Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(2): 137-148 (2025)
Copyright © Pakistan Academy of Sciences
ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-2)694

Pakistan Academy of Sciences

Research Article

Transformer Based Essay Generation and Automatic Evaluation Framework

Israr Hanif[1], Zoha Latif[1], Fareeha Shafique[1], Humaira Afzal[1], and Muhammad Rafiq Mufti[2*]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan
[2] Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari, Pakistan

Abstract: The purpose of Automated Essay Grading (AEG) systems is to evaluate and assign grades to essays efficiently, thereby reducing manual effort, time, and cost. The traditional AEG system mainly focuses its efforts on extractive evaluation rather than abstractive evaluation. The objective of this research is to explore the differences in the grading system of traditional and grammar schools. This research develops a transformer-based system that combines extractive and abstractive essay generation and evaluation. We utilize the Bidirectional Encoder Representations from Transformers (BERT) model for extractive essay generation and Quillbot for abstractive paraphrasing, and design a framework that evaluates both types of essays. To achieve this objective, we created the Long Essay Poets (LEP) dataset and evaluated this across four modes using four models. We compare the performance of four models: Random Forest, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a combined approach of CNN and LSTM. After performing the experiment, it is concluded that 46% of grades declined in Mode 3 and 44% of grades improved in Mode 4, and in the context of essay evaluation, the Random Forest model performs better in extractive and merging scenarios, and the Long Short-Term Memory (LSTM) Model outperforms in abstractive essay evaluation.

Keywords: Transformer, Essay Generation, BERT Model, Natural Language Processing, Automatic Essay Grading.

1. INTRODUCTION

Automated Essay Grading (AEG) systems are designed to evaluate the quality of written text. The term Essay is defined as well-organized text written about a specific topic. The study in this domain started in 1966 and is still an active research domain. The first AEG system known as the Project Essay Grader (PEG) system is designed to assign grades based on different features including sentence structure, dictation, and grammatical mistakes [1]. The advancement in Natural Language Processing (NLP) made a greater contribution toward the generation and assessment of these essays using text-based features. Continuous effort has been made to design a fair and reliable evaluation systems which help in reducing manual labor, save time and lessen bias.

These AEG systems are comprised of numerous domains that involve Education, NLP, and Linguistics and are used to assign grades based on certain features. The advent of transformers had a significant impact on the generation and evaluation of text. The self-attention mechanism utilized by the transformer model addresses complex challenges such as machine translation, short-answer questioning and sentimental analysis. Transformers are renowned for handling long-term dependency in a text comprising of two parts: 1) encoder, and 2) decoder. The encoder utilized a self-attention mechanism that gives weight to each word. The decoder also implements this mechanism to generate an output token and assures the relevancy of the output sequence. Parallel computation, capturing the context of text, and maintaining the connection of sentences to generate meaningful

Received: April 2025; Revised: May 2025; Accepted: June 2025

* Corresponding Author: Muhammad Rafiq Mufti <rafiq_mufti@ciitvehari.edu.pk>

output made transformers more powerful and useful in the domain of NLP [2]. Transformer models are trained on huge corpus and are capable of solving complex problems. These models include Generative Pre-Trained Transformer (GPT), Bidirectional and Auto-Regressive Transformers (BART) and Bidirectional Encoder Representations from Transformers (BERT). BERT was introduced by Google AI and is trained on internet raw text as well as Wikipedia text and different versions of GPTs are presented by Open AI and it has 175 billion parameters that generate text creatively [3]. BART, on the other hand, mostly used for summarization tasks [4]. These models generate text but the nature of the generated text varies between the two types. One is abstractive and the other is extractive. In abstractive generation, the model generates the text from input text but it generates the text in their own-word keeping the context of the original text. It majorly focuses on important aspects of the entire text. On the other hand, extractive generation is used to extract text as it is without any modifications. These models worked in an extractive manner [5]. Mostly GPT is used for abstractive and the BERT model is used for extractive essay generation. The BERT model is known for Essay generation tasks and its bidirectional features help in understanding the context of sentences. The BERT model is notable for retaining the correlation of words in a sentence and having different sizes, parameters and layers [6]. The generated text is evaluated by two different ways. The first is exact, whereas the second is approximate. Exact evaluation means the written text is exactly matched with the referenced text or not. On the other hand, approximate evaluation is done when certain phrases and important words are matched with original text and it also uses own wording keeping the context of the referenced text in mind. This study aims to design a framework that evaluates both types of Essays. It also focuses on assessing the differences in the essay grading system of traditional and grammar school systems. For this, the Long Essay Poets (LEP) dataset is created. This dataset consists of both types of Essays having 100 instances and 8 features. To check the similarity of the referenced text with the generated text, we use different similarity measures that include Cosine similarity, Jaccard similarity, Longest-common sequence, and N-grams. The Essays are taken from Wikipedia, which is an open-source database that consists of articles. The essay is generated in response to the prompt. The majority of the research emphasizes a single type of Essay, which can be either abstractive or extractive. There is a scarcity of automatic evaluation systems that handle both types of Essays. For evaluation and assigning grades to Essays, different machine and deep learning algorithms will be implemented.

