Enhancing Automatic Answer Evaluation Using Bi-Directional Long Short-Term Memory-Based Deep Learning

T.S Adharsh [1*] and M.K Jeyakumar [2]

[1]Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Kanyakumari(dist), TamilNadu, India. adharshts@gmail.com

[2]Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Kanyakumari(dist), TamilNadu, India. drjeyakumarmk@gmail.com

[*]Corresponding: adharshts@gmail.com

HOW TO CITE:
Amitava Sil, Sourav Dandapat, Rishav Raj (2026). Enhancing Automatic Answer Evaluation Using Bi-Directional Long Short-Term Memory-Based Deep Learning. International Journal of Special Education, 41(3s), 495-511.

ABSTRACT:

Automated Essay Scoring improves the effectiveness, consistency and fairness of evaluation through the use of deep learning models such as Bi-Directional Long Short-Term Memory in order to overcome some of the shortcomings associated with manual evaluation in large-scale educational testing situations. Automated Essay Scoring systems apply artificial intelligence and natural language processing technology to assess written responses quickly and without biasIt addresses the challenges in the education sector concerning the efficient evaluation of diverse and complex student responses due to the increasing number of learners. Conventional automated scoring approaches, which believe on lexical and syntactic pattern matching, fail to capture the deeper semantic and contextual nuances of short and essay-type answers. To resolve these limitations, the proposed framework employs a deep learning architecture based on the Automated Student Assessment Prize Short Answer Scoring, dataset, encompassing stages such as data exploration, feature extraction, model training, and evaluation. The system performs text structure and lexical feature analysis through visualization, while feature extraction involves tokenization, elimination of stop words, application of Term Frequency–Inverse Document Frequency, and truncated Singular Value Decomposition for effective dimensionality reduction. The comparative analysis between unidirectional Long Short-Term Memory and Bi-Directional Long Short-Term Memory demonstrates that the latter processes sequences in both forward and backward directions, thereby capturing contextual information more comprehensively and improving accuracy. The constructed Bi-Directional Long Short-Term Memory model, consisting of nineteen layers with dropout mechanisms to avoid overfitting, exhibits superior performance by achieving lower mean absolute error and validation

loss compared to the conventional Long Short-Term Memory model. The inclusion of callback functions during training allows dynamic performance monitoring and parameter optimization. Positioned within the broader context of automated essay scoring research, this study highlights the significance of deep learning and context-aware architectures in enhancing grading accuracy beyond simple lexical and syntactic evaluation. The findings reveal that the Bi-Directional Long Short-Term Memory based scoring system is an effective and reliable method that closely aligns with human assessment, offering a robust solution for modern educational evaluation.

Keywords: Answer scoring, answer evaluation, deep learning, Long Short-Term Memory, Bi-directional Long Short-Term Memory, Word Embedding

1. Introduction

The Automatic Essay Scoring (AES) system has become the most important product of the modern educational technology which has been evolved under the stress of universal education and the demand for some more rapid and objective forms of examination. Manual marking relies for its authority upon human intelligence and judicial power, but the experience is that it becomes more and more unmanageable and impossible in those vast educational systems which have developed owing to its slowness and liability to human inaccuracy and power of subjective judgment [1]. This produces lack of uniformity, and this of course makes it impossible to judge correctly the written answers to examination papers when such numbers have to be dealt with. During the early days of the introduction of automatic essay marking, it has been confined to the study of such external matters, as the frequency of words, the correctness of grammar, the number of words or sentences that the essay contains, or even such things as the average length of words in comparison to the number of words in the essay, and produced all the inaccuracies and imperfection that that phase of marking has shown in the past [2]. The results were slight advances realized, but it was soon found that this elementary form of text processing as a whole was inadequate to really estimate the depth of cognition required for this purpose, in that logical coherence, effective power of reasoning, structure of arguments or forms of language were matters educationally essential for the simulated marking of human beings [3]. The discrepancy between the results achieved by these elementary mechanical scoring systems and the necessary qualitative work to be done in the future by real examiners rendered it necessary that we should have a much better instrumentation process than one merely predicated on counting writing features, but something adequately qualified to elucidate both surface or obvious bubbles in the texts in question, as other lines of attack were producing heat but little or no light [4]. Recently, deep learning systems have taken a powerful change effect of late in the possibility of increasing and enhancing the capacities for the provision of AES in that the emphasis has changed from simply counting the writing features to be discovered in compositional style to the equivalent possibility of realising the meaning inherent in the text. The research given in the present paper, therefore, is a presentation of the composer of a new AES system which is based upon Long Short-Term Memory (LSTM) networks by Bidirectional LSTM Networks (Bi-LSTM) in the front the realms of modern human development of these mechanical aids. At the heart of this methodology is the idea that because the Bi-LSTM processes sequences of text in both

directions forward and backward richer contextual representations are produced giving a substantial improvement in the assessment of essays [5]. The bidirectional treatment of sequences is in fact indicative of the way humans read they remember content previously read and anticipate content to come to be able to decipher the semantic, contextual, and syntactic subtleties of superior writing. The Bi-LSTM is then a flexible and adaptable way of discovering relational mapping in essays and is therefore a strong analogue of human understanding and judging [6]. In order to discursively test this view the method used here is structured within the overall methodology of the Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) dataset which has a long history of being readily used for comparison between scores obtained through algorithmic means and scores set by human beings as to said answers supplied. The investigative project will be presented in a discrete series of stages starting with the overall exploratory data analysis phase [7]. This will entail a process which notably has the drawing of word clouds, examinations of length distributions of the text and of complexes of sentences involved so that the structural, linguistic patterns within the dataset will be the subject of investigation. Following this will come the feature extraction process where the text data contained in the unstructured essays will be represented by numerical vectors. In this phase important processes will involve the tokenisation of the text data, the removal of the irrelevant non-informative stop words, and the performation of Term Frequency–Inverse Document Frequency (TF-IDF) weightings on the data so that semantically rich terms will take precedence over more lexically contained ones [8]. As natural language data is characteristically high-dimensional, the dimensionality reduction is accomplished via truncated Singular Value Decomposition (SVD), thus keeping the most important semantic features and decreasing the computer cost and difficulties incorporating them into large vector space models. The comparative experimental design juxtaposes a standard one-way LSTM model with the proposed nineteen-layer Bi-LSTM. A number of these layers include the regularization technique dropout to avoid overfitting and to increase generalization. In addition, the utilization of state-of-the-art dynamic callback functions track important performance metrics such as validation loss and accuracy during the training process so that learning parameters can be adjusted dynamically to optimize convergence. The comparison of models is carried out using quantitative metrics from the test data set. The results demonstrate that the Bi-LSTM model has lower, Mean Absolute Error (MAE) and more regular trends in validation loss through epochs than the standard unidirectional LSTM base models [9]. The improvement in performance is consistent with the increase in prediction accuracy and stability in scoring reliability.

