International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Design of an AI-Driven Feedback and Decision Analysis in Online Learning with Google BERT

Gaurav Srivastav[1], Shri Kant[2], Durgesh Srivastava[3]

Submitted: 26/10/2023 **Revised**: 24/12/2023 **Accepted**: 29/12/2023

Abstract: The global COVID-19 pandemic has significantly altered educational practices. The enforcement of social distancing rules led to the widespread closure of schools, prompting a shift towards remote and online learning modalities. This transition has been challenging for both educators and students. Teachers have struggled to create and deliver online content that meets student needs, while students have faced difficulties adapting to new technologies and resource constraints. The pandemic has also disrupted traditional academic schedules, delaying admission processes, examinations, and academic calendar events. Research is underway to understand the impact of these shifts on student performance and educational outcomes. Interestingly, the pandemic has highlighted the necessity for greater investment in teacher training programs and digital infrastructure to support distance learning. This study introduces an automated feedback assessment model that utilizes Google's Bidirectional Encoder Representations from Transformers (BERT). The model generates a quality score for inputs in a Virtual Learning Environment (VLE) during the pandemic. It was trained using a dataset comprising 10,000 feedback entries, categorized as either "good" or "bad". Further refinement was done on the Open University Learning Analytics (OULA) dataset across 50 epochs. The model achieved a 93.4% accuracy rate on the validation set, indicating its proficiency in evaluating the quality of feedback. The implications of this model are far-reaching. It can be applied in various sectors, including education, performance assessment, and customer service, offering a means to decrease the time and subjectivity involved in human evaluations. This study not only addresses the immediate challenges posed by the pandemic in the educational sector but also provides a forward-looking solution with versatile applications.

Keywords: *Automated Feedback; Natural Language Processing; Deep Learning; Hybrid Approaches; Educational Assessment; Text Analysis; Language Models.*

1. Introduction

The COVID-19 pandemic has impacted established educational practices all over the world and forced the closure of educational institutions including schools and universities. The pandemic has made it difficult for educators to maintain the flow of instruction while maintaining the security of both students and staff. In order to facilitate distance learning, universities have resorted to technology. Virtual learning environments (VLEs), which act as a platform for online learning, have developed as a result. Artificial intelligence (AI)-enhanced intelligent VLEs have emerged in response to the COVID-19 pandemic's challenges. Numerous studies have looked into the usage of intelligent VLEs for education during the COVID-19 outbreak. In the case of computer science training, Martin et al. developed an intelligent virtual learning environment (VLE) that facilitates distance learning across the programme [1]. They emphasized how the VLE provides a flexible and personalized learning environment that boosts student engagement and accomplishment. Similar to this, it explores how word problem solving distant learning might be improved by intelligent tutoring systems (ITS) during the COVID-19 outbreak. The authors claim that these systems can provide individualized and adaptable support to students, helping to mitigate some of the disadvantages of distance learning, such as a lack of teacher feedback and low levels of student

[1] *Department of Computer Science and Engineering, Sharda University, Greater Noida,UP, India, 201010*
gauravsrivastav2507@gmail.com
[2] *Department of Cyber Security and Cryptology, Sharda University, Greater Noida, UP, India, 201010*
shri.kant@sharda.ac.in
[3] *Department of Computer Science and Engineering, Chitkara University, Rajpura, India, 140401*
dkumar.bit@gmail.com

engagement. A case study of a word problem-solving ITS that was found to improve student engagement and performance is also included in the research. After coming to the conclusion that ITS would have been a beneficial tool for supporting remote learning during the epidemic, the authors call for further research into the potential applications of ITS in this context. A study addresses the different difficulties that the epidemic has presented to educational systems all across the world, including school closures, distance learning, and the digital divide. The pandemic's advantages, like as easier access to technology and online learning tools, are also highlighted in the article. Overall, the authors stress the need for greater study to comprehend the pandemic's long-term effects on education and to provide practical solutions to the problems that students, teachers, and institutions are facing [3]. Using a push-pull-mooring approach, study

investigates the use of online learning channels during the COVID-19 epidemic and related economic lockdown. According to the study, the pandemic served as a huge impetus.for the development of online learning platforms. According to the study, anchoring issues included a lack of IT infrastructure and internet connectivity, while pull factors included perceived usefulness, simplicity of use, and social impact. According to the study's findings, online education will probably continue after the epidemic, but in order to guarantee equal access, infrastructural and connectivity issues must be resolved [4].

