
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 5, pp. 4005–4012, DOI 10.12694/scpe.v25i5.3153

RESEARCH ON THE DESIGN OF A SYSTEM BASED ON MACHINE LEARNING ALGORITHMS FOR AUTOMATIC SCORING OF ENGLISH WRITING ABILITY

SHAN ZHAO*

Abstract. The development and implementation of an innovative system designed to automatically score English writing ability using advanced machine learning algorithms is challenging task. The core objective of the study is to establish a reliable and efficient method for assessing written English, which is crucial in educational and professional settings. The paper begins with an overview of the existing methods of English writing assessment, highlighting their limitations, such as time consumption and potential biases in human evaluation. The main focus of the study is the design and testing of a machine learning-based system. Various algorithms, including Natural Language Processing (NLP) techniques and neural network models, are explored and integrated to assess writing quality, grammar, coherence, and content relevance. The system's architecture is detailed, explaining how these algorithms work in tandem to evaluate and score writing. An experimental setup is described where the system is trained and validated using a large dataset of English writing samples, ranging from beginner to advanced levels. The performance of the system is measured against traditional scoring methods, with emphasis on accuracy, consistency, and the ability to handle diverse writing styles and complexities. The results demonstrate the system's proficiency in accurately scoring English writing, with a notable reduction in scoring time compared to human evaluators. The paper discusses the implications of these findings for educational institutions and language testing organizations, suggesting that this system could revolutionize how English writing is assessed.

Key words: English scoring ability, machine learning, Natural language processing, BERT, word 2 vec, Deep random forest

1. Introduction.

1.1. Context and Background. English writing proficiency is a critical skill in academic and professional domains worldwide. Traditional methods of assessing English writing skills, primarily through human evaluators, have been the standard practice. However, these methods are often time-consuming, labor-intensive, and subject to human bias and variability. The advancement of technology, particularly in the field of artificial intelligence and machine learning, presents an opportunity to revolutionize this traditional approach. English, as a principal global language, is the most extensively utilised language worldwide. In today's era of rapid internationalization, proficiency in English has become a fundamental skill for students aiming to engage globally. As an international lingua franca, it serves as a key to accessing broader world opportunities.

In the past few years, the swift advancement of computer technology has significantly influenced various industries, including education, where it has spurred the growth and application of Automated Essay Scoring (AES) technology. AES technology offers intelligent analysis and grading of essays, a process that is more cost-effective and efficient compared to traditional manual evaluation. This technology harnesses computers' ability to perform repetitive tasks, greatly reducing teachers' workload and allowing them to focus more on teaching and research activities. Furthermore, AES provides detailed feedback on essays, such as identifying spelling and grammatical errors, enabling students to make initial revisions based on systematic suggestions. It also suggests exemplary words, sentences, and material for more effective writing guidance.

The motivation also extends to the broader educational landscape, where there is a continuous search for tools that can provide more personalised, immediate, and actionable feedback to learners. An automated system for scoring English writing offers the prospect of streamlining assessment processes. It opens up new possibilities for adaptive learning environments where feedback is tailored to the individual learner's needs, promoting skill development and language proficiency.

Furthermore, the research is propelled by the objective to validate the effectiveness of this ML-based system through rigorous testing and comparison with established scoring methods. By demonstrating the system's

*Zhengzhou University of Industrial Technology, Xinzheng,451150, China (shanzhaoresearc@outlook.com)

ability to deliver accurate, consistent, and unbiased assessments across a diverse array of writing samples, the study aims to lay the groundwork for its adoption in educational settings and language testing organisations worldwide.

Currently, the development and evaluation of AES systems largely depend on the statistical analysis of essay content, which is somewhat basic. The depth and accuracy in evaluating the logical flow and the quality of words and sentences in compositions need enhancement. Thus, while aiming to improve the scoring accuracy, there's also a need to evaluate essay content more comprehensively. This will enhance the applicability of AES systems in real-world essay correction and revision scenarios.

1.2. The Problem Statement. Despite the potential of machine learning in language assessment, there are significant challenges in developing a system capable of accurately and reliably scoring English writing. Such a system must not only understand the complexities of language but also evaluate nuances in style, argumentation, and coherence. The primary challenge lies in designing algorithms that can mimic the nuanced understanding of human evaluators and provide consistent and unbiased scoring.

