International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET229257

Essay Scoring Model Based on Gated Recurrent Unit Technique

Eluwa J.*, Kuyoro S., Awodele O., Ajayi A.
Department of Computer Science, Babcock University, Ilishan-Remo, Ogun State, Nigeria

Article Info
Volume 9, Issue 2
Page Number : 323-330

Publication Issue :
March-April-2022

Article History
Accepted : 15 April 2022
Published: 30 April 2022

ABSTRACT

Educational evaluation is a major factor in determining students' learning aptitude and academic performance. The scoring technique that relies solely on human labour is time consuming, costly, and logistically challenging as this rating is usually based on the opinion of "biased" human. Several studies have considered using machine learning techniques with feature extraction based on Term Frequency (TF) - Part of Speech (POS) Tagging without consideration to global vectorization (GloVe). These solutions require the process of selecting deterministic features that are directly related to essay quality which is time-consuming and needs a great deal of linguistic knowledge. Gated Recurrent Unit (a variation of Recurrent Neural Network) deep learning technique with focus on morphological analysis of essays for content-based assessment has therefore shown the capability of addressing the challenges posed by other AES techniques by building more abstract and complete linkages among features.

Deep learning algorithms such as Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) were used to learn the model with performance evaluation on metrics such as validation accuracy, training time, loss function, and Quadratic Weighted Kappa. The performance results showed that MLP, LSTM and GRU had average Quadratic Weighted Kappa (QWK) values of 0.65, 0.86 and 0.88 respectively with each algorithm having an average training time of 61.4, 62.68 and 67.86 seconds respectively. The loss functions for MLP, LSTM and GRU were 0.296, 0.24 and 0.126. This meant that GRU had the best estimate of the difference between the actual and forecasted scores. MLP, LSTM, and GRU had average validation accuracy of 0.48, 0.537, and 0.511 respectively. GRU was shown to be the optimal classifier and was used in the development of the essay scoring model.

Keywords : Deep Neural Network (DNN), Global Vectorization (GloVe), Hyper Text Mark-up Language (HTML), Machine Learning (ML), Natural Language Processing (NLP).

I. INTRODUCTION

Deep Learning (DL) is based on the structure and capabilities of an Artificial Neural Network ANN), which is a type of human neuron. ANN outperforms most other machine learning techniques because of its ability to use supervised, semi-supervised, and unsupervised learning on a variety of data [1]. Deep learning algorithms gradually learn from high-level features and they are better suited to issues involving large volumes of data.

AES stands for "automated essay scoring," which is "software that can predictably assess essays using a pre-trained computer model" [2]. The assessment is done through a computer system that assigns scores to a student's response built on a set of qualities or characteristics. Features are important to train AES models, especially in machine learning and neural networks [3]. Automated essay scoring (AES) systems has gotten increased attention to alleviate the strain of scoring.

Educational evaluation is a major factor in determining a student's learning aptitude and academic performance [3]. Essay being one of the most important factors used by teachers in evaluating student's intelligence and learning requires the examiner to put in a great amount of effort into scoring because of its subjective character [4]. The subjectivity in essay assessment leads to variation in grades and bias because of human factors. Currently, assessment is most done manually the limitation of which include weariness, interference, and discrepancy of scoring over time. To resolve this challenge, assessment is done through automation which expedites the process and reduces human rater efforts in scoring essays as close to human’s decision [2].

However, the simple programming languages and techniques are not adequate for essay evaluation as more responses are expected from the students with various answers and all the answers must be evaluated. As a result, more complex processes, and techniques, such as machine learning, natural language processing and deep learning, are becoming increasingly important [3].

Despite the inherent advantages of traditional AES, there exist many challenges that must be solved before it can be extensively deployed in the field. Such challenges are frequently caused by difficulty in picking adequate features and determining the scoring systems' interpretability. As a result, the focus of this study is to develop an essay scoring model, capable of reducing bias in human raters, improving objectivity and performance; and ensuring rater consistency, by building more abstract and complete relationships among features using Gated Recurrent Unit (GRU) deep learning technique.