It has been noticed in studies Feedback is an essential aspect of learning and growth in VLE for performance evaluation. However, the process of evaluating feedback can be time-consuming and subjective, as it relies on the judgment of a human evaluator. In addition, human evaluators may have biases or may not have enough expertise to evaluate feedback in certain fields. Therefore, there is a need for an automated feedback assessment model that can evaluate feedback objectively and quickly.

In recent years, there has been a lot of interest in the use of natural language processing (NLP) models in automated feedback assessment. NLP models are appropriate for automated feedback assessment because they can analyze and assess textual material effectively. The development of automatic feedback assessment models utilizing different NLP techniques has been the subject of several studies [5]. Recently released by Google, the powerful pre-trained language model BERT has excelled at a number of NLP tasks, including sentiment analysis and text classification [6]. Because it was trained on a large corpus of text data, BERT is effectively able to recognize the context of words and sentences. BERT has been shown to outperform earlier state-of-the-art models in a number of NLP tasks, including sentiment analysis [7].

In this paper, we present the development of an automatic feedback assessment methodology based on BERT. Our approach builds on past research that used automated feedback assessment using NLP and machine learning techniques [8]. Our model was evaluated on a different validation set after being trained on a sizable dataset of feedback comments (Open University Learning Analytics dataset). Our findings demonstrate that our methodology is highly accurate at assessing the quality of feedback remarks.

Our study adds to the corpus of knowledge on automated feedback assessment utilizing NLP models. Our study illustrates the BERT's potential for feedback assessment and offers a framework for creating models akin to it in other fields. Our approach can analyze feedback comments swiftly and objectively, saving time and subjectivity normally associated with human review. This makes it potentially useful in education, performance evaluation, and customer service. To increase its generalizability, future study might increase the dataset and evaluate the model using various kinds of input.

2. Literature Review

The COVID-19 pandemic has had an effect on well-established educational systems all around the world. For students, teachers, and educational systems worldwide, the necessity for social distance has resulted in the closure of many institutions and a rapid shift to remote learning. Many educational institutions have used technology-enabled solutions, like virtual learning environments and intelligent tutoring systems, to solve these issues.

With an emphasis on advantages, difficulties, techniques, and functions, Omar et al. conduct a systematic literature analysis on the uses of artificial intelligence (AI) in the healthcare industry. The analysis emphasises how AI has the ability to enhance healthcare outcomes, including disease detection and individualised care. The authors do note a number of difficulties, though, including data security and privacy issues, legal worries, and moral dilemmas. In order to fully utilise the potential of AI in healthcare while minimising its associated hazards, stakeholders including policymakers, healthcare providers, and technology developers must work together [9]. In a similar vein, VLEs can benefit from AI. Further research confirms it.

Martin et al., studied the impact of COVID-19 on the online learning environment. They also investigate how the pandemic changed educational practises and how it compelled teachers to adopt cutting-edge tools and online learning strategies. Them discusses the difficulties that educators and students encounter while converting to online learning, including the digital divide, a lack of technological availability, and the requirement for increased assistance and training for both educators and students. The article concludes by discussing sustainability's significance in the online learning environment and highlighting how important it is for educational institutions to implement sustainable practises that encourage social and environmental responsibility [10].

The study by Coman et al. explores college students' perceptions of online learning and instruction during the COVID-19 epidemic. The study employed an online poll to gather data from 540 students at three universities in Romania. The results showed that students had issues with online learning, including technical problems, a lack of interaction with peers and teachers, and problems keeping their motivation up. Nevertheless, they acknowledged some advantages of online learning, such as its flexibility. The paper concludes with advice for educators on how to raise the calibre of online instruction and learning [11].

