Yigal, and Jill Burstein. "Automated essay scoring with e-rater V. 2." The Journal of Technology, Learning and Assessment *4.3 (2006).*

[4] Kaggle. "Develop an automated scoring algorithm for student-written essays." (2012).

[5] Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of Documentation *60.5 (2004): 503-520.*

[6] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." EMNLP. *Vol. 14. 2014.*

[7] Herman, Joan, and Robert Linn. "On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia. CRESST Report 823." National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (2013).

[8] De Boer, Pieter-Tjerk, et al. "A tutorial on the cross-entropy method." Annals of Operations Research *134.1 (2005): 19-67.*

[9] Davidian, David. "Feed-forward neural network." U.S. Patent No. 5,438,646. 1 Aug. 1995.

[10] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural Computation *9.8 (1997): 1735-1780.*