In addition to the numerical superiority of the Bi-LSTM, there is further evidence that its superiority stems from its ability to better understand context, maintain semantic coherence over lengthy sequences of text and to cope with long-range dependencies, which is a well-known problem with language models. More significantly, the existence of a strong relationship between the scores produced by the Bi-LSTM and those awarded by human raters would suggest that it is not, in fact, merely a competent estimate of human evaluative judgment. It also bridges the worlds of machine assessment and human interpretation [10]. This offers validation to the model as a method of assessing essays which is reliable and fair and which reduces the variance between assessors. These implications tend toward the scalability of AES and suggest that the application of complex, bidirectional models to large sample sizes is possible without loss of quality of measure, and that they an be a possible alternative to human grading in educational situations where large numbers of students must be assessed. This study thus places Bi-LSTM as a considerable advance in the science of automated essay scoring, since it can be said to consist of both extensibility and power, providing assessments which combine the rapidity of computation with a great depth of

analysis of semantics. Socio-interpretational analysis almost human-like in its critical expression, the model indicates a giant leap towards a fairer and worthier and more responsive education. The increased accuracy and sensitivity in foreclosure voiced by Bi-LSTM offer the prospect of diminishing the labor of evaluators, a standardisation of evaluative outcomes and a basis of useful pedagogical suggestions to pupils. This is a development which represents the dawn of an AES paradigm, capable of simulating the integrity, insight and salient knowledge of expert human examination, the beginning of a new era in terms of computerized examination.

2. Methodology

The automated scoring system suggested in this work is an efficient, robust and scalable assessment scheme, capable of meeting the ever-changing demands of the modern education system. It employs innovative techniques based on machine learning and Natural Language Processing (NLP) and is implemented in four different phases, namely data exploration, feature extraction, model training and evaluation. The purpose of data exploration is to analyse the textual data in such a way that the linguistic features such as frequency distributions of words, semantic groups of words, and indicators of textual complexity, such as average length and coherence of sentences, come to the surface. The final products of visualising tools such as word clouds and mathematical analyses are the results of an initial treatment of the data set that provide the basis for the preconditions for further treatments of the data itself. Once these insights have been gained, the features extraction stage of the process consists of creating machine-readable representations of raw textual data through the processes of tokenisation, removal of stop words to cleanse the data of spurious effects, and the implementation of TF-IDF to measure the importance of terms in the context of the corpus studied. For computational efficiencies to be gained, however, truncated SVD is then used, which reduces the existential dimensionality of these feature vectors and makes their application to models easier in the training processes.

The automated scoring system suggested in this work is an efficient, robust and scalable assessment scheme, capable of meeting the ever-changing demands of the modern education system. It employs innovative techniques based on machine learning and NLP and is implemented in four different phases, namely data exploration, feature extraction, model training and evaluation. The purpose of data exploration is to analyse the textual data in such a way that the linguistic features such as frequency distributions of words, semantic groups of words, and indicators of textual complexity, such as average length and coherence of sentences, come to the surface. The final products of visualising tools such as word clouds and mathematical analyses are the results of an initial treatment of the data set that provide the basis for the preconditions for further treatments of the data itself. Once these insights have been gained, the features extraction stage of the process consists of creating machine-readable representations of raw textual data through the processes of tokenisation, removal of stop words to cleanse the data of spurious effects, and the implementation of TF-IDF to measure the importance of terms in the context of the corpus studied. As a result of the increasing necessity of scalable and convenient assessment systems in the education industry, AES systems eliminate the manual grading system and provide the learners with timely feedback. Enhanced with NLP, which has proven relatively effective both in text classification and automated grading activities, this strategy is implemented in a four-step pipeline data exploration, feature extraction, model training, evaluation, and it provides the structure of the research as in Figure 1. All the stages are very essential in building a modern, reliable system of evaluation in the modern education.

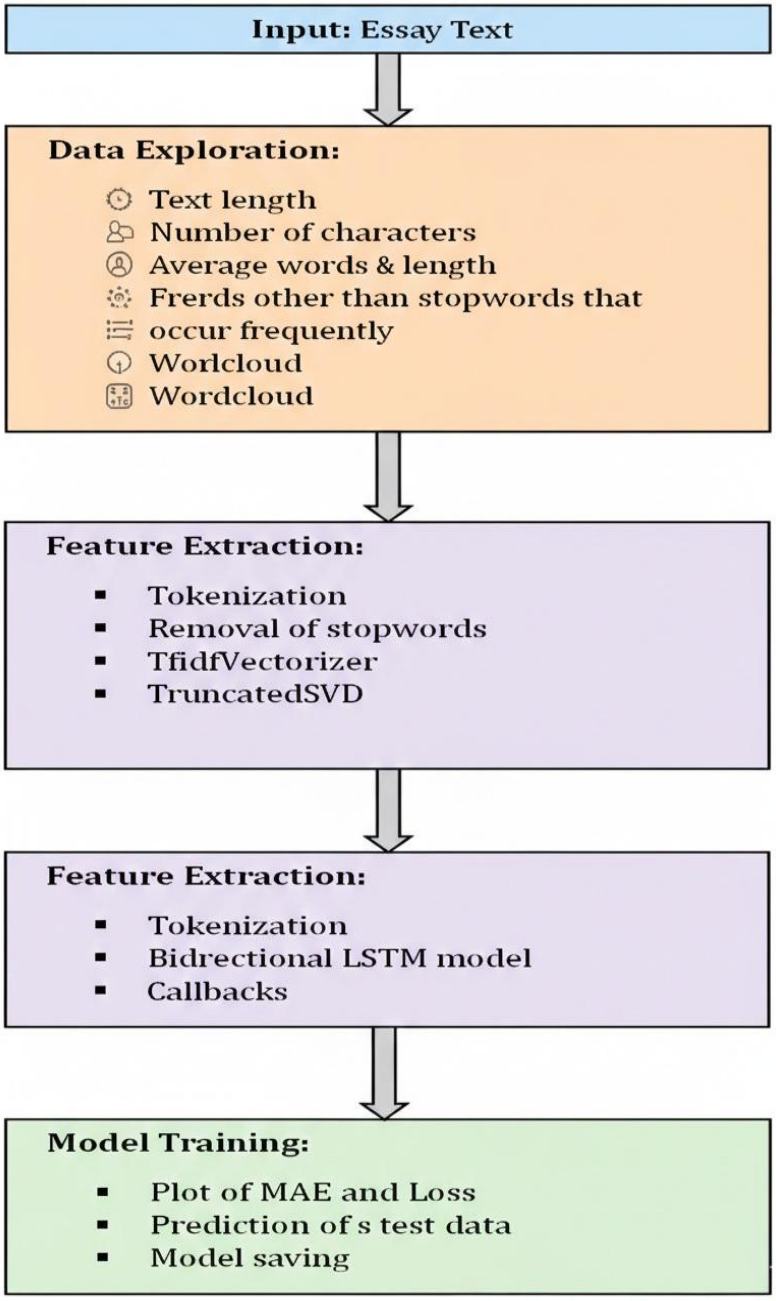


Figure 1: Architecture of the proposed system

For computational efficiencies to be gained, however, truncated SVD is then used, which reduces the existential dimensionality of these feature vectors and makes their application to models easier in the training processes.The training of models is based upon these representations through the LSTM networks and more preferably, the Bi-LSTM which works with the textual sequences in both forward and backward directions to produce more contextual information. The dropout layers at this point counteract overfitting, and dynamic callback processes allow monitoring and improvement of performance in real-time, which prevents the model to change effectively over the course of learning. The performance of the system is scientifically tested in the final evaluation phase and measured in terms of mean absolute error (MAE) and validation loss with the aim to have automated scores as close to those of human graders as possible. This validates the purpose of

creating an effective and dependable automated grading system.