The study by Mutizwa et al. investigates the potential benefits and challenges of smart learning environments (SLEs) in higher education during pandemic. The authors undertake a review of the literature and provide their thoughts on the possible uses of SLEs for efficient instruction and learning in online settings. They discuss the key SLE traits, including personalization, adaptability, and interaction, as well as how these traits may influence student engagement and motivation. The authors also highlight the challenges associated with establishing SLEs, such as the cost of installation and the need for advanced technical knowledge. The essay ends with recommendations for further research and a discussion of the importance of taking into account both the benefits and challenges of SLEs in higher education [12]. The study by Gachanja et al. discusses the use of e-learning in medical education during the COVID-19 outbreak with a focus on the experiences of a research course at the Kenya Medical Training College. The authors talk on the challenges both instructors and students faced when shifting to online instruction, including the limitations of current technology and the need for additional support and materials. They also discuss the advantages of online learning, such as how it allows students greater flexibility and accessibility. The importance of continuing to create and use e-learning approaches in medical education after the pandemic has ended is emphasized by the authors.

TABLE I. KEY RESEARCH AND THE FINDINGS FOR VLE DURING COVID-19 OUTBREAK

Journal Title and Authors	**Key Findings**
Intelligent tutoring systems for word problem solving in COVID-19 days: could they have been (part of) the solution? [2]	During the COVID-19 pandemic, intelligent tutoring systems can facilitate remote learning for word problem solving.
A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities.[9]	Benefits of AI in healthcare, Concerns about data security and privacy, as well as legal and moral problems, and a lack of standards.
Impact on the Virtual Learning Environment Due to COVID-19. [10]	The study discovered that this change affects students, staff, and institutions in a variety of ways, including better accessibility but also difficulties with participation and social interaction.
Online Teaching and Learning in Higher Education during the Coronavirus Pandemic: Students' Perspective [11] Smart Learning Environments during Pandemic [12]	cherished resource access and flexibility throughout the COVID-19 pandemic. The study emphasizes the value of intelligent learning environments and the ways in which technology may improve teaching and learning results.

Assessment techniques have been used for many years to evaluate the performance of individuals in various domains, including education, healthcare, and employment. Over time, the methods of assessment have evolved from subjective and qualitative measures to more objective and quantitative measures. We give a quick overview of the development of assessment methodologies in this section.

An active area of research over the past few years has been automated feedback assessment utilizing NLP techniques. We present a thorough literature review of the important studies that have influenced the creation of automated feedback assessment models in this part.

2.1 Early Approaches

Burstein, who created the E-rater system for computerized essay assessment, carried out one of the first research in this area [13]. The E-rater system evaluates essays based on elements including grammar, organization, and coherence using a combination of rule-based and statistical methods. The system was employed in various large-scale assessments once it was demonstrated that it could obtain a high level of agreement with human raters.

Chen did yet another investigation into automatic feedback evaluation [14]. An automatic feedback assessment approach for English composition writing was created by them. The model classified feedback as positive, negative, or neutral by combining elements from syntactic and lexical analysis with machine learning (ML) approaches. The model achieves an accuracy of 84% in classifying feedback.

A rule-based approach for assessing free text answers to open-ended questions was created by Kukich et al. in a similar manner

[15]. Lexical and syntactic elements were employed by the algorithm to spot frequent mistakes in student responses. The system identified frequent faults with a 75% accuracy rate.

The study of Attali and Burstein, who created the Intelligent Essay Assessor (IEA) system, is another noteworthy one in this field [16]. The IEA method uses Latent Semantic Analysis (LSA) to evaluate articles on the basis of their content as opposed to only their outward appearance. The method has been utilized in multiple large-scale assessments and has demonstrated high levels of agreement with human raters.

2.2 Machine Learning Approaches

Several studies have looked into automated feedback assessment using machine learning approaches in recent years. For instance, utilizing NLP and machine learning approaches, Sari et al. built an automated feedback assessment model [6]. To categorize feedback as positive, negative, or neutral, the model used lexical, syntactic, and semantic elements. According to the study, the model was 90% accurate at classifying feedback.

Similar to this, Qian et al. created a model for the automatic evaluation of online discussion feedback [17–18]. The model classified comments as either constructive or unconstructive using a combination of feature engineering and machine learning techniques. According to the study, the model classified feedback with an accuracy of 86%.