II. LITERATURE REVIEW

The foremost AES system, Project Essay Grader (PEG), gathers linguistic features [5] from previously graded essays and selects the relevant weighted features using a multiple linear regression technique [6]. The system takes a statistical approach by focusing solely on the writing style characteristic rather than content or text semantics. The Intelligent Essay Assessors (IEA) were created to address the PEG's weakness. It is built on the foundation of Latent Semantic Analysis (LSA), which considers the essays' content, style, and mechanics with no consideration for word order [7].

In quick succession, the Educational Testing Service (ETS 1) was created to handle small sentences by building domain-specific and concept-based lexicon from training data with NLP and Information Retrieval (IR) algorithms used to extract relevant features from the texts. E-Rater (Electronic Essay Rater), was created utilizing NLP and statistical techniques based on style and content [8]. The Conceptual Rater (C-Rater) was created using NLP to assess the accuracy of essay content [8]. It is not necessary to submit many previously graded essays; rather, a single assessed response is sufficient. It is utilized in reading comprehension and algebra with other rating models [4]. Based on the example of E-rater and C-Rater, many tools, such as Bayesian Essay Test Scoring System (BETSY), Automark, Intelligent Essay Marking System (IEMS), and Schema Extract Analyze and Report (SEAR), can award scores based on both style and content using a variety of techniques such as statistical approaches, NLP, rule based expert systems, and so on [4].

In 2016, [9] employed convolutional neural network (CNN) with word embedding to automatically learn features for in domain and cross domain adaptations. Essay scoring task was considered as regression task with a two-layer CNN model. One convolutional layer was used to extract sentences representations, while the other was stacked on sentence vectors to learn essays representations. A model based on RNN was developed by [10] without any feature engineering, to learn the relationship between an essay and its assigned score. For the goal of automated essay scoring, many neural network models such as CNN, RNN, GRU, CNN + LSTM and LSTM were investigated to gain some insights into the models. A study by [11] utilized Bi-directional Long Short-Term Memory (LSTM) and Hierarchical Attention Network (HAN) using Word2Vector to represent each work as word embedding and Skip-gram model to transfer words into their vector forms. Bi-LSTM to analyse the extraction of semantic relations between the word vectors. The essays are converted to a list of word embedding fed into the word embedding (input) layer as an alternative approach to LSA, which is used to represent a word by a vector of real numbers. To capture the weighted vectors calculated by the proportion of words in the essay, semantic analysis was used to harness a bi-LSTM network with the attention process.

In 2018, [12] employed a hierarchical recurrent neural network paired with an attention mechanism on the sentence and document levels using two layers of Bi-LSTM to learn the content representation of the essay and the topic while the relevance of each word in a sentence and each sentence in a document is learned via the attention mechanism. A study by [13] proposed a qualitatively upgraded deep convolution recurrent neural network to compute the quality of a piece. The model not only uses pre-trained word or sentence representations of text, but it also considers qualitatively enhanced properties including lexical variety, in formativeness, coherence, well-formedness. A symmetrical neural network AES model that can accept the input pair was proposed by [14]. The Siamese Bidirectional Long Short-Term Memory Architecture (SBLSTMA) model can capture not only the semantic aspects of the essay, but also the information about the rating standards. A hybrid model for scoring summaries that combines state-of-the-art recurrent neural networks with textual complexity indexes was developed by [15]. The study utilized the Amazon Mechanical internet research service where 636 summaries for 30 texts were collected. Two trained researchers graded the corpus summaries on two primary dimensions: main ideas and main idea correctness. The network is fed the summary and the original text, which are represented by pretrained Glove word embeddings of size 100, with non-vocabulary words ignored. To exchange network weights for the summary and the entire text, a Bi-GRU Siamese architecture was utilized. The forward-backward concatenated outputs from each cell are max-pooled.