1.3. Research Objectives. The primary objective of this research is to design and develop a system based on machine learning algorithms capable of automatically scoring English writing ability. This involves:

1. Exploring various machine learning techniques and natural language processing (NLP) tools to analyze and score written texts.
2. Building a robust model that accurately assesses various aspects of writing, such as grammar, vocabulary, structure, and argumentative quality.
3. Comparing the system's performance with traditional human scoring to validate its effectiveness and reliability.

1.4. Significance of the Study. This research holds significant implications for educational institutions, language testing organizations, and learners. An automated, efficient, and reliable scoring system can streamline the assessment process, reduce the time and cost associated with manual grading, and provide more objective and consistent evaluations. Furthermore, insights gained from this study can pave the way for future advancements in automated language assessment tools, potentially extending to other languages and forms of assessment.

2. Literature survey. Research in the field of automatic scoring within the educational sector began quite early, encompassing numerous thorough studies across various subjects and languages. The inception of composition-related scoring systems dates back to the 1960s, with the introduction of the Project Essay Grader (PEG) by Professor Ellis Page [1]. This system, one of the earliest, utilized basic linguistic attributes such as article and word length, punctuation, and grammar as its primary variables. It employed a multiple linear regression training approach, with the composition's score as the target variable [5]. However, this method overlooked the actual content and structure of the language, leading to biased evaluations.

Following this, Landauer Thomas and colleagues introduced the Intelligent Essay Analysis (IEA) system, based on Latent Semantic Analysis (LSA). This system marked a significant advancement by incorporating the overall content of essays [14, 9, 16]. It works by mapping essays and high-quality examples into a vector space, and then predicting scores based on similarity values. Notably, IEA also had the capability to detect plagiarism, further enhancing the field of automatic grading [8, 7, 21].

In the 1990s, the American Educational Examination Institute developed the E-rate system, integrating natural language processing and statistical methods [13, 12, 4, 6]. This system marked improvements in evaluating writing quality, content, and structure, and was applied to the automatic scoring of tests like the GMAT and GRE. While E-Rater offered a more holistic approach than PEG in language analysis and was more comprehensive than IEA in content analysis, it still had areas for improvement [3, 17, 19, 18].

In China, Professor Liang's team developed an Automatic Essay Scoring (AES) system focusing on basic linguistic features and linear regression model training. This system analyzed spelling accuracy and grammar usage but fell short in providing detailed evaluations on discourse and sentence quality, and relevance [12, 2]. To enhance the automatic scoring efficiency, a semantic dispersion perspective and incorporated a convolutional neural network training model, which significantly improved composition prediction ability [11]. Qiu's research

involved evaluating composition fluency and integrating it into the AES model to enhance scoring effectiveness [14]. Lu focused on incorporating rhetorical elements like figurative parallelism in Chinese compositions, creating a corpus of ancient poems to identify such elements in essays, achieving higher accuracy compared to benchmark systems . Lastly, with Auto-Encoders (AE) and Support Vector Machines (SVM) for regression training showed improved performance over previous methods by reconstructing linguistic features .

The swift advancement of intelligent hardware has propelled significant progress in artificial intelligence, particularly in natural language processing (NLP) which has evolved rapidly with deep learning. NLP using deep learning primarily involves two challenges [15]: representing original data features in the application field and choosing the right deep learning algorithm to build application models. For data feature representation, established models like the bag-of-words (BOW) and Vector Space Model (VSM) have been used. However, these methods have limitations. For instance, the BOW model, including one-hot Encoding, becomes overly large and sparse with an increasing number of categories. Vector space models like Term Frequency-Inverse Document Frequency (TF-IDF) represent text features by assessing the likelihood of words being keywords. Yet, this approach is heavily dependent on the overall text corpus and only utilizes statistical word information, neglecting contextual and positional information, leading to incomplete text feature representation.

Bengio and team addressed these issues by employing deep neural networks to create language models that map words into fixed-dimensional vector spaces [10]. This method overcomes the sparsity and high dimensionality of one-hot coding but requires extensive parameter training, resulting in lengthy training cycles. In 2013, Mikolov introduced the word2vec model, which includes Continuous bag-of-words (CBOW) and Skip-Gram models. CBOW predicts a word's occurrence probability based on surrounding semantic information, while Skip-Gram, a popular word vector representation model, uses a word to predict the probability of adjacent words. Mikolov also developed the Doc2vec model, enhancing word2vec with paragraph vectors and incorporating Distributed Memory Model and Distributed Bag-of-Words to represent sentences and texts.