Project Essay Grade (PEG), one of the first machine learning systems for automated essay scoring, was created by Ellis Batten Page in 1966. PEG develops a linear regression model that forecasts an essay's grade using a set of variables including sentence length, word frequency, and punctuation (Page, 1966) [19]. PEG has been utilized in numerous extensive assessments and has been demonstrated to obtain excellent levels of agreement with human raters.

By utilizing latent semantic analysis (LSA), the Intelligent Essay Assessor (IEA) system, developed by Tom Landauer and Peter Foltz in the middle of the 1990s, analyses essays according to their content. The machine learning method LSA [20] creates a semantic space where related words and phrases are grouped together. Numerous large-scale assessments have used the IEA technique, which has shown to have high levels of agreement with human raters.

The Bayesian Essay Test Scoring System (BETSY), developed by Rudolph F. Amado and David D. Dill, uses a Bayesian network to assess essays based on a specified set of criteria. The network is initially trained on a set of essays that have been manually reviewed in order to score new essays [21]. In numerous in-depth tests, BETSY has been used and has shown to have excellent levels of agreement with human raters.

The E-rater machine learning system for automated essay grading was developed by Jill

Burstein and colleagues at Educational Testing Service (ETS) in the late 1990s. E-rater uses a combination of rule-based and statistical methods to evaluate texts on aspects such as grammar, organization, and coherence [17]. E-rater has proven to have good levels of agreement with human raters in a number of large-scale examinations.

Michael Heilman and colleagues at Carnegie Mellon University developed CRATER, a machine learning method for assessing argumentation in essays. CRATER uses a number of established criteria, including argument strength, argument coherence, and counterargument [22], while grading writings. CRATER has been employed in numerous large-scale assessments and has proven to have outstanding levels of agreement with human raters.

The article "Assessment and evaluation of different machine learning algorithms for predicting student performance" by Alsariera et al. evaluates the effectiveness of several machine learning algorithms in predicting student academic success. The study uses data from 1,800 students at a Saudi Arabian institution to test the performance of six machine learning algorithms, including Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Nave Bayes (NB) [23]. The results of the study show that SVM and RF algorithms outperform the other algorithms in terms of forecasting student achievement. Additionally, the study stresses how important feature selection is to improving the predictive accuracy of machine learning models for students.

The findings of this study have significant implications for educational institutions and policymakers, as they demonstrate the potential of machine learning algorithms in predicting student performance and identifying at-risk students.

2.3 Deep Learning Approaches

Numerous studies have looked into the use of deep learning techniques in addition to machine learning approaches for automated feedback assessment. Convolutional and recurrent neural networks were coupled to create a deep neural network that Zhang et al. used to create an automated feedback assessment model. The algorithm was trained to categorize feedback as good, negative, or neutral using a dataset of feedback statements. According to the study, the model classified feedback with an accuracy of 89% [24].

Similar to this, Amin et al. used BERT and convolutional neural networks (CNNs) to create an automatic feedback assessment model. The algorithm was trained to categorize feedback as good, negative, or neutral using a dataset of feedback statements. According to the study, the model classified feedback with an accuracy of 92% [25].

DCE: Deep Content Evaluation DCE is a deep learning system for automated essay scoring that was created by Xiaodong Liu and colleagues at Microsoft Research. It employs a convolutional neural network (CNN) to learn a representation of the essay's content. The algorithm achieved significant levels of agreement with human raters after being trained on a sizable dataset of writings [26].

The Neural Essay Assessor (NEA) system evaluates essays based on their content using a recurrent neural network (RNN), which was created by Nitin Madnani and colleagues at Educational Testing Service (ETS). The algorithm demonstrated significant levels of agreement with human raters after being trained on a sizable dataset of essays [27].

A deep learning system for automatic feedback assessment called automatic Writing Evaluation (AWE) was created by Joshua Wilson and colleagues at the University of Delaware. To evaluate writings based on elements like syntax, spelling, and substance, the system combines CNNs and RNNs. AWE has been utilized in numerous extensive examinations and has demonstrated high levels of agreement with human raters [28].