A study by [16] suggested an automatic essay scoring approach based on a mix of CNN and Ordinal Regression (OR) for the characteristics of automatic scoring mechanisms. The loss function of this experiment is selected using the Adam optimization approach, and the output data is OR processed to construct an automatic scoring model. In 2019, [17] developed a scoring model based on feature scoring and RNN using training dataset from Kaggle. Data preparation was carried out to remove unwanted data and avoid problems during runtime. In order to improve human grader subjectivity which may be incorporated unintentionally into the essay scores during training, [18] developed a model where essays were partitioned into subsets that were a representation of similar graders' essays using explanation approach and clustering. PCA was used to reduce the dimensionality of essay attributes, and a k-means clustering approach was employed to group relevant writings together. The scoring models were created using random trees and SVM, and scores were assigned to the second level cluster at random. A fully automatic essay grading model based on a google word2vec two-layer BLSTM model was implemented by [19]. The model adopted a two-layer LSTM architecture, with the first layer learning fundamental features and the second layer learning more abstract features at a higher level utilizing google word2Vec for feature extraction. Finally, BLSTM combines the forward hidden layer and the backward hidden layer, allowing it to abstract both previous and subsequent contexts.

A study by [20] proposed a DNN-AES architecture that incorporates Item Response Theory (IRT) models to address rater bias in training data. A CNN-LSTM-based model with Lookup table layer to convert each word in an essay into a D dimensional word-embedding representation, where similar words with related meanings have similar representations. The CNN layer retrieves n-gram level characteristics from a sequence of word embedding vectors. Experimental dataset was Automated Student Assessment Prize (ASAP) extracted from Kaggle. A bespoke multi-model neural network using Keras Functional API, which consists of a two-layer neural network for processing the numerical representation of the essay and a word vector neural network for processing the sequence from Keras tokenizer was constructed by [21]. The tokenized sequences are then evaluated using an LSTM neural network, and the vector representation is evaluated using a 2-layer neural network. The results were combined, and the final score is predicted using a 2-layer neural network.

III. METHODS AND MATERIAL

A comprehensive examination of literature was conducted to characterize and determine the features according to the content-based assessment utilizing the morphological classification technique [22]. This technique takes into cognizance the inflectional and derivational forms of language at the pre-processing phase. Previous studies on essay scoring system were examined to characterize the features in determining the variables, which represent features used in existing scoring process. Then the features were characterized into rubrics based on the method, technique, tools and concepts.

Secondary dataset was collected from Kaggle provided by the Hewlett Foundation to aid semantic exploration and Part of Speech (POS) tagging. The total essay collected was split into 5970 training dataset and 1493 testing dataset and saved as pickle file.

Also, an assemblage of Computer Science questions and answers were collected to create a more robust dataset in order to ensure high reliability. The research utilized three deep learning techniques: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), all of which were thoroughly investigated and examined to determine which strategy was best for constructing a model for this study. Each of the deep learning algorithms were evaluated for objectivity in the scoring process using the selected metrics such as validation accuracy, epoch on training time, loss function, and Quadratic Weighted Kappa.

The Essay Scoring model was built in a web-format using Python, Scikit-Learn, Flask, Cascading Styling Sheet (CSS) and HTML using the best classifier with SQLite as the database management system.

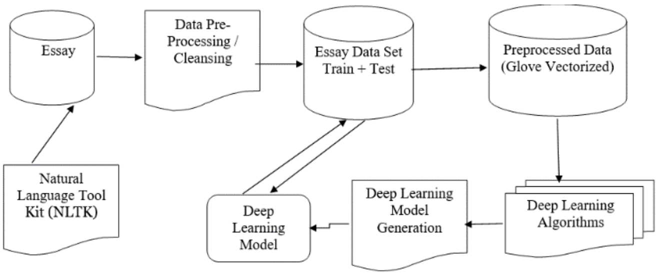
Essay
Data Pre-Processing / Cleansing
Essay Data Set Train + Test
Preprocessed Data (Glove Vectorized)
Natural Language Tool Kit (NLTK)
Deep Learning Model
Deep Learning Model Generation
Deep Learning Algorithms

Figure 1: Conceptual framework of developed model

IV. RESULTS AND DISCUSSION

The performance results showed that MLP, LSTM and GRU had average Quadratic Weighted Kappa values of 0.65, 0.86 and 0.88 respectively with each algorithm having an average training time of 61.4, 62.68 and 67.86 seconds respectively. The loss functions for MLP, LSTM and GRU were 0.296, 0.24 and 0.126. This meant that GRU had the best estimate of the difference between the actual and forecasted scores. MLP, LSTM, and GRU had average validation accuracy of 0.48, 0.537, and 0.511 respectively, while using an epoch of 1 to 20. GRU was shown to be the optimal classifier in the development of the essay scoring model.