The first transformer model was introduced as a 'Fast weight controller' in the Early's 1990. Even though it captures association among words in a sentence, this model has some drawbacks and is not accepted worldwide [7]. Then Lu *et al.* [2] presented its model and named it transformer**,** defining the elements on which it depends. Its self-attention mechanism handled the biggest challenge in NLP by providing weightage to each word in a sentence and understanding its importance. This model helped in solving complex problems. One of the models known as BERT was introduced by Devlin *et al.* [6]. It is a pre-trained model based on Masked Language Model (MLM) and Next Sentence Prediction (NSP) making it a powerful model. The objective of MLM and NSP is to predict missing words in a sentence based on context. Apart from capturing sequence and connection in a sentence, BERT also captures contextual meanings in long sequence problem handling. Its vast application includes fine-tuning pre-trained models and parameter adjustment for text generation tasks. It also solves the challenges of text summarization, machine translation, and question answering [8]. Many researchers performed different experiments and generated Essays through the transformer BERT model. Qu *et al.* [9] used the BERT model and GPT-2 for text generation and automatic question-answering generation. This research presents a comparative study of these two models implemented on two large datasets named LLK and BaiduBaike having sizes of 848M and 5.3G. Analysis is performed to see the performance of these pre-trained models and it concludes that GTP-2 generates creative, novel and long sentences, whereas BERT model performs better in an extractive generation like in the case of question-answer generation. Another new approach presented by Chen *et al.* [10], is known as Conditional Masked Language Modeling (C-MLM) which fine-tunes the BERT model to generate more coherent text. This study provides a new dimension of the BERT model which is not only used for language understanding but also for generation task. IWSLT German-English and English-Vietnamese MT datasets are used in the experiment and this new approach outperforms generation tasks like machine translation and text summarization. The result shows that this approach generates a coherent and high quality that gives meaning properly. In another study of Essay generation, a new model was proposed by Lin *et al*. [11] called as PC-SAN model, which represents a Pretraining-Based Contextual Self- Attention Network. Topic Essay Generation used the PC-SAN model and generated high-quality essays that were informative and relevant to the topic. In the encoder layer, the BERT model is used to create Essays. It helps in maintaining the contextual and semantic meaning of sentences. Two Chinese corpus named ESSAY and ZhiHu are used to train the model. ZhiHu contains 50k training and 5k testing records and ESSAY comprises 300k training and 5k testing dataset. As an outcome, the PC-SAN model generates improved quality content and maintains topic consistency with the help of the BERT model. Another study conducted by Chan and Fan [12] for question generation was employed for a pre-trained BERT model. Two models were made by reconstructing the BERT model, in addition to utilizing the original BERT model. The BERT architecture consists of a multi-layer and bi-directional transformer. The three models developed were BERT-QG, BERT-SQG, and BERT-HLSQG. The SQuAD dataset is used for the training of the model. The BERT-HLSQG model performs well in question-generation tasks. Another goal of this study is to identify the difference between the original questions and the questions generated by the model which achieved higher performance on paragraph and sentence-level input. One of the language models is created known as AraGPT2 by Antoun *et al*. [13] for Arabic language. This study highlights the role of transformers in NLP and text generation, and also develops the N-grams language model (LM) which can handle a large corpus of text on the internet and 1.46b parameters of news and articles. A machine learning detector was also developed giving 98% accuracy. However, this model has some limitations as it can comprehended between human and machine written text.

On the other hand, evaluation is also an important and challenging task. The research started in 1960 and to date is an active and hot topic due to the presence of massive online courses, and many assignments are subjective based. So, one of the studies presents AES based on individual fairness which means ‘similar people should get similar treatment’ [14]. The dataset used for the AES system for individual fairness is the Automated Student Assessment Prize (ASAP). A total of 1569 responses from grade 7 students were collected each with 187 words. Sentence-BERT and LASER are used in this study instead of BOW and TF-IDF. The paraphrased Essay was analyzed for how well it maintains a similarity ranking. In terms of text representation, BERT and LASER are better and for scoring, gradient boosting is best according to this study. Another study presents short answer grading, a text-mining model proposed by Suzen *et al*. [15]. The distance between student response model sentences determines the completeness of the Essay. The important role is played by model vocabulary for both grading and feedback. A correlation of 0.81 is obtained from both responses. For the evaluation of automatic short answer scoring, the Kaggle short answer scoring dataset of approximately 10.000 is used [16]. This model is pre-trained on Word2Vec, Google news corpus, and Wikipedia dataset. Weightage is applied to each word in the first part and then word count, unique word, and average length are computed with statistical models. Quadratic Weighted Kappa (QWK) score of 0.78 is obtained from the random forest model and this model has application in diverse English and Science topics. An automated assessment system for the objective system can be easily developed to assess student-acquired knowledge but for the subjective system. It has limitations as it cannot compare correct answers semantically for automatic assessment and giving grades. To measure semantic similarity, an AG system was built by Hameed and Sadiq [17] in which input features like structural knowledge, and syntactical and sentence semantics were used in the Support Vector Machine (SVM) model to check similarity. This method’s performance surpassed the previous methods having root means square error of 0.83 and Pearson’s correlation coefficient of 0.63. Essay evaluation is not an easy task as it is time-consuming and several factors affect its accuracy. So, a new evaluation system was proposed by Ramamurthy, M. In this study, student written and reference text similarity is measured by the CS method and to generate document vector space, various methods are utilized. The dimensionality reduction technique is also used on document vector space along with new proposed synset-based word similarity model. The MSR paraphrase corpus and Li's benchmark datasets were used in this experiment. Kaggle short-answering dataset model performance is also evaluated [18].