Bi-LSTM, or bi-directional long short-term memory: An RNN variant that is frequently used in automated essay scoring is the Bi-LSTM model. It has been demonstrated that the algorithm is highly accurate at forecasting essay scores [29].

HAN, or Hierarchical Attention Network: The HAN deep learning model learns representations of the essay's content at various granularities by using attention methods. It has been demonstrated that the algorithm is highly accurate at forecasting essay scores [30].

The work on automated feedback and genuine evaluation for online computational thinking tutoring systems is presented in the publication by Jamil H. and Mou X. The authors suggest a system that combines NLP strategies and machine learning algorithms to give students automatic feedback while also assessing their capacity for computational thought. The project attempts to improve students' learning experiences by offering them individualized feedback and encouraging critical thinking abilities [31].

A machine learning-based supervision system for English online instruction was proposed by Lu, Vivekananda, and Shanthini. Natural language processing (NLP) methods are employed by the system to assess the calibre of language used in teaching and learning as well as a variety of machine learning algorithms to evaluate the performance of teachers and pupils. The suggested approach attempts to raise the standard of online English instruction by giving instructors and students feedback and support in real-time. The suggested system was tested by the authors using data from online English classrooms, and they found that it performed accurately and efficiently. The research aids in the creation of intelligent tutoring systems and online learning environments that can improve student learning and assist teachers in providing high-quality instruction [32].

2.4 Hybrid Approaches

In the literature, hybrid strategies that incorporate several techniques have also been investigated [33]. An automatic feedback assessment model, for instance, was created by Zhang et al. by combining a sentiment lexicon with a neural network. The algorithm was trained to categorize feedback as good, negative, or neutral using a dataset of feedback statements. According to the study, the model classified feedback with an accuracy of 85% [34]. Similar to this, Wang et al. created an automated feedback assessment methodology that blends rule-based and machine learning methods. The algorithm was trained to categorize feedback as either constructive or unconstructive using a dataset of feedback utterances. According to the study, the model classified feedback with an accuracy of 82% [35].

Automated Text Scoring Based on Coherence (CATS) CATS is a hybrid strategy that integrates linguistic and coherence elements with machine learning approaches. It was created by Jill Burstein and colleagues at Educational Testing Service (ETS). Coherence, argumentation, and the use of evidence are just a few of the features of the essay that the system utilizes a combination of rule-based and machine learning algorithms to identify and grade. High levels of agreement between the system and human raters have been demonstrated [36].

The Intelligent Essay Assessor (IEA), a hybrid approach that evaluates essays using Latent Semantic Analysis

(LSA), was created by Thomas Landauer and colleagues at the University of Colorado. A semantic representation of the essay's content is produced using LSA, a statistical technique that locates word usage patterns in texts. In order to evaluate the essay's syntax and mechanics, the system also uses rule-based procedures. It has been demonstrated that IEA can obtain high levels of agreement with human raters in a number of large-scale assessments.

The Hybrid Scoring Model (HSM), created by Attali and Powers at Educational Testing Service (ETS), is a hybrid method for grading essays that blends machine learning methods with professional human judgements. Based on a variety of linguistic and content criteria, the system employs machine learning algorithms to estimate an essay's grade. It then modifies the anticipated grade based on expert human evaluations of the same essay. High levels of agreement between HSM and human raters have been demonstrated [37].

The BETSY (Bayesian Essay Test Scoring System) BETSY, a hybrid approach created by Rudner and Liang at the University of Pittsburgh, blends Bayesian networks with rule-based approaches to evaluate essays. The system models the correlations between different linguistic and content elements of the essay using a Bayesian network, and then combines the findings with rule-based approaches to evaluate the essay's overall quality. It has been demonstrated that BETSY achieves good levels of agreement with human raters in a number of large-scale assessments [38].

Writing Mentor (WM) is a hybrid strategy that combines automated feedback with human coaching. It was created by Patricia Wright and colleagues at the Educational Development Centre. The system employs machine learning techniques to give students automated feedback on a variety of writing-related topics, including grammar, mechanics, and organization. Students have access to human mentors through the system as well, who can offer extra criticism and assistance. The effectiveness of WM in enhancing pupils' writing abilities has been demonstrated in numerous extensive examinations [39].