2.1 Data Exploration

In order to render these analytical results even more interpretable and useful, they are represented in graphic form, that is bit maps and word clouds are employed, in order graphically to illustrate the trends in the lexical changes that occur. The word clouds show the relative weight of various words by altering both their size and color of meaning according to their frequency as to word and as to context, thus representing theses words graphically as it were, being in black and high words in red, or light yellow, for instance in the various hues of exclamations oaths, etc. The word clouds are, therefore, a powerful diagnostic means, as well as a fascinating visual presentation for the isolation and understanding of recurring motifs, underlying themes and covert lexical choices in a corpus of texts. This visualization gives at once a rapid and interpretable summary, in order that the gulf may be bridged between qualitative information and quantitative study and so that the important content patterns and semantic relationships may be directly isolated and exhibited. So that taking it as a whole, the exploratory phase gives us a fair flat, multi-dimensional basis whereon we may advance to the actual data and elicit deeper feature extraction and modelling, in order to have the really high quality textual features isolated and introduced into the arsenal in order to increase the effectiveness of the mechanized rating system for essays. The exploratory analysis of the great mass of essays written by students might be regarded as data is the first step in the work on the automated essay scoring project. This preliminary process is a necessary starting off place in order that an investigation should be made of the lexical elements that are available and what hidden variations are to be found in evidence as pertaining to the successful construction of a scoring machine. This analysis is made on levels by carrying out such simple quantitative measurements as counting the number of letters, the number of words and finding the average length of words that are counted, and after these have been completed there are other and valuable measurements made by an analysis of the linguistic phenomena.

Among the advanced linguistic phenomena used to an advantage, or possibly sufficient to create some needed exploration and development of vocabulary or, systems of meanings, is the elimination of high frequency stop words which ordinarily have but little meaning-per- word because it gives a meaning to the various descriptive and pertinent message units. Typical evidences of high or low writing values are what might be called structural variety which are measured by the varied lengths of sentences and paragraphs, some indices of textual cohesion including a local and global organization, date of writing lexical variety measured in terms of frequency of variables and their mean age of acquisition, syntactical phrase variety in parts of speech and forms of rhetorical structure. Some of the structural varieties are measured; the mean clause length and sentence length which can be easily found. The structure can be utilized to express the values of the clause structure sub-clause structure and the multi-clause sentence involved in the contributions for the essay-tests.

2.2 Feature Extraction

Once the data exploration has been performed, the next step in the methodology is feature extraction, the step in which raw textual data is transformed necessitating structure or information that can be fed in a form suitable for either machine learning or natural language processing methods. This process begins with tokenization from where the essay text is dissected into smaller pieces reputed tokens, which may be words, subwords or punctuation marks and this is the basis of the future computation analysis therein. Most tokens are designated through the influence of whitespace, are formed into high dimensional embedding vectors which preserve the semantic and syntactic style of the text [11]. They can be computed by unsupervised methods which imply co-occurence frequencies of successor words or supervised

methods specific to particular applications of NLP. Good tokenizing is done to ensure the intergrity of the complicated structures of linguistics of the whole essay. After this step has been accomplished, the step follows of doing away with the stopwords words which add little or uncompelling meaning such as is, have, the, etc. these effectively join up significant meanings in the data. If they have been removed it makes the remaining features thereby much clearer and much more associative and the quality of writing of the essays may be achieved to any individual essay with much greater understanding. In cases of need, detokenization is used to regenerate the text in order that the text as was, may be more easily analyzed using the terms of specific languages [12]. TF-IDF vectorization is a common method of numerical encoding the text that has been processed, giving all words relative importance within a single essay and the corpus as a whole; contextually relevant words become prominent and words that occur frequently, but are less informative are deemphasized. The last operation is dimensionality reduction by truncated SVD, which reduces the set of features to the most salient features, which improves calculation efficiency and allows the model to learn better in later steps.

2 .3 Model Training

The model is initially trained using a LSTM network and subsequently with a BiLSTM to generate the scoring for the given essays based on their textual features. The training process also incorporates a callback function to optimize model performance and prevent overfitting during training [13]. The LSTM operates on a feature-extracted input sequence represented as x=(x1,x2,…,xn) where *n* denotes the length of the sequence. Its architecture is designed to capture long-term dependencies within sequential data and overcome the limitations of traditional recurrent neural networks (RNNs) that struggle with vanishing or exploding gradients. The LSTM cell consists of three control gates—forget, input, and output—that collaboratively manage the flow of information within the network [14]. The forget gate determines how much of the previous memory state *ct−1* should be preserved for the current state ct while the input gate regulates the extent to which new information from the current input *xt* is integrated into the memory cell. The output gate subsequently dictates how much of the updated cell state *ct* is exposed to produce the hidden state *ht* which serves as the output for the current time step. Each gate is fully connected and uses the logistic sigmoid activation function to generate values between 0 and 1, effectively controlling the volume of information allowed to pass through at each stage. The proposed model architecture integrates a single-layer LSTM with a sigmoid activation function *σ*, weight matrices (*W*), and bias vectors (*b*) corresponding to the three gates, as well as word vector representations (*xt*) and hidden states (*ht*). Furthermore, dropout and dense layers are applied to enhance generalization capability and produce efficient and accurate essay scoring outcomes. Figure 2 shows **s**tructure of the long short-term memory.

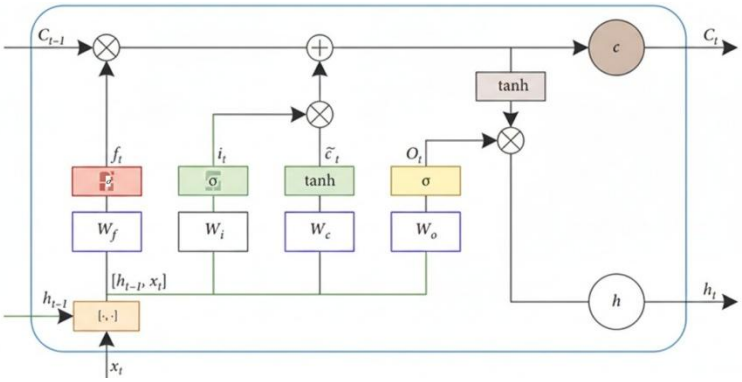


Figure 2: Structure of the long short-term memory

Input gates: $i_t=\sigma(W_{ix}x_t+W_{ih}h_{t-1}+b_i)$

Forget gates: $f_t=\sigma(W_{fx}x_t+W_{fh}h_{t-1}+b_f)$

Output gates: $o_t=\sigma(W_{ox}x_t+W_{oh}h_{t-1}+b_o)$

Cell states: $f_t*h_{t-1}+ i_t*\tanh(W_{cx}x_t+W_{ch}h_{t-1}+b_c)$

Cell outputs: $h_t=o_t*\tanh(c_t)$

(a)

lstm_input	input:	[(None, 1, 300)]
InputLayer	output:	[(None, 1, 300)]
lstm	input:	(None, 1, 300)
LSTM	output:	(None, 1, 300)
lstm_1	input:	(None, 1, 300)
LSTM	output:	(None, 64)
dropout	input:	(None, 64)
Dropout	output:	(None, 64)
dense	input:	(None, 64)
Dense	output:	(None, 1)