For essay scoring, the merger of LSTM with transformer also presented by Johnsi and Kumar [32]. This study utilizes the sequential ability of the LSTM model with a transformer for an essay grading system. The ASAP-AES dataset is used in an experiment that contains 12k essays on eight different topics. To understand the context of long-generated essays and their relationships among words across sentences, the Bi-directional LSTM model is used with an attention pooling layer for generating vectors for each word. It uses a sigmoid function for classification and makes a comparison with other classification models. The evaluation metrics used in this study are the Mean Squared Error, which is used for checking regression quality, and QK for comparing scoring between humans and models. The combined approach yields a QWK of 0.86, and the model explicitly shows coherence and structure. Neural network approaches are also used in scoring Essays. Xia *et al.* [19] designed an automatic Essay scoring system. To get semantic and contextual meaning, the Bidirectional Long Short-term Memory (BLSTM) model is implemented on the ASAP dataset. The performance of different word embeddings was evaluated, and it was found out that Google Word2Vec was far better than standard Word2Vec method. It obtained a QWK score of 0.87, thus saving costs and minimizing the manual effort. The researcher emphasizes evaluating the answer to free-text questions. In order to address this, an intelligent auto-grading system has been designed. The BLSTM model is used to catch semantic meanings and create an attention layer to Essay to capture information correctly. The model is trained on the ASAP dataset available on Kaggle and focuses on critical words, avoiding unnecessary words, maintaining the logical semantics of sentences, and predicting grades. Skip-gram model is used as a word embedding and the Softmax function as an activation function to give weightage to each word and achieve a QWK score of 0.83 [20]. In another study, neural network architecture by using CNN and Long Short-Term Memory (LSTM) layer with word embedding as input, is designed by Riordan *et al.* [21] for evaluation of short answer scoring (SAS). A certain size of window features is extracted by the LSTM model and transferred to the aggregate layer, which chooses only accurate and crucial word windows. These models are applied to three different datasets named ASAP-SAS, Power-grading, and SRA dataset. The highest accuracy resulted in a QWK of 90%. Another unique concept of memory-augmented neural network is proposed for Essay evaluation. It comprises four layers. In these layers, the Essay is represented in vector form, assigned weights to words, and scored as output in the last output layer. Different models are implemented on the automated student assessment prize dataset and the best result is given by the LSTM+CNN model with an accuracy of 76% which is higher than previous studies [22]. Attention-based Essay scoring system is designed by Dong *et al.* [23] by using a combination of CNN+LSTM and the results showed that sentence level document model leads to be more effective in the case of long Essays. More importance is provided to the attention-based model by evaluating Essays. The output of layers is a sentence vector which assigns weights to each sentence and after the CNN layer, the attention layer of LSTM is attached to give scoring to each essay. The obtained results show that the model has an average QWK score of 0.764.

The contributions of this research work are as follows. A novel LEP dataset is formulated due to the absence of potential features and long Essays. Essay generation is done through transformer-based mode. A framework is designed that evaluates and assigns grades to both extractive and abstractive Essays. The aim is to explore the difference in grading system of the traditional and grammar school.

2. PROPOSED FRAMEWORK

The proposed framework is illustrated in Figure 1. This is the step-by-step approach for extractive and abstractive essay generation, and it applies different similarity metrics and implement different models for essay evaluation.

2.1. LEP Dataset

In the first stage of our study, we created a new dataset for solving the existing problems. Current datasets often have limitations, such as:

Short Texts: Unfortunately, most available datasets are made up of smaller text samples that are not

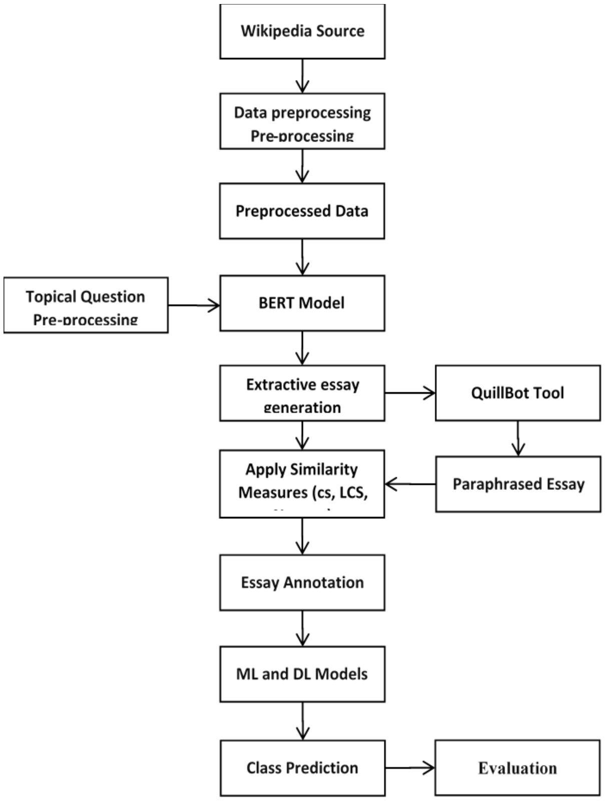


Fig. 1. Schematic presentation of proposed framework.

adequate for our purposes.

Language Barriers: Some datasets are collected in native languages and are not translated into English at all.

Feature Limitations: Many existing datasets are missing important features and do not contain both extractive and abstractive essays in one dataset.

In order to address these challenges and achieve the goals of the present research, we created a new dataset called the Long Essay Poets (LEP) Dataset. The source of the dataset is Wikipedia. As an open-source platform, Wikipedia provides a huge collection of articles of various domain and in various languages. It follows some structural guidelines and is written without discrimination of gender, religion, biasness, and religion. Wikipedia articles are reliable as one can check their validity by looking at the references and one can read talk links and discussion in case of wrong information was provided. The LEP dataset given in Table 1 has 8 features including id, Essay, question, answer, LCS, CS, N-Gram, and grade. Essays on Muslim

Table 1. Preview of LEP dataset.