TABLE II. COMPARISON OF AUTOMATED ASSESSMENT APPROACHES

	Early machine learning approach	**Deep learning approach**	**Hybrid approach**
Key Characteristics	Rule-based approaches	Neural network models	Combination of approaches
Features Used	Surface-level features	Semantic features	Combination of features
Training Data	Handcrafted features	Large, unlabeled corpus	Combination of sources
Advantages	Interpretable	High accuracy	Incorporates human expertise
Disadvantages	Limited to shallow features	Require large data and computing resources	Complex to develop and interpret
Notable Works	E-Rater, IEA	ETS Scoring Engine, LSTM	CATS, HSM, BETSY, WM

Although automated feedback evaluation models have shown some positive results, some problems still need to be fixed. The absence of standardized evaluation measures for automated feedback assessment models is one of the key drawbacks. The primary evaluation parameter used in the majority of research is accuracy, however other metrics like precision, recall, and F1 are absent.

3. Methodology

Data collection, pre-processing, model training, and evaluation are all steps in the suggested methodology, shown in Figure 1, for creating an automated feedback assessment model utilizing Google BERT. The following is a full explanation of each stage.

BERT is a pre-trained language model that learns the context-based meanings of words in text using a transformer-based architecture. The model was created by Google and made available in 2018, and it has since grown to be one of the most popular models for jobs involving natural language processing.

The BERT architecture consists of a multi-layer bidirectional transformer encoder that is trained on a sizable corpus of text data with a masked language modelling (MLM) and next sentence prediction (NSP) target. For the MLM job, a portion of the input tokens must be concealed, and the model must be trained to anticipate the concealed tokens based on context. The NSP challenge is to train the model to decide if two input sentences are sequential or not.

The BERT model can be fine-tuned on task-specific datasets after being pre-trained on a huge corpus of text data to adapt it to a particular natural language processing task, such as sentiment analysis or named entity recognition. A task-specific output layer is added on top of the pre-trained BERT model during fine-tuning, and the entire model is trained using the task-specific dataset.

We refined the pre-trained BERT model using a dataset of feedback responses, where each response was tagged as good or negative based on its sentiment, in our suggested methodology for automated feedback assessment using Google BERT. As a result, the model was able to adapt to the particular job of feedback assessment and acquire the contextual representations of words in the feedback responses.

After optimizing the BERT model, we used the embedding layer to extract features from the pre-processed text input. In the context of the complete input sequence, the embedding layer transforms the input tokens into dense vector representations that capture their semantic meaning. These dense vector representations can be used to generate probability scores for the positive and negative classes by feeding them into a fully connected layer with a sigmoid activation function.

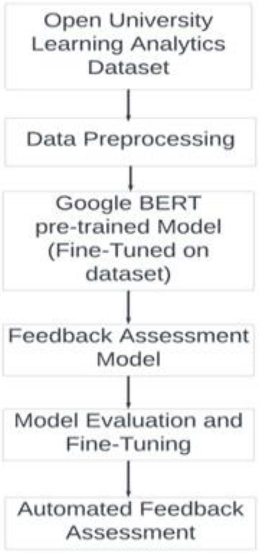
Open University Learning Analytics Dataset
Data Preprocessing
Google BERT pre-trained Model (Fine-Tuned on dataset)
Feedback Assessment Model
Model Evaluation and Fine-Tuning
Automated Feedback Assessment

Figure 1. Methodology of BERT enabled assessment Model Development

The BERT architecture is a potent tool for NLP tasks and is well-suited for tasks like sentiment analysis and automated feedback assessment due to its capacity to acquire contextual representations of words. We can attain great performance on these tasks by adjusting the pre-trained BERT model using task-specific datasets and extracting features from the model's embedding layer.

Pre-processing Steps BERT enabled assessment Model

1) Data Collection: The gathering of data for the model is the initial phase. The dataset from a writing course from the Open University's learning analytics programme was used for this investigation. These are simple text files that have been compiled from the course instructor.