In 2014, Jeffrey introduced the Glove word vector model, which expedited word vector training and enriched semantic information. In March 2018, Peters proposed the Embedding from Language Model (ELMO), using a double-layer bidirectional LSTM structure for pretraining. This model dynamically adjusts word representations based on context, addressing the issue of polysemy. Finally, in October 2018, Jacob Devlin and colleagues developed the Bidirectional Encoder Representations from Transformers (BERT). Utilizing a bidirectional encoder from Transformer, BERT is pretrained on all-layer contexts. Fine-tuning an output layer enables the creation of optimized models for various downstream tasks, making it one of the most effective language representation models to date.

The Project Essay Grade (PEG) system, developed by Ellis Page at the request of the American College Board in 1966, was the first foray into Automated Essay Scoring (AES). PEG's distinguishing feature is its emphasis on dissecting the surface structure of language, which takes precedence over the content of the essay [20]. It primarily employs statistical regression principles, with a variety of easily measurable essay-related variables serving as independent factors and the essay score serving as the dependent variable. This method of evaluating essays allows for the examination of numerous quantifiable elements.

Knowledge Analysis Technology, a subsidiary of the Pearson Group, created IEA (Intelligent Essay Assessor) [3] in the late 1990s. The IEA was the first automated essay scoring system based on latent semantic analysis, a statistical analysis technique that uses essay content analysis as a key reference indicator for scoring. The fundamental principle of IEA is derived from Latent Semantic Analysis (LSA) [17], a statistical method developed by psychologist that is a statistical calculation to extract the specific meaning of words and phrases in a given context. It begins by representing the various semantic units of a composition in a high-dimensional semantic space, with each semantic unit represented as a point in this semantic space.

3. Proposed methodology. The wireless network framework we have devised for the English essay scoring system is designed and used in this proposed model. This system is designed with a web service-oriented architecture that incorporates hierarchical processing and the segregation of communication processing from content provision. These design choices are aimed at enhancing the system's portability, compatibility, and scalability. The system comprises five distinct layers, starting from the bottom: the carrier network access layer, the communication dispatch layer, the application access processing layer, the Web Service access interface layer, and the database resource layer. The carrier network access layer pertains to the underlying network

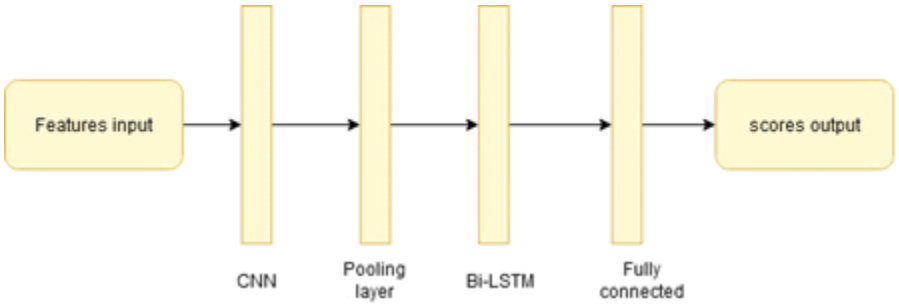

Fig. 3.1: Architecture of proposed model

Table 3.1: Dataset of Testing and Training details

Category	**Training set**	**Test set**
Type of learners in mother tongue	16	14
Type of essay's topic	27	5
Number of essays	1142	96
Number of essay words	878766	75551
Average scores	27.82	27.47
Lowest scores	0	13
Median scores	28	26
Highest scores	40.5	40
Standard deviation	5.5	5.96

infrastructure essential for system data communication, encompassing wireless communication networks like GSM and CDMA. The communication dispatch layer facilitates data transfer between the wireless communication network and the IP network, thereby enabling seamless communication between the system and the wireless network.

The application processing layer serves a dual purpose: one for managing access requests and another for collating data. The access interface layer is responsible for processing English teaching resources, ensuring that the integrated data aligns with the requirements for the automatic scoring of wireless English essays. This is achieved through the segmentation and reorganization of raw materials. Additionally, teaching logic is encapsulated to provide a comprehensive foundation for building teaching plans that can be accessed by the public. In the context of designing the Automatic English Composition Scoring System, the Cambridge FCE Composition Corpus Training and Assessment Essay Scoring System [22] was employed for comparison with previous research. Figure 3.1 shows the proposed architecture in detail.