ID	Essay	Question	Answer	LCS	CS	N gram	Grade
1	Syed Muhammad ibn Yousuf al-Hussaini (7 August 1321,10 November 1422), commonly known as Khwaja Banda Nawaz Gesudaraz, was a Hanafi Maturidi scholar and Sufi saint from India of the Chishti Order.Gaisu….	Write about Muhammad ibn Yousuf al-Hussaini?	In 10 November 1422, commonly known as khwaja banda nawaz gesudaraz, was a hanafi maturidi scholar and sufi saint…….	0.82	0.98	0.83	A
2	Ibrahim Qutb Shah Wali (1518 – 5 June 1580), also known by his Telugu names Malki BhaRama and Ibharama Cakravarti, was the fourth monarch of the kingdom of Golconda in southern India…….	Tell us about Ibrahim Qutb Shah Wali?	Raya. ibrahim is known for patronizing telugu extensively because he was moved by a genuine love for the language……	0.30	0.92	0.87	B
3	Khanzada Mirza Khan Abdul Rahim (17 December 1556 – 1 October 1627), popularly known as simply Rahim and titled Khan-i-Khanan, was a poet who lived in India during the rule of Mughal emperor Akbar……	Detail about Khan-i-Khanan?	Rahim was known for his hindustani dohe (couplets) and his books on astrology. abdul rahim was born in delhi, the son of bairam khan, akbars….	0.33	0.92	0.88	B+

poets are taken and we cover their background, education, their interests in poetry and details about written books. We named this feature of the dataset as 'Essay'. In our data preprocessing steps, we removed citations by using the Chrome extension 'Remove citation'. We also removed hyperlinks, and single and double quotes and eliminated other language words from the 'Essay' columns. The third column is about the question. Open-ended questions are raised against essays. Open-ended questions are those questions which are not specific but are generalized, and the answers are in detail-oriented manners. The next feature is answer which is either generated through the BERT model or paraphrased by using the paraphrasing tool. Different similarity metrics are utilized such as Cosine Similarity, Longest Common Subsequence, and N-Gram to calculate the similarity score and the last column is the 'grade' column. Grades are assigned based on predefined criteria set by our traditional and grammar school systems.

2.2. Extractive and Abstractive Essay Generation

For extractive essay generation, we used BERT, a transformer model that is currently one of the most sophisticated stances. The favourite feature of BERT is the understanding of context and the use of long-time dependencies in the sentence. To this end, we employed the BERT-base uncased model since it has relatively better performance and less computational overhead. In extractive essay generation, BERT understands the question, finds the coherent context of the answer within the referred Essay and returns the answers without any modification. For the abstractive Essay, we again used this generated Essay and rephrased it by using Quillbot. It changes the sentence structure from active to passive, gives synonyms, and rephrases the sentence while preserving the original meaning. We did this to see how the grading system of traditional and grammar schools is affected when we made certain changes in the answer. The Figure 2 depicts how BERT model works in extractive manner.

2.3. Similarity Metrics

Similarity metrics are used to measure the similarity between two sentences, paragraphs, and documents. In NLP, the most commonly used metrics are Cosine similarity, LCS, and N-Gram. Other forms of similarity matching algorithms are available but they cannot be used when matching short texts with long text. This is because our referenced answer is longer and the generated answer is not always equal in length to the referenced answer. So as a result, those metrics give too much lower scoring. Among the similarity metrics, we have Cosine similarity which is used to determine the angle of cosine between two vectors in the multi-dimensional space placing the text into the vector form before calculating the angle of similarity [24]. We have a higher score of similarity near to 1 when the angle of cosine is smaller. We also calculate the similarity score of the referenced essay with the generated essay using the longest common subsequence. It is used for the purpose of finding the longest segment between two sentences, based on sentence formation. The LCS is calculated by the total length of LCS to the geometric means of length of two sequences. The values closest to 1 are more similar and closest to 0 are less similar. The similarity score can be calculated by using N-gram feature. Thus N-gram contains n numbers of words in the sequence [25]. Using only one similarity metric is not suitable for all cases, therefore we combine Cosine, LCS, and N-Gram similarity into an average similarity score, and then have grades given depending upon the similarity score.

2.4. Traditional Grammar and School Grading System

The research work aims to identify the differences in the grading system of traditional and grammar schools. Our traditional school grading criteria are extractive in nature as higher similarity score would mean higher grade. On the other hand, grammar school obliged abstractive essays also. It also assigned higher grades if answer is relevant and is

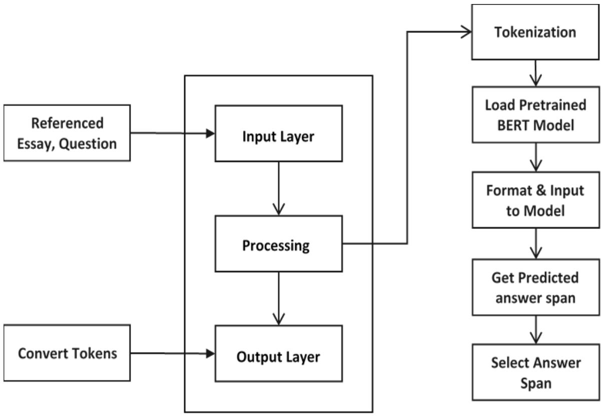


Fig. 2. Layout of the Essay generation through BERT.

from the given context although it is not exactly from the referenced essay. So, in our LEP dataset, the first 50 instances are generated from the BERT model in extractive manners and assigned grades based on predefined criteria set by the Punjab Board grading system and that is implemented in traditional schools. The last 50 rows are paraphrased by using Quillbot and thus assigned based on grammar school criteria given in Figure 3.