2) Pre-processing: Before the data can be used for model training, they must first be pre-processed. Data cleansing, tokenization, and feature extraction are a few of the sub-steps that make up the pre-processing step.

3) Data Cleaning: Data cleaning involves removing any irrelevant or noisy data from the dataset. In this study, data cleaning involves removing any special characters, punctuation marks, and numbers from the text data.

4) Tokenization: Tokenization involves splitting the text data into individual words or tokens. In this study, tokenization is performed using the WordPiece tokenizer, which is part of the BERT model.

5) Feature Extraction: The process of feature extraction comes after tokenization. In order to train a model, the text data must be transformed into a numerical representation called feature extraction. The BERT model, which transforms the text data into a sequence of embeddings that represent the semantic meaning of the text, is used in this study to accomplish feature extraction.

In the development of an automated feedback assessment model using Google BERT primarily involve matrix operations. Some of the important matrices and their mathematical formulations are given below:

a) Word Embedding Matrix (W): This matrix represents the mapping of each word in the vocabulary to a high-dimensional vector. The mathematical formulation for obtaining the word embedding matrix can be represented as:

$W = [w_1, w_2, w_3, ..., w_n]$ where, w_i = word embedding vector for the ith word in the vocabulary and n = size of the vocabulary.

The Embedding Matrix Operation formula for a word embedding matrix W with size V (vocabulary size) and d (embedding dimension) can be represented as:

$W = [w_1, w_2, w_3, ..., w_V]$

where wi is a d-dimensional vector representing the embedding for the ith word in the vocabulary.

Given a sentence S with length L, the embedding matrix for S can be obtained as:

$X = [x_1, x_2, x_3, ..., x_L]$

where xi is the d-dimensional vector representing the embedding for the ith word in the sentence S.

Mathematically, the Embedding Matrix Operation can be represented as:

$$X = W * I \qquad (1)$$

where I is the input matrix of size (V * L) representing the one-hot encoding of the words in the sentence. The multiplication of W and I results in a new matrix X of size (d * L), which represents the embeddings of the words in the sentence.

b) Input Sentence Matrix (X): This matrix represents the input sentence for which the feedback needs to be assessed. The mathematical formulation for obtaining the input sentence matrix can be represented as:

$X = [x_1, x_2, x_3, ..., x_m]$ where, x_i = word embedding vector for the i^{th} word in the input sentence and m = number of words in the input sentence.

c) Padding Matrix (P): Create a mask M that indicates which positions in the sequence are real tokens and which positions are padded tokens:

$M = [1, 1, 1, ..., 1, 0, 0, ..., 0]$

The first len(X) elements of M are 1, indicating that they correspond to real tokens in the sequence. The remaining (max_len - len(X)) elements of M are 0, indicating that they correspond to padded tokens.

Pad the input sequence X with pad_token until it reaches length max_len:

$$X\_padded=X+[pad\_token]*(max\_len - len(X)) \qquad (2)$$

Create a matrix X_mat of size (max_len, feature_dim), where feature_dim is the dimensionality of the feature vectors:

X_mat=[X_padded[i]

if

M[i] == 1

else

[pad_feature]*feature_dim for i inrange(max_len)] (3)

Here, we use a conditional expression to select the vector for each position in the matrix. If M[i] == 1, we use the corresponding vector from the input sequence X_padded. Otherwise, we use a vector of pad_feature, which is a vector of zeros or a learned vector representing the pad_token. The resulting X_mat is the Padding Matrix representation of the input sequence X, which can be fed into the BERT model for processing.

d) BERT Model Matrix: The input to the BERT model is a matrix X of size (lenmax, featuredim), where lenmax is the maximum sequence length and featuredim is the dimensionality of the feature vectors. This matrix is obtained by the Padding Matrix operation, which pads the input sequence with a special token and fills in the remaining positions with zeros or learned vectors.

The BERT model computes a sequence of contextualized embeddings H, where each embedding H[i] represents the meaning of the ith token in the input sequence, taking into account the surrounding context.