(b)

bidirectional_15_input	input:	[(None, 1, 500)]
InputLayer	output:	[(None, 1, 500)]
bidirectional_15(lstm_15)	input:	(None, 1, 500)
Bidirectional(LSTM)	output:	(None, 1, 1000)
bidirectional_16(lstm_16)	input:	(None, 1, 1000)
Bidirectional(LSTM)	output:	(None, 1, 1000)
bidirectional_17(lstm_17)	input:	(None, 1, 1000)
Bidirectional(LSTM)	output:	(None, 1, 512)
bidirectional_18(lstm_18)	input:	(None, 1, 512)
Bidirectional(LSTM)	output:	(None, 1, 256)
bidirectional_19(lstm_19)	input:	(None, 1, 256)
Bidirectional(LSTM)	output:	(None, 128)
dropout_3	input:	(None, 128)
Dropout	output:	(None, 128)
dense_6	input:	(None, 128)
Dense	output:	(None, 64)
dense_7	input:	(None, 64)
Dense	output:	(None, 1)

M str

Bi-LSTM network has been employed to enhance the performance and accuracy of the essay scoring system by leveraging the contextual dependencies present in both directions of a text sequence. Unlike the standard unidirectional LSTM, which processes information only from past to present, thereby limiting its understanding of upcoming words and phrases, the Bi-LSTM architecture processes data in two directions—forward and backward—allowing it to obtain comprehensive information from both previous and subsequent contexts in a sentence [15]. This dual processing capability enables the model to capture more meaningful semantic and syntactic relationships, leading to improved contextual understanding. In natural language sequences, the meaning of a current word often depends not only on the words that precede it but also on the words that follow. For instance, the interpretation of a term can vary considerably when understood within its future context. Therefore, the Bi-LSTM network overcomes this limitation by combining two independent LSTM layers with opposite orientations into a unified structure that transmits information from both directions simultaneously. In the forward layer, the sequence is processed in its natural order, analyzing each token from left to right, while the backward layer processes the same sequence in reverse order, from right to left [16].

Bi-LSTM has its major advantage in the fundamental idea to extend the contextual meaning ex post facto by taking into consideration backward and forward from the sequence of input. At each time step a forward layer sequentially reads the text towards its end and a backward layer sequentially reads the text towards its end. From those two layers arises the different hidden states and those two different states are concatenated. This very all important consideration of the concatenation of the hidden

states produces a complete representation of the input which in relation to the output layer gives concomitant access to the contextual meaning both backward and forward. This two edged view is a direct reply to the principal shortcomings of the unidirectional LSTMs that is the lack of access to the future context. The importance of this context is most relevant in the application of essay scoring where logical flow and grammatical coherence are amongst the overriding factors since this particular context is indispensable for the full appreciation of meaning and structure of the text. Because of the forward and backward integration of the information, the Bi-LSTM architecture is enabled to embrace the long range dependencies and other subtleties of language use that obtain across sentences and paragraphs. Also by processing the input flows in both directions separately and before generating the output they get to be regarded, the architecture is thus retained by the necessary dependencies and furthermore greatly enhances the representability and learning capabilities of the model. Hence the Bi-LSTM neural architecture may grant more subtle and richer semantic representation of the contents of the essays so that they may relate to the total quality of the language in its entire content rather than in an isolated way and will give therefore relative and context meaningful qualities. Its evolved memory and also context encoding also ensures it can evaluate with accuracy factors like the essential ones in writing like the coherence of content of text, syntactic integrity, and the technical inner coherence or diversity of vocabulary which are necessary parts of an efficient computer essay evaluation system. The basis of the essay scoring process modeling is the Bidirectional LSTM, which was chosen to overcome the limitations of unidirectional models and reinforce automatic assessment. In a typical LSTM, sequences are only processed forwardly and the sequence makes use of contexts that occur before it, but it is also unable to use information that occurs ahead, and this can result in the sentence losing its actual meaning. Bi-LSTM architecture solves this by combining two separate layers of hidden state, which interpret the input sequence during both forward and backward direction of running the sequence. The design will enable this model to identify contextual dependencies of the past and the future of any particular word. The network can obtain the more detailed contextual representation by concatenating the results of these opposing layers, and therefore the final scoring would be informed by the whole sentence structure. This ability to absorb information going in both directions is the primary factor that makes bidirectional networks show significant better results in comprehending subtle language as compared to the unidirectional networks. The forward hidden sequence $\overrightarrow{h_t}=(\overrightarrow{h_1}, \overrightarrow{h_2}, \dots \overrightarrow{h_n})$ of a bidirectional LSTM considers the input in ascending order and the backward hidden layer $\overleftarrow{h_t}= (\overleftarrow{h_1}, \overleftarrow{h_2}, \dots \overleftarrow{h_n})$, which considers the input in descending order for the input sequence $x=(x_1,x_2,\dots x_n)$. Following is an illustration of how the two network directions behave independently up until the final layer, where their outputs are concatenated as $y_t=[\overrightarrow{h_t},\overleftarrow{h_t}]$:

$$\overrightarrow{h_t}=\sigma\left(W_{\overrightarrow{h}x}x_t+W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1}+b_{\overrightarrow{h}}\right)$$

$$\overleftarrow{h_t}=\sigma\left(W_{\overleftarrow{h}x}x_t+W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1}+b_{\overleftarrow{h}}\right)$$

$$y_t=W_{y\overrightarrow{h}}\overrightarrow{h_t}+W_{y\overleftarrow{h}}\overleftarrow{h_t}+b_y$$

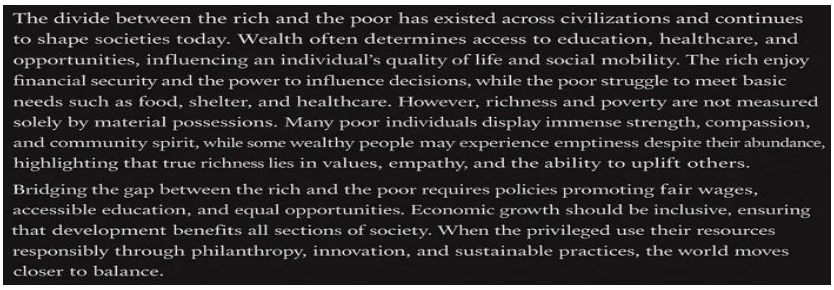
The output sequence of the hidden layer is expressed as $y_t=(y_1,y_2,...,y_n)$ $y_t = (y_1,y_2,...,y_n)$. The model architecture integrates a bidirectional long short-term memory (Bi-LSTM) framework composed of 19 layers, including three dropout layers for regularization and seven dense layers for feature mapping and classification. The structural configuration of this Bi-LSTM architecture is depicted in Figure 4, providing a detailed visual representation of its layered organization. Throughout the training iterations, a callback function is employed to monitor model performance, optimize learning efficiency, and ensure adaptive adjustments based on real-time validation outcomes.