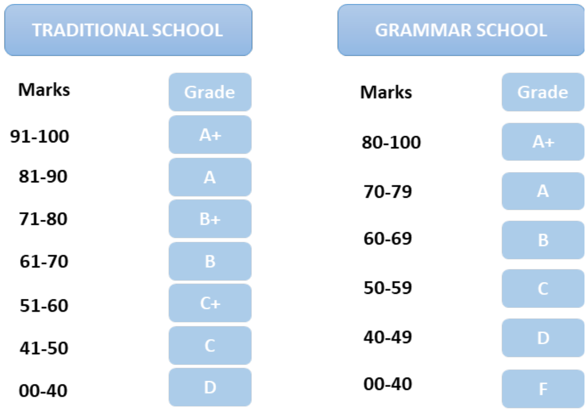
2.5. ML and DL Models for AGS

When output variable is categorical, the problem is said to be a classification problem. In the LEP dataset, labels or grades are already assigned. We implemented different models on the LEP dataset including random forest, CNN, LSTM and combined approach of CNN+LSTM.

2.5.1. Random forest model (RFM)

The RFM model is used to solve both classification and regression problems efficiently. During the training of the dataset, it makes multiple decision trees. By using the method of sampling with replacement, it makes a subset of the original dataset and hence solves the problem of overfitting. After making a subset and selecting important features, RFM model applies the ensemble technique of voting and gives predictions [26]. The experimental setup, including TfidfVectorizer() for word vectors and a train/test split, is 80/20 ratio. When RFM is implemented on LEP dataset, the following steps are followed:

- Data Preparation
- Feature Extraction (TF-IDF)
- Model Training
- Model Evaluation



TRADITIONAL SCHOOL

Marks	Grade
91-100	A+
81-90	A
71-80	B+
61-70	B
51-60	C+
41-50	C
00-40	D

GRAMMAR SCHOOL

Marks	Grade
80-100	A+
70-79	A
60-69	B
50-59	C
40-49	D
00-40	F

Fig. 3. Schematic presentation of essay grading criteria.

2.5.2. Convolutional neural network (CNN) model

The CNN model outperforms in many NLP applications. It is useful in identifying patterns and making predictions. This model consists of layers including convolutional layer, max pooling along a fully connected layer. Different activation functions such as Relu and Softmax along layers are used to determine objects, sentimental analysis, and text classification [27].When CNN is implemented on LEP dataset , several steps are followed as:
Input data → Feature extraction → CNN processing
Numeric feature processing → Prediction
The experimental setup includes a dropout rate of 0.5 with ADAM optimizer, batch size 16 and 20 epochs with cross entropy loss function, having 20% testing and 80% training data.

2.5.3. Long short-term memory model (LSTM)

An advanced version of the recurrent neural network is LSTM which handles long-term association in a sentence and retain time series data in memory. The LSTM model eliminates the issue of overfitting and vanishing gradient that comes in the RNN model. There are three gates in the LSTM model. First is the input gate used to input text then the forget gate which is used to keep relevant and important information and discard irrelevant information and last one is the output gate which presents output. Through these gates, we can keep and erase information. The most used activation functions in the LSTM model are sigmoid and Tanh which help in resolving the issue of long-term dependency [28]. The LSTM model is implemented with two layers, each containing 128 units, and a dropout rate of 0.5. The model is trained using the ADAM optimizer with a learning rate of 0.0001, over 20 epochs and a batch size of 32.

2.6. Evaluation Metrics

The performance of ML and DL models is measured using evaluation metrics. These metrics show how well a model performs on the given datasets, allowing us to choose the best model through comparative analysis.

2.6.1. Accuracy

Total correctly predicted instances out of total instances is termed as accuracy. It counts overall correct prediction, regardless of any class. It also evaluates the effectiveness of the classification model [29].

$$Accuracy = \frac{TCP}{TI}$$

where, TCP denotes total correct predictions and TI is total instances.

2.6.2. Precision

Precision measures how many predictions that model predicts positive are actually positive. It focuses on the accuracy of positive predictions [29]. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

where, TP represents true positive and FP denotes false positive.

2.6.3. Recall

Recall also known as sensitivity is calculated by identifying correctly positive instances out of the total instances of that class [29].

$$Recall = \frac{TP}{TP + FN}$$

where, FN represents false negative.

2.6.4. F1-score

F1-score is a balanced metric for both precision and recall. It is a harmonic mean of both metrics and is useful for uneven class distribution [29]. It is calculated as:

$$F1 - score = \frac{2}{1/precision\ + 1/recall}$$

3. RESULTS AND DISCUSSION

Four different models are implemented on the LEP dataset across four different modes. The analysis is divided into two parts: 1) Mode Setup, 2) Model-wise mode setup.

The LEP dataset is analyzed in four modes: 1) Evaluation of original instances in the context of Traditional School represented as Mode 1, 2) Evaluation of Paraphrased Instances in the context of Grammar School represented as Mode 2, 3) Merged instances in the context of the Traditional School represented as Mode 3, 4) Merged instances in the context of the Grammar School represented as Mode 4. The purpose of modes is to evaluate how the grades of abstractive and extractive essays are affected when grading criteria change. It also reflects the mindset of the examiner. In Mode 1, there are a total of 50 records. In these records, an answer column is generated from the BERT model, and grades are assigned based on traditional school criteria as given in Figure 3. There are also 50 records in Mode 2 in which the answer column is now rephrased by using Quillbot and assigned grades based on grammar school criteria. In Mode 3, there are a total of 100 records. In this mode, 50 records from the original and 50 records from paraphrased are combined and all these records are again evaluated on traditional school criteria. Similar to Mode 3, there are in 100 records in Mode 4 which are evaluated on grammar school criteria. As a result, 46% grades declined in Mode 3, and 44% grades improved in Mode 4. In all these modes, we calculate similarity scoring by using similarity metrics that are CS, LCS, and N-Gram. The values of similarity metric of Mode 1 are represented in Figure 4.