3.1. Dataset. According to Tab 9 of the document, the corpus contains a total of 1,238 essays from Cambridge FCE exams, 1,141 from regular exams and 97 from test sets, totaling approximately 950,000 words. Manual correction was used to evaluate and score each essay. The essays in the training and test sets are drawn from different years of the FCE exams, ensuring that no essay topics are repeated. 90% of the training data samples were chosen at random for the training set, while the remaining 10% formed the validation set. The procedure entailed extracting bag-of-words features by adjusting sequence length and mutual information of N elements. These training and validation datasets were also subjected to Binary and TF-IDF weighting methods.

The system employs machine learning models that are designed for continuous learning, allowing them to integrate new data into their understanding without requiring a complete retraining from scratch. Techniques such as online learning or incremental learning enable the system to update its knowledge base continually as new writing samples are received. To accommodate evolving language use and emerging linguistic trends,

the system periodically revisits and updates its evaluation criteria. This involves retraining the models on a combination of original and newly acquired data. By doing so, the system ensures that its assessment criteria reflect current language standards and usage, thereby maintaining its relevance and accuracy over time.

3.2. GLoVE word Embedding. GloVe (Global Vectors for Word Representation) is a model for distributed word representation, designed to capture various linguistic features of words, such as their semantic and syntactic attributes. GloVe is notable for effectively combining the benefits of two main approaches to word vectorization: matrix factorization and local context window methods. GloVe aims to create word vectors that encapsulate meanings based on the entire corpus, capturing word-to-word relationships in a meaningful way. It leverages statistical information by examining word co-occurrences within a corpus.

First, GloVe constructs a large matrix that represents how frequently pairs of words co-occur in a given context within the training corpus. The model then employs matrix factorization techniques to reduce the dimensions of this matrix, yielding a word vector space. Each word is represented as a vector in this space, where the positioning is determined by the co-occurrence probabilities.

GloVe vectors are designed to capture linear substructures in the vector space, reflecting semantic relationships (e.g., man-woman, king-queen). They can easily handle large-scale corpora efficiently. Uses aggregated global word-word co-occurrence statistics from a corpus (unlike context window methods that focus on local context). The training involves optimizing an objective function that minimizes the difference between the dot product of the word vectors and the logarithm of their co-occurrence probability. The process is unsupervised, requiring only a text corpus.

3.3. LDA Feature Extraction. Feature Extraction Using Latent Dirichlet Allocation. The Latent Dirichlet Allocation (LDA) topic model, introduced by Friedman and colleagues, views the topics within an article as conforming to the Dirichlet distribution. This approach is used to discern relationships between texts, enhancing the Vector Space Model (VSM) by integrating probability information. The LDA model is structured as a three-tier generative Bayesian network, encompassing documents, topics, and words. Its core probabilistic computation is illustrated in Formula (3.1).

$$p(wi \mid dj) \;=\; \Sigma k \; s = 1 \; p(wi \mid z \;=\; s) p(z \;=\; s \mid dj) \tag{3.1}$$

Here, $p(wi \mid z = s)$ denotes the likelihood of the word wi being associated with topic s, and $p(z = s \mid dj)$ signifies the probability of topic s in the specific short text dj. Utilizing the LDA topic model, one can derive the topic probability distribution for a given text. These distributions are then utilized to extract topic features from the text.

3.4. Sentence Recognition. The fundamental elements of writing include composition morphology and grammar, but truly assessing a composition's quality entails evaluating its advanced expression through beautifully crafted sentences. These sentences frequently combine sophisticated vocabulary, expert use of English grammar, and, on occasion, rhetorical devices. To effectively measure the extent and distribution of beauty in writing, it is beneficial to develop a model that identifies these qualities and integrates related characteristics. Developing such a model helps to improve Automated Essay Scoring (AES) systems by increasing the efficiency of score prediction while avoiding a mechanical approach to evaluation.