The student model is also tested regularly at various checkpoints with a development set during the training phase to monitor the

performance of the student model and ensure a constant level of learning. As well as this, the loss curve is observed to identify model accuracy and convergence variations during the training cycles. During this process a callback will be used which will be provided with the student model and also a training step. Continuation-passing style programming is a technique where functions are called in a specified order depending on the result that happened at the last function call. The training uses this in the way of employing callbacks at varying points of the training procedure. In this situation the distiller will keep the state of the student model by using a callback mechanism which is called at every `checkpoint` based on the parameters `num_train_epochs` and `ckpt_frequency`. This type of callback allows dynamic checking and storing of various parametrics and controlling at certain regular time intervals throughout the course of the whole training, although to allow maximum performance without overfitting or losing important aspects of the model.

3 . Result And Discussion

The proposed automated essay scoring system was evaluated using the public ASAP-SAS dataset which includes 10th grade student responses scaned to text via OCR. The evaluation started with the data exploration phase in which a number of textual characteristics such as sentence length, average word length, number of characters, and distribution of lexicons, were analyzed and also visualized by such visual techniques as word- clouds and frequency plots to expose useful trends. The automatic essay marking system was evaluated on publicly available ASAP-SAS dataset, which consists of the responses of Grade 10 students made digital through the OCR technology. The analysis of the textual characteristics to be examined in the evaluation includes the initial stage of the exploration of such essential textual features as the length of a sentence, the average length of words, the number of characters, and the distribution of lexical items, which was backed by the methods of visualization, including word clouds and frequency plots to determine the patterns of importance. Following the extraction of the features, which includes, tokenization, stopword elimination, TF-IDF vectorization and dimensionality reduction based on the truncated SVD, both, unidirectional LSTM and Bidirectional LSTM (Bi-LSTM) models were trained. LSTM model had 814,705 parameters trained during 150 iterations, but its fluctuations in the mean absolute error (MAE) and loss values were significant, thus scoring was not consistently stable. On the other hand, the Bi-LSTM model, with 19 layers and 13,411,393 parameters, performed better, as it analyzed the text input in both directions, thus, complex relationships between the context. The Bi-LSTM was trained to achieve both shorter (26 epochs) and longer (150 epochs) forms, and achieved the best performance metrics, the MAE of 0.6708 and the validation loss of 0.3503 in the short run, and steadily good trends in the long run. These results indicate that the bidirectional architecture provides a stable, context-sensitive scoring method, which is significantly accurate to human scoring. The evaluation started with the stage of data exploration, and a sample input essay is presented in Figure 5.


The divide between the rich and the poor has existed across civilizations and continues to shape societies today. Wealth often determines access to education, healthcare, and opportunities, influencing an individual's quality of life and social mobility. The rich enjoy financial security and the power to influence decisions, while the poor struggle to meet basic needs such as food, shelter, and healthcare. However, richness and poverty are not measured solely by material possessions. Many poor individuals display immense strength, compassion, and community spirit, while some wealthy people may experience emptiness despite their abundance, highlighting that true richness lies in values, empathy, and the ability to uplift others.

Bridging the gap between the rich and the poor requires policies promoting fair wages, accessible education, and equal opportunities. Economic growth should be inclusive, ensuring that development benefits all sections of society. When the privileged use their resources responsibly through philanthropy, innovation, and sustainable practices, the world moves closer to balance.

Figure 4: Sample input essay

The proposed automated marking system of essay assignments begins at the data exploration stage which will involve a sophisticated study of the statistical properties in order to produce a thorough account of the structural and lexical structure of the essays. In the first instance the total length of text produced is an indicator for general brevity and verbosity of response. Another important property in this analysis is the number of words per sentence which is an overall crack ontologoically of syntactical complexity and descriptive depth which means an immediate factor in the determination of readability, coherence and quality of writing in a global sense. The figures are exhibited graphically to determine the distributions, to observe outliers and variations which would give an indicator of inconsistency or imbalance in our developments of ideas in the different essays. On the basis of all of this, this analysis will also derive the average length of the words as a measure of the luxurient property of the words concerned, and the average word lengths in themselves are likely to be a good indicator of the grade of development in the use of the language employed and the complex constructions used. Finally there is a character count type of analysis which gives a count of the number of characters in each sentence and which provides a more discriminative account as to the length of the sentences established than that of the words since it covers all punctuation, word density and authorial style. These graphical accounts of the sentence length, word complexity and character count when combined would give a multi-dimensional and quantitative description of the subject which are necessary for the basis of subsequent feature extraction as modelling training.

Together, the visual image of these three measures an average words per sentence, average word length, and characters per sentence creates a structure profile of the essays as a whole. When these visualizations are compared to each other, one can identify the underlying patterns that are used to evaluate writing fluency, complexity and cohesion. These parameters are shown in figures 6(a), 6(b), and 6(c) respectively; they are average word count in a sentence, average word length and the number of characters per sentence.

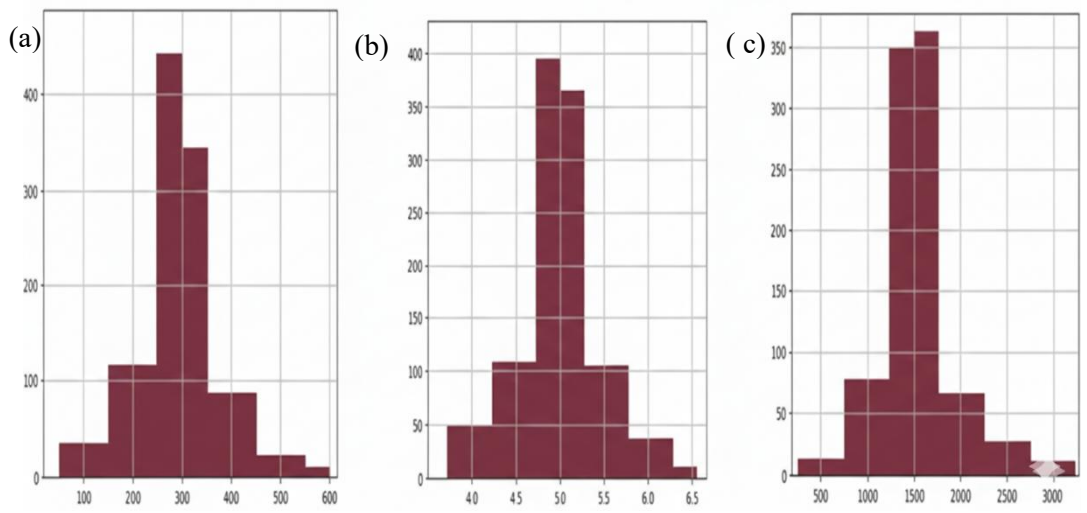


Figure 5: Average number of words in the sentence,(b) Average word length in each sentence, (c) Number of characters present in each sentence

In order to visually examine the textual information, a word cloud was created, which is represented in Figure 6(a). Such visualization method shows the frequency of words in the

essays, the size of the word also reflects its frequency and gives a direct idea of the most common words. This was followed by a refined analysis process after tokenization and the elimination of stopwords and the high-frequency keywords were then represented as high-frequency keywords in Figure 6(b). The model training was done in two phases after the data exploration and feature extraction. The first phase entailed training a default LSTM model, the architectural parameters, and the 814,705 trainable parameters of which are provided in Figure 6(c).