3.1. Model Wise Mode Evaluation

The Random Forest, CNN, LSTM and CNN + LSTM algorithms are applied on four different modes of LEP dataset.

3.1.1. RFM Evaluation

RFM Model is implemented by using four different modes and their performance is illustrated in Figure 5. Precision is higher as compared to accuracy F1-score and recall, which indicates good ability of the model to exclude almost all false positives. It means that our model carefully selected the features that are relevant to the context and thus improved precision. However, the recall is about 0.7, which suggests that some instances are not identified while accuracy and F1-score gives us a fair balanced performance [30]. With Mode 2, performance achieves much lower results in all metrics in comparison with Mode 1. According to the results measured by accuracy and precision, it is 0.4 and 0.15 respectively which shows a very poor performance and high imbalance toward

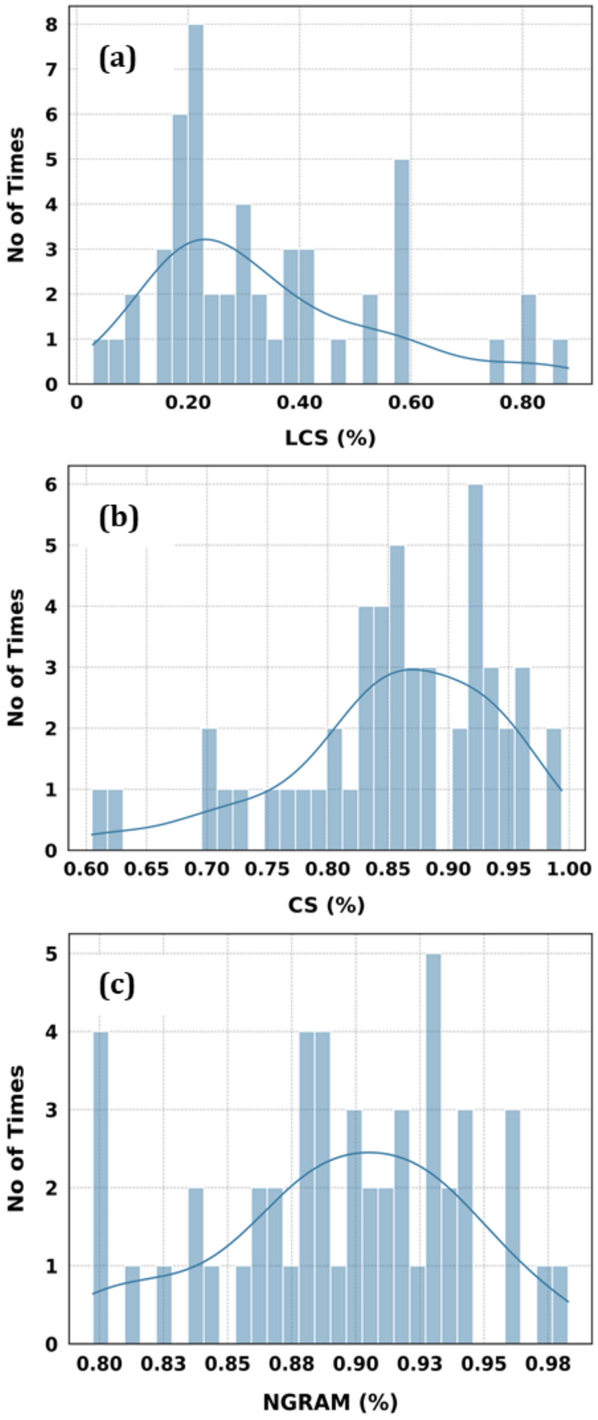


Fig. 4. Numeric features distribution: (a) LCS vs Number of Times, (b) CS vs Number of Times, and (c) Ngram vs Number of Times.

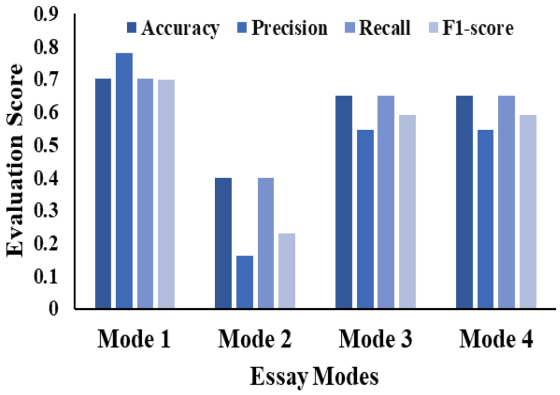


Fig. 5. RFM implementation by modes.

negative class. Recall is close to 0.4, which means moderate possibilities of identifying true positives only. As compared with Mode 2, performance increases significantly in Mode 3 and Mode 4. The accuracy gets to 0.65 which is the second best of all the modes. Precision and recall are 0.55 and 0.6, respectively, providing a better ratio of false positives per false negatives. We can therefore say that F1-score agrees with these values and thus it establishes balanced performance. From the results, it is evident that RFM performs well in Mode 1. It produces the best precision and moderate high accuracy as well as recall.