Sentence elegance recognition is a type of text classification. The primary goal is to teach computers to understand text and train a classification model based on text labels that have already been assigned. As a result, new input texts are classified. Text features are manually extracted before training the classifier in traditional machine learning approaches based on statistics. Manually identifying and creating perfect features that capture the nuanced beauty of language, on the other hand, is difficult. Deep learning methods, on the other hand, excel at capturing text characteristics by automatically selecting and combining features. Traditional statistical and rule-based methods rely on manually created sentence features, which frequently fail to capture the essence of well-constructed sentences, particularly those containing advanced grammar or stylistic devices such as metaphors and personification. In comparison, neural network models can learn semantic vectors from large amounts of data on their own, effectively representing sentence features in binary classification tasks.

The Convolutional Neural Network (CNN) is a widely used type of artificial neural network. It employs convolutional kernels to capture local information, which is then synthesized into global information via the

pooling layer. The core architecture of a CNN includes an input layer, convolutional layers, and pooling layers. In automatic essay scoring tasks, the input layer typically consists of a text representation matrix formed by word vectors. The convolutional layer allows for the setting of kernels of various sizes, enabling the capture of certain contextual and sequential information. One of the key advantages of CNNs over traditional neural network models is the introduction of weight sharing, which simplifies the network's complexity and accelerates training. In the pooling layer, a segment-wise maximum pooling approach is used to preserve the relative location of multiple local maximum values. This method is also capable of detecting the intensity of features if strong characteristics are repeated. However, it's important to note that while this approach retains coarse position information, absolute position details are lost. After the convolution and pooling processes, a representation of the sentence level is achieved.

One of the most significant benefits is the system's ability to provide immediate, personalized feedback to students on their writing. This instant feedback loop can significantly enhance the learning process, allowing students to identify and correct errors, refine their writing style, and better understand the criteria for high-quality writing. Immediate feedback is particularly valuable in large classrooms or distance learning scenarios where individualized attention from instructors may be limited. With the system taking on the task of assessing basic grammar, spelling, and syntax, educators can devote more time and resources to teaching higher-level writing skills. These include argumentation, critical thinking, and creative expression. Teachers can focus on developing students' abilities to construct well-organized, coherent, and persuasive texts, rather than spending excessive time marking mechanical errors.

Proposed BiLSTM -CNN model. Creating a model that combines Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) entails creating a system that takes advantage of the advantages of both architectures. BiLSTM excels at understanding context and dependencies in sequential data, such as text, whereas CNNs excel at extracting features, in this case, sentence structures, from data. A step-by-step procedure for creating such a model:

Convolutional Layer. Firstly, Identify features in the word vector matrix. Then it generates three different types of convolution kernels (e.g., 3x128, 4x128, and 5x128), each with 50 kernels. These kernels will aid in the extraction of various local features from word vectors. Finally, use these kernels to extract features from the word vector matrix. To introduce nonlinearity and accelerate convergence, use a ReLU activation function for each neuron.

Pooling Layer. To distill the features extracted by the convolution layer and to reduce the dimensionality of the feature space. Divide each feature map into chunks (e.g., three parts), and apply max pooling to each chunk. This approach helps preserve the relative order information and capture the strongest features.

Bi-LSTM Layer. After the pooling layer, the output is fed into a BiLSTM layer. It analyzes the sequence data (features extracted and pooled from the CNN) in both forward and backward directions. This is crucial for understanding the context and dependencies in the text data. The BiLSTM processes the sequence of features, capturing information from both past and future contexts.

Fully Connected and Output Layers. The output is Flatten from the Bi-LSTM layer to create a one-dimensional vector. Add two dense layers to allow the model to learn non-linear combinations of features in fully connected layer. The sigmoid activation function used in the output layer for binary classification tasks (like sentiment analysis) or softmax for multi-class classification.

Model Training and Optimization. An appropriate loss function (like cross-entropy) is choosen and an optimizer stochastic gradient descent is used. Model is trained using backpropagation and adjust the weights iteratively. finally, implement dropout or L2 regularization to prevent overfitting.