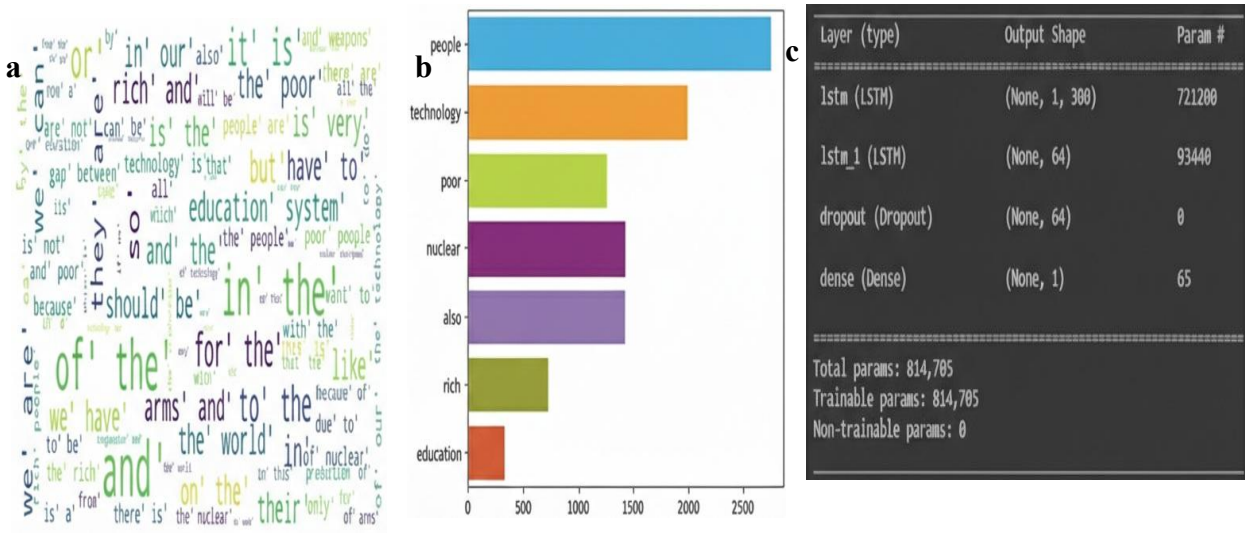


Figure 6:(a) Word cloud, (b) Statistics of frequently used words, (c) Parameters of the LSTM model

Figure 6 presents a comprehensive overview of the textual analysis and model architecture used in the essay evaluation process. Subfigure (a) displays a word cloud that visually represents the most frequently occurring words in the dataset, with terms such as "people," "education," "technology," "rich," "poor," and "nuclear" appearing prominently. These keywords suggest that the essays primarily revolve around societal issues, technological progress, and education-related themes. Subfigure (b) shows the statistical distribution of frequently used words, where "people" and "technology" occur most often, followed by "poor," "nuclear," and "education," highlighting the major areas of focus within the essays. This frequency analysis provides insights into the linguistic tendencies and thematic diversity of the dataset. Subfigure (c) illustrates the parameters of the LSTM model used for essay evaluation.

The suggested model architecture meant to learn semantic patterns to be able to score automated essays on a career-focused scale will include two layers of LSTMs (with 300 and 64 output units, respectively), a dropout layer to regularize the process and a dense output layer, resulting in a total of 814,705 trainable parameters which is a fairly complex model. Nonetheless, the visualized results of the performance in the form of the Mean Absolute Error (MAE) in Figure 7, and the loss values in Figure 8 indicate a major limitation. Both metrics have high levels of fluctuations and instability in their learning curves. This unstable waveform shows that the model has difficulties in convergence and generalization and it finally results in suboptimal and unreliable scoring output against its theoretical capability.

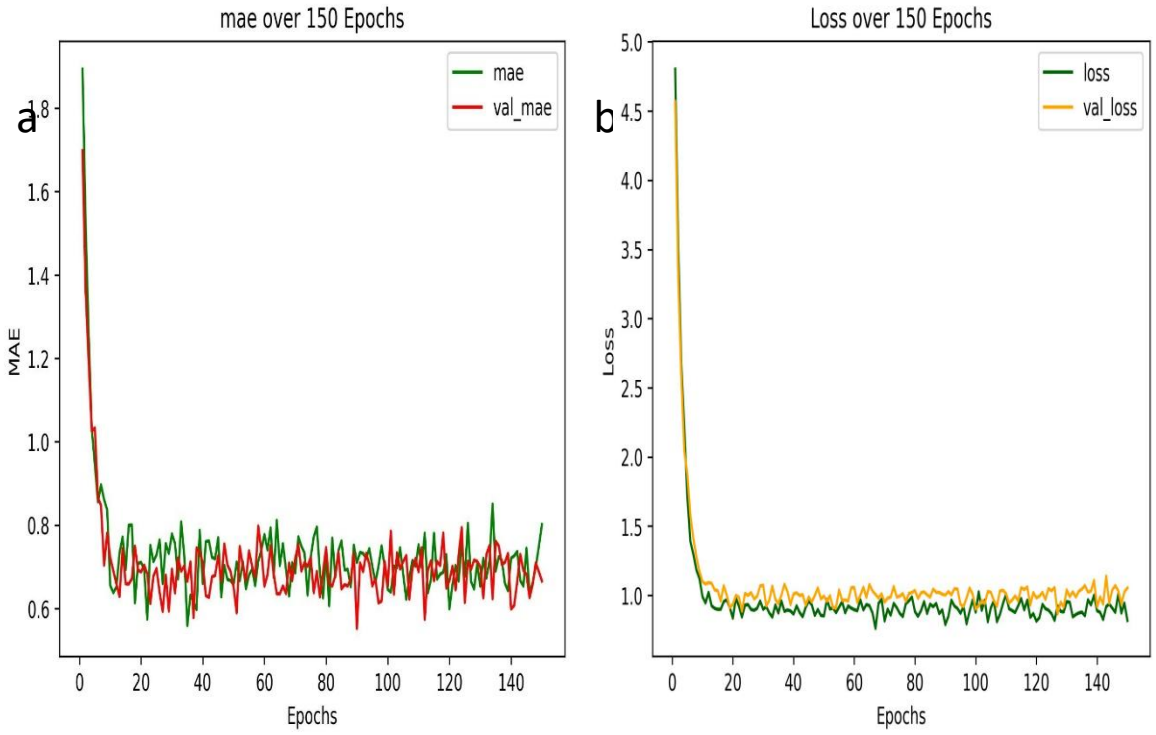


Figure 7: (a) Mean absolute error and (b) Loss value of the LSTM system

As Figure 7a shows, the Mean Absolute Error (MAE) of the LSTM model has a significant volatility in the course of the 150 epochs, and hefty fluctuations are observed in the training and validation curve. Such non-convergence to a consistent minimum means that it has low generalization ability, and this is one of the major reasons that it scores lowly. The ongoing oscillations indicate some inherent challenge in creating a powerful mapping of textual entry and score, and this also shows the need to employ a more complex modeling strategy. This is possibly due to the fact that the bidirectional LSTM architecture is more successful in alleviating the problem of such instability, providing a smoother convergence process and more accurate predictions.