Table 1: Evaluation Based on Loss Function

Algorithm	Essay Set 1	Essay Set 3	Essay Set 4	Essay Set 5	Essay Set 6
MLP	0.25	0.5	0.5	0.2	0.23
LSTM	0.32	0.25	0.33	0.2	0.1
GRU	0.21	0.2	0.1	0.02	0.1

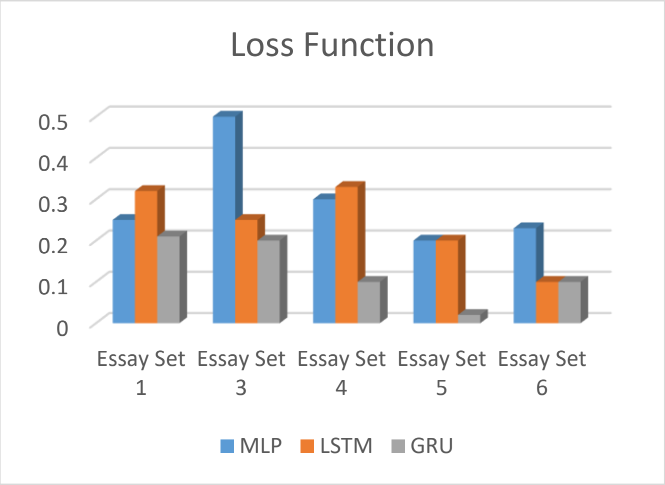

Figure 2 : Loss Function evaluation graph

Table 2: Evaluation Based on Validation Accuracy

Algorithm	Essay Set 1	Essay Set 3	Essay Set 4	Essay Set 5	Essay Set 6
MLP	0.60	0.50	0.40	0.4	0.50
LSTM	0.80	0.42	0.50	0.47	0.48
GRU	0.70	0.50	0.48	0.47	0.7

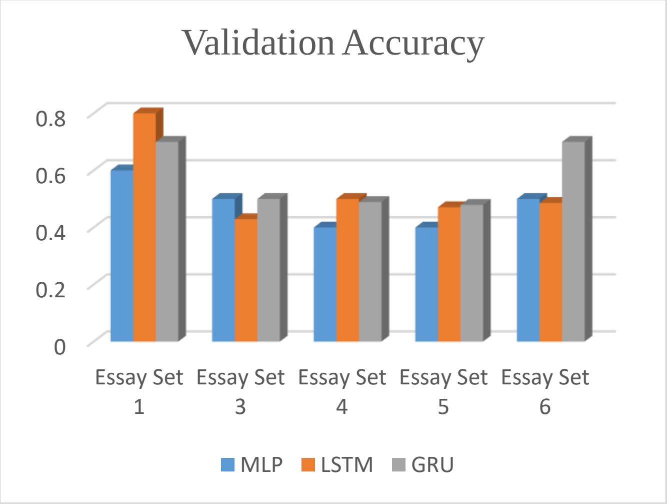

Figure 3: Validation Accuracy evaluation graph

Table 3: Evaluation based on Training Time

Algorithm	Essay Set 1	Essay Set 3	Essay Set 4	Essay Set 5	Essay Set 6
MLP	170.3	25.5	11.7	66.4	33.1
LSTM	157	35	30	39.1	52.3
GRU	177.3	31.4	33.1	42.1	55.4