The graph 2 representing the CNN model is the shortest, suggesting that it has the lowest accuracy among the three models presented. The accuracy percentage is approximately 85%, which indicates that while the CNN model is relatively accurate, there may be room for improvement in feature analysis tasks. The CNN-LSTM model, is significantly greater than the first, implying a noticeable increase in accuracy. The model's accuracy is around 90%, showing that combining CNN features with LSTM, which can capture sequential information, offers a substantial improvement over the plain CNN model. The Proposed CNN-BiLSTM model is the highest of all, indicating the highest accuracy among the three models on feature analysis. The accuracy is just under 91.6. The bidirectional LSTM allows the model to access information from both past and future

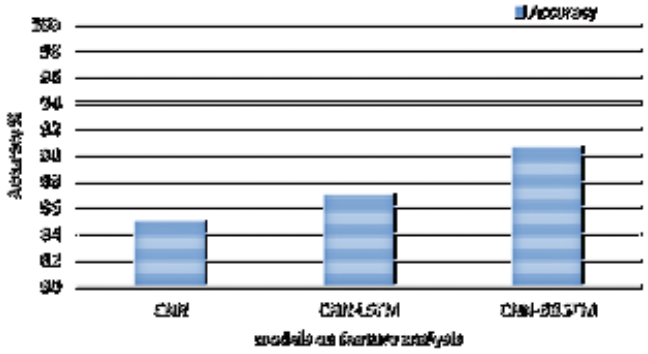

Fig. 3.2: The accuracy of feature analysis

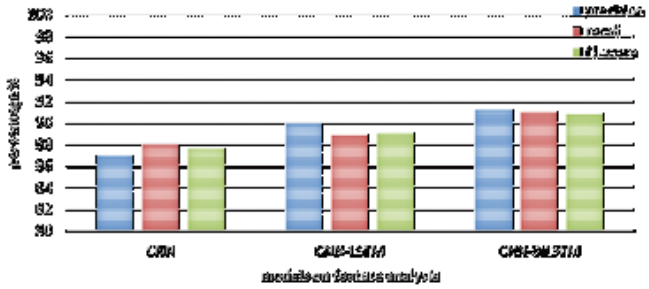

Fig. 3.3: The performance comparison of feature analysis model

states, potentially giving it an advantage in understanding the context within data, leading to higher accuracy

Convolutional neural networks (CNNs) have a strong capability for extracting intricate features from sentences. To investigate this, experiments were conducted with three distinct approaches to feature extraction: manual feature engineering, CNN-based feature extraction, and a hybrid method combining manual and CNN. When examining the trends in accuracy and recall as illustrated in Figure 3.3, the performance metrics for all three methods were broadly comparable. However, it's notable that manual feature engineering CNN alone resulted in the lowest accuracy, specifically at 87%, . Conversely, the CNN-based approach yielded the lowest recall rates in feature classification, and consistently, both accuracy and recall were lower for CNN and CNN-LSTM. This suggests an inherent challenge within the algorithm's ability to discern sentences, and highlights a discrepancy with human perception in real-world applications.

The data indicates that BiLSTM and CNN-based feature extraction methods significantly enhance the differentiation between the sentences. Additionally, it's observed that the performance metrics for sentences are consistently lower across all categories, underscoring the complexity of assessing the aesthetic quality of sentences—a factor that is also reflective of an individual's writing skill level.

In pursuit of refining the experiment, consideration was given not only to the blend of feature extraction techniques from machine learning but also to the analysis of different network classifications based on various feature combinations. This included looking at linguistic and semantic feature integration, as well as the incorporation of difficult features. The results, as evident in Figure 3.3, features outperformed using text-CNN and BiLSTM models. When comparing models with equivalent feature types, the LSTM model surpassed the text-CNN in performance, demonstrating its superior capability for memory learning in text-mining applications. Finally, an enhanced version of LSTM, known as Bi-LSTM, achieved the best results in the second set of experiments. This improvement is attributed to Bi-LSTM's adeptness in capturing temporal dependencies from different directions, thereby obtaining more temporally relevant sentence features.

Continuous evaluation and benchmarking against industry standards and datasets ensure that the system's

performance meets the expected criteria for accuracy, fairness, and reliability. These evaluations guide further refinements and adjustments to the system.

4. Conclusion. The research highlights an approach to automating the assessment of English writing proficiency using cutting-edge machine learning algorithms. Addressing the inefficiencies and biases inherent in traditional evaluation methods, this research outlines the development of an intelligent system that employs Natural Language Processing and neural network models to deliver swift, consistent, and objective analysis of written English. The system's architecture, which harnesses a synergy of algorithms to evaluate various aspects of writing, is rigorously tested against a vast corpus of English samples. The findings are clear: the proposed machine learning-based system not only rivals but also potentially surpasses human raters in terms of accuracy and speed, marking a significant advancement in the field of language assessment. This breakthrough holds considerable promise for educational and professional domains, offering a scalable, reliable alternative that could fundamentally transform the assessment landscape of English writing ability.