Figure 7b illustrates the loss variation of the LSTM model over 150 epochs for both training and validation datasets. Initially, a steep decline is observed, indicating that the model rapidly learns during the early epochs. However, after around 20 epochs, the loss values begin to stabilize, fluctuating slightly but remaining relatively consistent thereafter. The persistence of minor oscillations suggests limited convergence stability and possible overfitting tendencies. The overall trend shows that while the model achieves partial learning efficiency, it fails to attain optimal loss minimization, implying that the LSTM architecture alone is insufficient for achieving strong generalization in essay scoring tasks.

(a)

Layer (type)	Output Shape	Param #
bidirectional_35 (Bidirectional)	(None, 1, 1000)	4004000
bidirectional_36 (Bidirectional)	(None, 1, 1000)	6004000
bidirectional_37 (Bidirectional)	(None, 1, 512)	2574336
bidirectional_38 (Bidirectional)	(None, 1, 256)	656384
bidirectional_39 (Bidirectional)	(None, 128)	164352
dropout_7 (Dropout)	(None, 128)	0
dense_14 (Dense)	(None, 64)	8256
dense_15 (Dense)	(None, 1)	65

Total params: 13,411,393
Trainable params: 13,411,393
Non-trainable params: 0

(b)

	Unnamed: 0	promptId	uniqueId	essay	predicted_score
0	0	1	1_315	Curriculum has been adopted in many schools. T...	3.0
1	1	1	1_214	I strongly agree with the statement , The tig...	3.0
2	2	1	1_196	Imagination and creativity is the most importa...	3.0
3	3	1	1_178	In our eduction system leaves no room for imag...	2.0
4	4	1	1_201	I will agree at some what extend, because if w...	3.0
...	...	...	...	...	...
300	300	5	5_146	Earth is a creation of God and everything that...	3.0
301	301	5	5_65	production of arms and weapons in this present...	3.0
302	302	5	5_151	Race to become more powerful can destroy the e...	3.0
303	303	5	5_404	In its attempt to harness the power of the ato...	3.0
304	304	5	5_360	Racein the production of arms and weapons in t...	2.0

305 rows × 5 columns

(c)

	Unnamed: 0	promptId	uniqueId	essay	evaluator_rating
0	0	1	1_323	At present age, our education system is not go...	3.0
1	1	1	1_238	I am agree the tightly defined curriculum of o...	4.0
2	2	1	1_212	I strongly agree with the statement that tight...	2.0
3	3	1	1_117	Our education system is nice quitely but i dis...	2.0
4	4	1	1_229	i am totally agree with the statement that tig...	3.0
5	5	1	1_226	I am totally disappointed(Not Agree) with our ...	2.0
6	6	1	1_99	the education system is very rough it does not...	3.0
7	7	1	1_240	I agreed with the tightly defined curriculum o...	2.5
8	8	1	1_127	Now a days, education system is very important...	2.0
9	9	1	1_176	In the current scenario,education plays a vita...	4.0

Figure 8: (a) Parameters of Bidirectional LSTM model, (b) Predicted score for the test essay, (c) Evaluator rating

Figure 8 presents the structural details and performance outputs of the Bi-LSTM model applied for automated essay scoring. 8(a) outlines the model's architecture, comprising multiple bidirectional LSTM layers with decreasing neuron sizes (1000, 512, 256, and 128), followed by a dropout layer and dense layers for final prediction. The model contains a total of 13,431,393 trainable parameters, indicating high learning capacity suitable for complex text analysis. 8(b) illustrates the predicted scores generated by the Bi-LSTM model for test essays across various prompts. The predictions exhibit an aptness in scoring trends which indicates the ability of the model to effectively analyse in a fair manner the subjects of the essays and to input scores into the scoring which reflects the deficiencies in the writing of the essays. The evaluator scores of these essays are shown in figure 8(c) representing the ground truth or what is the benchmark for comparative purposes. The nearness of the predicted scoring with those given by the evaluator show the correctness and the power for generalisation of the proposed model. These figures taken all together show the efficiency with which the Bi-LSTM model captures the semantic contexts and linguistic models within the essays. The hierarchical model used permits a deep extraction of features to take place as represents the practical value in the said predictions illustrating the reliability of the model with regard to the tasks in the classroom of the automated scoring of essays thus cutting down the amount of manual work to be performed and thus allowing proper and consistent values for scoring to be obtained from which to evaluate consistently the results obtained. The mean absolute error (MAE) and loss values for the Bi-

LSTM model were examined after 26 epochs and 150 epochs, which are shown in Figures 9(a) and 9(b) respectively. These values produce the optimum values for the MAE value which is found to be 0.6708, and the validation loss which attains the value 0.3503 after the mentioned 26 epochs. The values for the MAE and validation loss over 150 epochs of training are accordingly illustrated in Figures 9(c) and 9(d). These plots represent the trend of results for both the MAE and validation loss, which will lead to a better understanding of the model performance over prolonged training along with a guide as to the convergence properties and stability which is exhibited over differing numbers of epochs or training iterations for the model.