3.1.2. CNN Implementation

Table 2 show the performance of the convolutional neural network (CNN) in different modes. As for mode 1, CNN model demonstrates the highest specific accuracy and F1-score, which is the balanced measure of diversity that reflects both precision and recall. However, the moderate precision suggests the requirement to further decrease the number of false positives for increased accuracy. Unlike Mode 2, it was identified that the accuracy, precision, and F1-score were relatively lowest, implying the model had many false positive cases. Mode 3 and Mode 4 have very similar metrics, which are moderate accuracy and F1-

Table 2. Evaluation score through CNN.

S. No.	Mode	Accuracy	Precision	Recall	F1-score
1	Mode1	0.7	0.4899	0.7	0.5764
2	Mode2	0.4	0.16	0.4	0.2285
3	Mode3	0.5	0.25	0.5	0.3333
4	Mode4	0.5	0.25	0.5	0.3333

score, however, less precision indicates problem with reliability of predictions. From the results, it can be noted that the CNN model outperforms in Mode 1 it is because CNN is proficient in capturing local contextual patterns.

3.1.3. LSTM Outcome

The performance of a LSTM model is shown on all modes of the LEP dataset in Figure 6. According to Mode 1, the accuracy and the recall are greater than 0.4, this shows that its performance of correctly identifying positive cases is reasonably well. However, the less precision and F1-score indicating that the model has a tendency to generate a large number of false positives. On the other hand, Mode 2 exhibits the poorest performance, particularly in terms of precision and accuracy, due to its high misclassification rate and inability to efficiently capture positive cases. Mode 3 and Mode 4 have approximately the same accuracy and recall rates, even though the precision and F1 scores are somewhat lower; this means that there might be a small trade-off to the overall accuracy of the model regarding the management of the positive cases. Among them, Mode 3 is likely to be more balanced in such trade-offs and potentially more suitable for generalizing tasks.

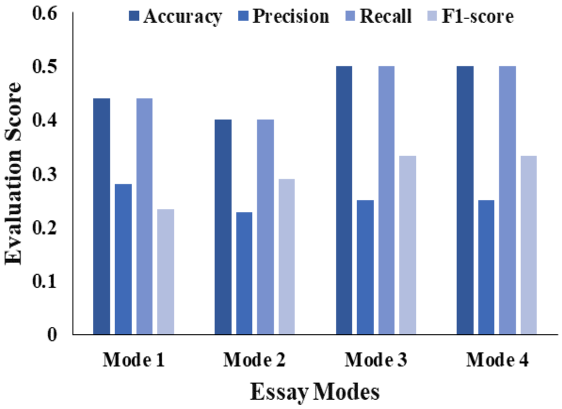


Fig. 6. LSTM outcome by Modes.

3.1.4. CNN + LSTM Implementation

The integration of CNN and LSTM architectures was proved to be efficient and provided very high performance in Mode 1. This is because the CNN model performs better in selecting the right contextual window span, while the LSTM model captures dependencies between sentences, which helps achieve higher accuracy compared to other models [31]. Table 3 shows a comprehensive picture of the model's performance across key evaluation parameters.

The final finding of our research is:

- By Performing mode analysis, it has been seen that 46% of grades are declined in Mode 3, and 44% of grades are improved in Mode 4.
- Among all models, the RFM model demonstrates better results on the LEP dataset.
- In the case of extractive essay evaluation and merging scenario, the RFM model gives higher accuracy, precision, and F1-score.
- The LSTM model outperformed other models in abstractive essay grading and achieved better results.
- Our proposed framework achieved higher accuracy in extractive essay grading as compared to abstractive.

4. CONCLUSIONS

The research work aims to design a framework that assigns grades on both types of essays and evaluate the difference in the grading system of traditional and grammar schools. To achieve the research objective, we designed the LEP dataset to generate essays through the BERT model and assigned grades based on predefined criteria. The performance of paraphrased essays is evaluated in traditional mode to measure how it affects the grade. The dataset is analyzed by four models across four modes and as a result RFM model achieved a higher accuracy of 70% in extractive essay and 65% in merging modes and the LSTM model achieved 40% in abstractive essay grading. In future work, we can enhance the

Table 3. Evaluation Score through CNN + LSTM.

S. No.	Mode	Accuracy	Precision	Recall	F1-score
1	Mode1	0.7	0.4899	0.7	0.5764
2	Mode2	0.4	0.16	0.4	0.2285
3	Mode3	0.5	0.25	0.5	0.3333
4	Mode4	0.5	0.25	0.5	0.3333

paraphrased part of the LEP dataset to achieve better performance in abstractive essay grading. We can also expand the scope of the LEP dataset by incorporating it into various domains.

5. CONFLICT OF INTEREST

The authors declare no conflict of interest.