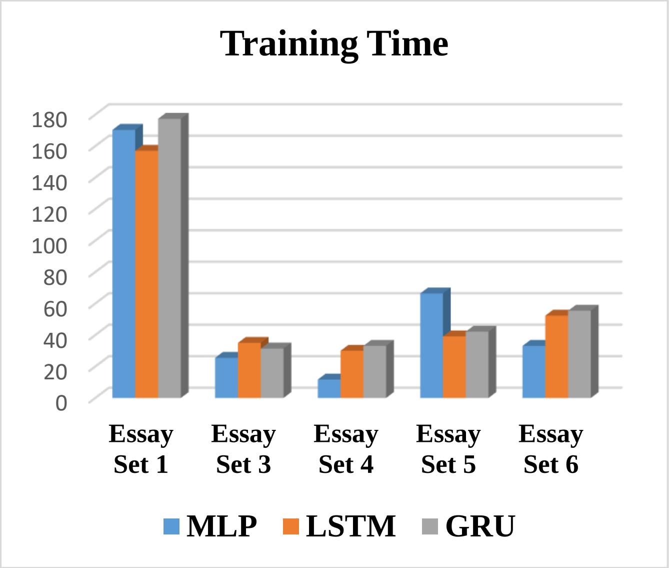

Figure 3: Training time evaluation graph

Table 4: Evaluation based on QWK

Algorithm	Essay set 1	Essay set 3	Essay set 4	Essay set 5	Essay set 6
MLP	0.75	0.64	0.63	0.66	0.65
LSTM	0.80	0.82	0.85	0.84	0.86
GRU	0.82	0.80	0.87	0.86	0.88

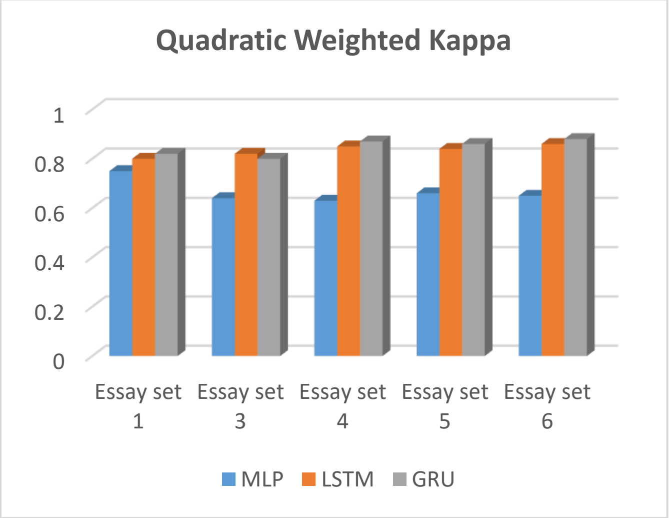

Figure 5: QWK evaluation graph

Table 5: Comparative Analysis Evaluation

Algorithm	Average Loss Function	Average Validity Accuracy	Average QWK (%)	Average Training Time (s)
MLP	29.6	48	65	61.4
LSTM	24	53.7	86	62.68
GRU	12.6	51.1	88	67.68

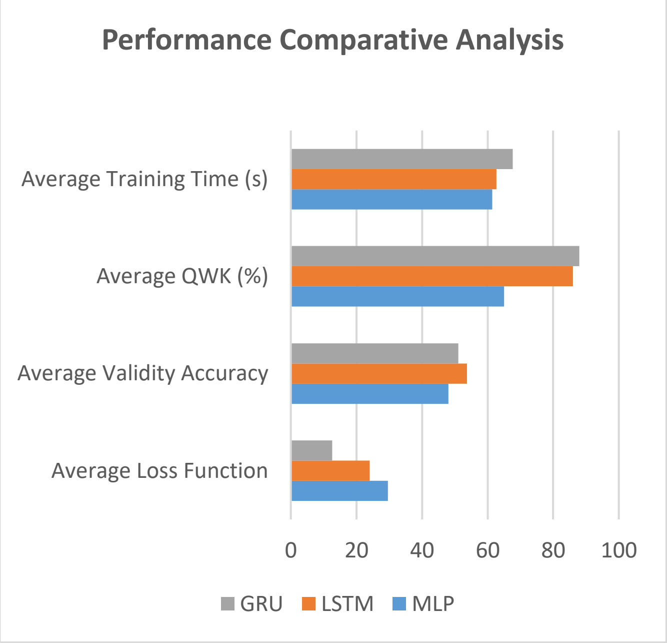

Figure 6: Performance Comparative Analysis graph

V. CONCLUSION

The study concluded that the GRU model with 88% QWK is suitable and was used for the development of an efficient essay scoring model with high precision and a low loss function. The study suggested that further research could make use of more dataset as testing data and cross validated for effective evaluation performance. Research on various optimization techniques can be explored to determine a wide range of relevant datasets and dimensionality reduction in order to enhance classification performance.

VI. REFERENCES

[1]. Shetty, S. & Siddiqa, A. (2019). Deep Learning Algorithms and Applications in Computer Vision. International Journal of Computer Sciences and Engineering. https://doi.org/10.26438/ijcse/v7i7.195201.