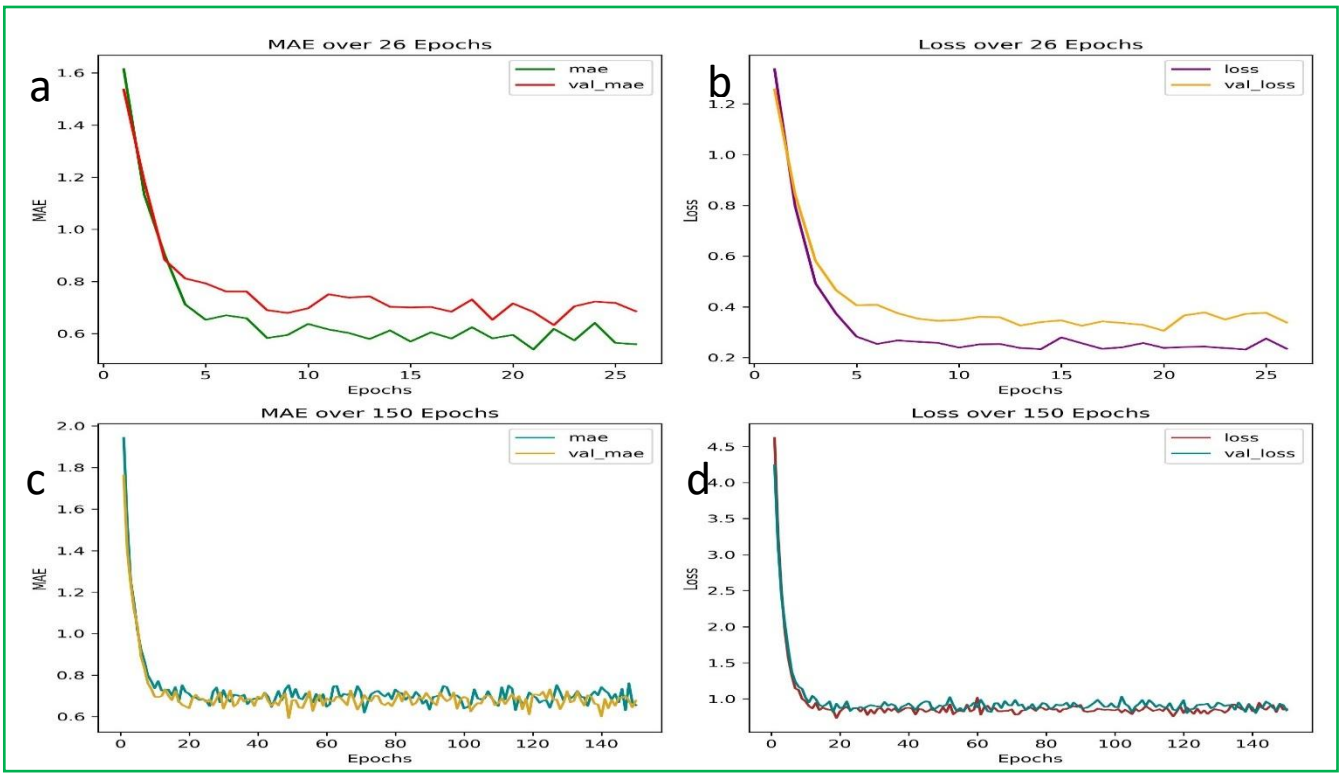


Figure 9: (a) MAE over 26 epochs on the Bi-LSTM model, (b) Loss over 26 epochs on the Bi-LSTM model, (c) MAE over 150 epochs on the Bi-LSTM model, (d) Loss over 150 epochs on the Bi-LSTM mode

The convergence patterns in both MAE and loss curves highlight that the Bi-LSTM effectively captures temporal dependencies in the data, achieving optimal learning early and maintaining steady accuracy over time. Overall, the figures collectively indicate that the Bi-LSTM model is well-trained and robust, with strong agreement between training and validation metrics throughout the training process.

5. Conclusion

This research analyzes the implementation of machine learning techniques to deal with the scoring of student essays. This preliminary analysis of the subject used the analysis of an LSTM model, which proved deficient, and an advanced Bi-LSTM model equipped with callback functions were constructed. The advanced model succeeds in improving the accuracy and reliability of the grading procedure, and produces a superior instrument with which to evaluate student answers. Due to the limitation in LSTM for optimal scoring, the Bi-

LSTM model has been developed with callbacks to obtain the student answer scoring system. The study highlights that traditional manual grading is increasingly challenging due to a growing student population, and older automated methods relying on lexical and syntactic patterns often fail to capture deep semantic contexts in student responses. This research paper shows that a deep learning-driven system involving a nineteen-layer Bi-LSTM network with dropout regularisation is a very effective system to use in automated essay marking. The model, which is trained and tested on the ASAP Short Answer Scoring dataset, uses its bidirectional processing with the ability to process forward and backward text sequences to extract more contextual information out of student essays and greatly enhances the accuracy of grading. Quantitatively, the Bi-LSTM had a significantly lower Mean Absolute Error (MAE) of 0.6708 and a validation loss of 0.3503 than the unidirectional LSTM, which highlights the high performance of the Bi-LSTM. The model also incorporated dynamic call back functions which increased the robustness of the system as the performance could not only be monitored in real time, but parameters could be tuned as well during the training. The system proved to be very efficient and this was aided by the elaborate data preparation pipeline, which presented the data optimally structured for deep learning analysis by way of an elaborate exploratory analysis process, tokenization, stop-word elimination, TF-IDF weighting and dimensionality reduction using truncated SVD. In the overarching framework of automated scoring, the results indicate the necessity for more advanced, complicate context aware architecture, such as Bi-LSTM, rather than the more simplistic lexicon based architecture, since this is a powerful and scalable and reliable approach that can be aligned closely with the human criterion of evaluation, lessening the work load of the evaluators and maintaining the level of invariance of grading, consistency of grading, integrity of grading and level of interpretation.

References

Alhusseini, M. M., & Feizi-Derakhshi, M. R. (2025). Benchmarking ChatGPT and DeepSeek in April 2025: A Novel Dual Perspective Sentiment Analysis Using Lexicon-Based and Deep Learning Approaches. *arXiv preprint arXiv:2509.19346*.

Deepender, & Walia, T. S. (2025). Optimizing Automated Scoring for Hindi Responses: A Hybrid Framework with Advanced Feature Integration. *International Journal of High Speed Electronics and Systems*, 2540807.

Hamidi, H., & Ghelichi, M. (2025). A hybrid deep learning technique for depression detection of Persian tweets using using Bi-directional LSTM-CNN model. *Multimedia Tools and Applications*, 1-31.

Tran, Q. D. L., & Le, A. C. (2023). Exploring bi-directional context for improved chatbot response generation using deep reinforcement learning. *Applied Sciences*, *13*(8), 5041.

Dada, I. D., Akinwale, A. T., & Tunde-Adeleke, T. J. (2025). A Structured Dataset for Automated Grading: From Raw Data to Processed Dataset. *Data*, *10*(6), 87.

Itriq, M., & Noor, M. H. M. (2025). Arabic hate speech detection using deep learning: a state-of-the-art survey of advances, challenges, and future directions (2020–2024). *PeerJ Computer Science*, *11*, e3133.

Dada, I. D., Akinwale, A. T., & Tunde-Adeleke, T. J. (2025). A Structured Dataset for Automated Grading: From Raw Data to Processed Dataset. *Data*, *10*(6), 87.

Ahamed, A., Tarafder, M. T. R., Rimon, S. T. H., & Ahmed, N. (2025). Bidirectional Deep Learning and Extended Fuzzy Markov Model for Sentiments Recognition. *IECE Transactions on Neural Computing*, *1*(1), 11-29.

Ahamed, A., Tarafder, M. T. R., Rimon, S. T. H., & Ahmed, N. (2025). Bidirectional Deep Learning and Extended Fuzzy Markov Model for Sentiments Recognition. *IECE Transactions on Neural Computing*, *1*(1), 11-29.

Zhang, S. (2024, May). Study on Deep Learning-Based Personalised Product Recommendation Model for Autonomous Question-and-Answer Robot. In *The World Conference on Intelligent and 3D Technologies* (pp. 59-72). Singapore: Springer Nature Singapore.

Jie, Q., & Huang, C. (2025). A Dual-Path Gated Attention-Based Deep Learning Model for Automated Essay Scoring Using Linguistic Features. *International Journal of Advanced Computer Science & Applications*, *16*(7).

Devi, S. A., Ram, M. S., Dileep, P., Pappu, S. R., Rao, T. S. M., & Malyadri, M. (2025). Positional-attention based bidirectional deep stacked AutoEncoder for aspect based sentimental analysis. *Big Data Research*, *39*, 100505.

Siddalingappa, R., & Sekar, K. (2022). Bi-directional long short term memory using recurrent neural network for biological entity recognition. *IAES International Journal of Artificial Intelligence*, *11*(1), 89.

Azam, S., Azam, Z., Amjad, T., Ubaid, M., & Azam, M. (2025). Sentiment Analysis Of Music Reviews Using Deep Learning: A Bidirectional Lstm Approach. *Contemporary Journal of Social Science Review*, *3*(3), 1734-1751.

Dağkurs, B., & Atacak, İ. (2025). Deep learning-based novel ensemble method with best score transferred-adaptive neuro fuzzy inference system for energy consumption prediction. *PeerJ Computer Science*, *11*, e2680.

Palaniyappan, B., & Vinopraba, T. (2025). Grid-to-prosumer (G2P) interactions: Using bi-directional LSTM techniques to enhance the smart grid network through a demand response scheme. *Measurement: Energy*, 100067.