6. REFERENCES

1. H.M. Alawadh, T. Meraj, L. Aldosari, and H.T. Rauf. An efficient text-mining framework of automatic essay grading using discourse macrostructural and statistical lexical features. *SAGE Open* 14(4): 1-14 (2024).
2. B.R. Lu, N. Haduong, C.Y. Lin, H. Cheng, N.A. Smith, and M. Ostendorf. Efficient encoder-decoder transformer decoding for decomposable tasks. *arXiv* 2403: 13112 (2024).
3. G. Yenduri, M. Ramalingam, G.C. Selvi, Y. Supriya, G. Srivastava, P.K. Maddikunta, G.D. Raj, R.H. Jhaveri, B. Prabadevi, W. Wang, and A.V. Vasilakos. GPT (generative pre-trained transformer) - a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access* 12: 54608-54649 (2024).
4. N. Delpisheh and Y. Chali. Improving faithfulness in abstractive text summarization with EDUs using BART (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence, (20th - 27th February, 2024), Vancouver, British Columbia, Canada* (2024).
5. W. Sun, C. Fang, Y. Chen, Q. Zhang, G. Tao, Y. You, T. Han, Y. Ge, Y. Hu, B. Luo, and Z. Chen. An extractive-and-abstractive framework for source code summarization. *ACM Transactions on Software Engineering and Methodology* 33(3): 75 (2024).
6. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *(3rd - 5th June 2019), Minneapolis, Minnesota, USA* (2019).
7. I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. *International Conference on Machine Learning, (18th - 24th July 2021), Vienna, Austria* (2021).
8. M.V. Koroteev. BERT: A review of applications in natural language processing and understanding. *arXiv* 2103: 11943 (2021).
9. Y. Qu, P. Liu, W. Song, L. Liu, and M. Cheng. A text generation and prediction system: pre-training on new corpora using BERT and GPT-2. *10th IEEE International Conference on Electronics Information and Emergency Communication, (17th-19th July 2020), Beijing, China* (2020).
10. Y.C. Chen, Z. Gan, Y. Cheng, J. Liu, and J. Liu. Distilling knowledge learned in BERT for text generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (5th - 10th July 2020), Washington, USA* (2020).
11. F. Lin, X. Ma, Y. Chen, J. Zhou, and B. Liu. PC-SAN: Pretraining-based contextual self-attention model for topic essay generation. *KSII Transactions on Internet and Information Systems* 14(8): 3168-3186 (2020).
12. Y.H. Chan and Y.C. Fan. A recurrent BERT-based model for question generation. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, (4th November 2019), Hong Kong, China* (2019).
13. W. Antoun, F. Baly, and H. Hajj. AraGPT2: Pre-trained transformer for Arabic language generation. *6th Arabic Natural Language Processing Workshop, (19th April 2021), Kyiv, Ukraine* (2021).
14. A. Doewes, A. Saxena, Y. Pei, and M. Pechenizkiy. Individual fairness evaluation for automated essay scoring system. *15th International Conference on Educational Data Mining, (24th - 27th July 2022), Durham, United Kingdom* (2022).
15. N. Süzen, A.N. Gorban, J. Levesley, and E.M. Mirkes. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science* 169: 726-743 (2020).
16. A. Kumar, M. Sharma, and R. Singh. Automatic question-answer pair generation using deep learning. *3rd IEEE International Conference on Inventive Research in Computing Applications, (11th - 13th July 2021), Coimbatore, India* (2021).
17. N.H. Hameed and A.T. Sadiq. Automatic short answer grading system based on semantic networks and support vector machine. *Iraqi Journal of Science* 64(11): 6025-6040 (2023).
18. M. Ramamurthy and I. Krishnamurthi. Design and development of a framework for an automatic answer evaluation system based on similarity measures. *Journal of Intelligent Systems* 26(2): 243-262 (2017).
19. L. Xia, J. Liu, and Z. Zhang. Automatic essay scoring model based on two-layer bi-directional long-short term memory network. *3rd International Conference*

on Computer Science and Artificial Intelligence, (6th-8th December 2019), Beijing, China (2019).

20. Z. Wang, J. Liu, and R. Dong. Intelligent auto-grading system. *5th IEEE International Conference on Cloud Computing and Intelligence Systems, (23rd - 25th November 2018), Nanjing, China* (2018).
21. B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee. Investigating neural architectures for short answer scoring. *12th Workshop on Innovative Use of NLP for Building Educational Applications, (8th September 2017), Copenhagen, Denmark* (2017).
22. S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan. A memory-augmented neural model for automated grading. *4th ACM Conference, L@S 2017, (20th - 21st April 2017), Cambridge, MA, USA* (2017).
23. F. Dong, Y. Zhang, and J. Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. *21st Conference on Computational Natural Language Learning, (3rd - 4th August 2017), Vancouver, Canada* (2017).
24. S. Sheng, J. Jing, Z. Wang, and H. Zhang. Cosine similarity knowledge distillation for surface anomaly detection. *Scientific Reports* 14(1): 8150 (2024).
25. D. Jurafsky and J.H. Martin (Eds.). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition. *Prentice Hall*, *Pearson Education, India* (2000).
26. N. Jain and P.K. Jana. LRF: A logically randomized forest algorithm for classification and regression problems. *Expert Systems with Applications* 213(18): 119225 (2023).
27. M. Krichen. Convolutional neural networks: A survey. *Computers* 12(8): 151 (2023).
28. H. Alizadegan, B. Rashidi, A. Radmehr, H. Karimi, and M.A. Ilani. Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction. *Energy Exploration & Exploitation* 43(1): 281-301 (2025).
29. A. Sbei, K. ElBedoui, and W. Barhoumi. Assessing the efficiency of transformer models with varying sizes for text classification: A study of rule-based annotation with DistilBERT and other transformers. *Vietnam Journal of Computer Science* 2024: 1-28 (2024).
30. H. Wang, Q. Liang, J.T. Hancock, and T.M. Khoshgoftaar. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data* 11: 44 (2024).
31. R. Nallapati, B. Zhou, C.N. dos Santos, Ç. Gülçehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. *arXiv:1602.06023* (2016).
32. R. Johnsi and G.B. Kumar. Enhancing automated essay scoring by leveraging LSTM networks with hyper-parameter tuned word embeddings and fine-tuned LLMs. *Engineering Research Express* 7(2): 025272 (2025).