REFERENCES

[1] S. Bonthu, S. Rama Sree, and M. Krishna Prasad, *Automated short answer grading using deep learning: A survey*, in Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5, Springer, 2021, pp. 61–78.

[2] R. G. Garroppo, M. G. Scutellà, and F. d'Andreagiovanni, *Robust green wireless local area networks: A matheuristic approach*, Journal of Network and Computer Applications, 163 (2020), p. 102657.

[3] Y. Guo, *A study of english informative teaching strategies based on deep learning*, Journal of Mathematics, 2021 (2021), pp. 1–8.

[4] V. Kumar and D. Boulanger, *Explainable automated essay scoring: Deep learning really has pedagogical value*, in Frontiers in education, vol. 5, Frontiers Media SA, 2020, p. 572367.

[5] V. S. Kumar and D. Boulanger, *Automated essay scoring and the deep learning black box: How are rubric scores determined?*, International Journal of Artificial Intelligence in Education, 31 (2021), pp. 538–584.

[6] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, *A deep learning approach to assessing non-native pronunciation of english using phone distances*, ISCA, 2018.

[7] Y. Li, *Deep learning-based correlation analysis between the evaluation score of english teaching quality and the knowledge points*, Computational Intelligence and Neuroscience, 2022 (2022).

[8] Y. Liu and R. Li, *Deep learning scoring model in the evaluation of oral english teaching*, Computational Intelligence and Neuroscience, 2022 (2022).

[9] C. Lu and M. Cutumisu, *Integrating deep learning into an automated feedback generation system for automated essay scoring.*, International Educational Data Mining Society, (2021).

[10] X. Lu and R. Hu, *Sense-aware lexical sophistication indices and their relationship to second language writing quality*, Behavior research methods, 54 (2022), pp. 1444–1460.

[11] O. Lyashevskaya, I. Panteleeva, and O. Vinogradova, *Automated assessment of learner text complexity*, Assessing writing, 49 (2021), p. 100529.

[12] H. Meisheri, R. Saha, P. Sinha, and L. Dey, *Textmining at emoint-2017: A deep learning approach to sentiment intensity scoring of english tweets*, in Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017, pp. 193–199.

[13] F. Qin, *College english intelligent writing score system based on big data analysis and deep learning algorithm*, Journal of Database Management (JDM), 33 (2022), pp. 1–26.

[14] D. Ramesh and S. K. Sanampudi, *An automated essay scoring systems: a systematic literature review*, Artificial Intelligence Review, 55 (2022), pp. 2495–2527.

[15] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, *Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring*, arXiv preprint arXiv:2008.01441, (2020).

[16] J. Sawatzki, T. Schlippe, and M. Benner-Wickner, *Deep learning techniques for automatic short answer grading: Predicting scores for english and german answers*, in International Conference on Artificial Intelligence in Education Technology, Springer, 2021, pp. 65–75.

[17] S. Thara and P. Poornachandran, *Social media text analytics of malayalam–english code-mixed using deep learning*, Journal of big Data, 9 (2022), p. 45.

[18] M. Uto, *A review of deep-neural automated essay scoring models*, Behaviormetrika, 48 (2021), pp. 459–484.

[19] Z. Wang, H. Huang, L. Cui, J. Chen, J. An, H. Duan, H. Ge, N. Deng, et al., *Using natural language processing techniques to provide personalized educational materials for chronic disease patients in china: development and assessment of a knowledge-based health recommender system*, JMIR medical informatics, 8 (2020), p. e17642.

[20] T. Xia and X. Chen, *A weighted feature enhanced hidden markov model for spam sms filtering*, Neurocomputing, 444 (2021), pp. 48–58.

[21] S. Yuan, T. He, H. Huang, R. Hou, and M. Wang, *Automated chinese essay scoring based on deep learning*, CMC-Computers Materials & Continua, 65 (2020), pp. 817–833.

[22] Z. Yuan, *Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm.(retraction of vol 40, pg 2069, 2020)*, 2021.

Edited by: Rajanikanth Aluvalu
Special issue on: Evolutionary Computing for AI-Driven Security and Privacy: Advancing the state-of-the-art applications
Received: Jan 6, 2024
Accepted: Feb 9, 2024