[2]. Lim, C., Bong, C., Wong, W. & Lee, N. (2021). A Comprehensive Review of Automated Essay Scoring (AES) Research and Development. Pertanika Journal of Science & Technology. 29 (3): 1875 – 1899. https://doi.org/10.47836/pjst.29.3.27.

[3]. Ramesh, D. & Sanampudi, S.K. (2021). An Automated Essay Scoring System: A system literature review. https://doi.org/10.1007/s10462-021-10068-2.

[4]. Srivastava, K., Dhanda, N., & Shrivastava, A. (2020). An Analysis of Automatic Essay Grading Systems. International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, 8(6).

[5]. Page, E. B, (1966). "The Imminence of Grading Essays by Computer". Phi Delta Kappan, 48:238-243.

[6]. Hearst, M. A. (2000). The debate on automated essay grading. IEEE Intelligent Systems and their applications. 15(5), 22-37.

[7]. Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In New horizons in university teaching and learning: Responding to change. Centre for Educational Advancement, Curtin University. 173-184.

[8]. Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. Journal of Information Technology Education: Research, 2(1), 319-330.

[9]. Dong & Zhang (2016). Automatic Features for Essay Scoring – An Empirical Study. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 1072–1077.

[10]. Taghipour, K., Ng, H. (2016). A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing. 1882–1891. https://doi.org/10.18653/v1/D16-1193.

[11]. Wang Z., Liu J., & Dong R. (2018). Intelligent Auto-grading System. Proceedings of CCIS.

[12]. Chen and Li (2018). Relevance-Based Automated Essay Scoring via Hierarchical Recurrent Model. In: 2018 International Conference on Asian Language Processing (IALP). 378–383. https://doi. org/ 10. 1109/ IALP. 2018. 86292 56.

[13]. Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. 93–102.

[14]. Liang G, On B, Jeong D, Kim H & Choi G. (2018). Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture. Symmetry, 10(12), 682–. https://doi.org/10.3390/sym10120682.

[15]. Ruseti, S., Dascalu, M., Johnson, A., McNamara, D., Balyan, R., McCarthy, K., & Trausan-Matu, S. (2018). Scoring summaries using recurrent neural networks. In: International Conference on Intelligent Tutoring Systems. 191–201.

[16]. Chen, Z. & Zhou, Y. (2019). Research on Automatic Essay Scoring of Composition Based on CNN and OR. In: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). https://doi.org/10.1109/ICAIBD.2019.88370 07.

[17]. Cai C. (2019). Automatic essay scoring with recurrent neural network. In: Proceedings of the 3rd International Conference on High Performance Compilation, Computing and Communications. https://doi.org/10.1145/3318265.3318296.

[18]. Zupanc, K., & Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 118–132.

[19]. Xia, L., Liu, J., Zhang, Z. (2019). Automatic essay scoring model based on two-layer bidirectional Long and Short-Term Memory Network. In: Proceedings of the 2019 3rd International Conference on Computer Science and Artificial intelligence. https://doi.org/10.1145/3374587.3374596. 133-137.

[20]. Uto, M. & Okano, M. (2020). Robust Neural Automated Essay Scoring Using Item Response Theory. In: Artificial Intelligence in Education. AIED 2020. (12163). https://doi.org/10.1007/978-3-030-52237-7_44.

[21]. Zhu, W. & Sun, Y. (2020). Automated essay scoring system using multi-model Machine Learning. Computer Science & Information Technology. 109-117. https://doi.org/10.5121/csit.2020.101211.

[22]. Kuyoro, S., Eluwa, J., Awodele, O. & Ajayi, A. (2021). Characterization of Essay Content for Content-Based Assessment Using Morphological Classification Technique. International Journal of Scientific and Engineering Research. 12(1). ISSN:2229-5518.

Cite this article as :

Eluwa J., Kuyoro S., Awodele O., Ajayi A., "Essay Scoring Model Based on Gated Recurrent Unit Technique", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 2, pp. 323-330, March-April 2022. Available at doi : https://doi.org/10.32628/IJSRSET229257
Journal URL : https://ijsrset.com/IJSRSET229257