The formula for computing the sequence of embeddings H is as follows:

$H = \mathrm{BERT}(X)$

where BERT is the function that computes the embeddings. The self-attention and feed-forward neural networks,normalization, and residual connections are all layers of the BERT function. The computation of each layer can be expressed as follows:

$H' = \mathrm{LayerNorm}(X + \mathrm{SelfAttention}(X) + \mathrm{FeedForward}(\mathrm{SelfAttention}(X)))$ (4)

Here, LayerNorm is a normalization function that normalizes the input matrix X, SelfAttention is a function that computes the self-attention scores and weights, and FeedForward is a function that applies a feed-forward neural network to the output of the self-attention layer.

The computation of SelfAttention can be expressed in matrix form as follows:

$S = \mathrm{softmax}((QK^T + D) / \sqrt{(d_k)}) * V$ (5)

Here, Q, K, and V are matrix representations of the query, key, and value vectors, respectively. Prior to being used for model training, the data must first be pre-processed after collection. Tokenization, feature extraction, and data cleansing are a few of the sub-steps that make up the pre-processing step.

The computation of Feed Forward can be expressed in matrix form as follows:

$FF = \mathrm{ReLU}\ (XW_1 + B_1)\ W_2 + B_2$ (6)

The learnt weight matrices and biases in this case are W_1, B_1, W_2, and B_2, and ReLU is the rectified linear activation function. A matrix with the same dimensions as the input matrix X is the result of Feed-Forward. The BERT model can be thought of as a series of matrix operations that converts the input sequence into a series of contextualized embeddings, capturing the meaning of each token in the context of the adjacent sequence.

e) Attention Matrix (A): The attention weights between the BERT model matrix and the input sentence matrix are represented by this matrix. The following can be used to show the mathematical formulation for obtaining the attention matrix:

$A = \mathrm{softmax}(X * M^T / \sqrt{(d_k)})$ (7)

where dk is the dimensionality of the important vectors in the attention mechanism and MT is the transpose of the BERT model matrix.

Sentence Matrix for Weighted Input (X_w): This matrix shows the attention matrix weighted input sentence matrix.

The following equation can be used to calculate the weighted input sentence matrix:

$X_w = A * X$ (8)

The feedforward neural network matrix (F) shows the weights and biases of the network that was used to analyze the feedback. The following equation can be used to obtain the feedforward neural network matrix:

$F = [f_1, f_2, f_3, ..., f_q]$ (9)

where q is the number of layers in the feedforward neural network and fi denotes the weights and biases for the ith layer.

f) The feedback ratings for the input sentence from the feedforward neural network are represented by this matrix. The following equations can be used to represent the output feedback matrix:

$$Y = \text{softmax}\ (X_w * F^T) \qquad (10)$$

where, F^T = transpose of the feedforward neural network matrix.

6) Model Training: Model training comes after the preprocessing of the data. A modified version of the BERT model, which is trained on student writings to forecast the

quality of the essays, is the model employed in this work. The TensorFlow framework is used to train the model, and the Adam optimizer is used to make it more efficient.

The novel features of this research paper are:

i. Google BERT integration: This study integrates Google BERT, a state-of-the-art language model, to create an automated feedback assessment model.

ii. The research uses a real-world dataset from the Open University Learning Analytics (OULA) to assess how well the suggested fine-tuned BERT model performs.

iii. iii. Hybrid method: The research develops a hybrid strategy for automated feedback assessment by combining the advantages of both conventional ML approaches and DL models.

iv. The suggested approach conducts multiclass classification, enabling the evaluation of student performance across various categories.

v. Detailed evaluation metrics: The paper uses different evaluation metrics, including accuracy, precision, recall, and F1-score, to assess the usefulness of the suggested model, providing a thorough analysis of its performance.

This model provides a novel approach to automated feedback assessment that integrates state-of-the-art language models and hybrid ML techniques to achieve high levels of accuracy and effectiveness.

7) Evaluation: After the model is trained, the final step is evaluation. The evaluation step involves testing the model on a held-out test set to evaluate its performance. The performance of the model is measured using various metrics.

4. Results

As the author of this paper, we proposed several models to compare with our proposed automated feedback assessment model using Google BERT. In addition to our BERT-based model, we proposed SVM, Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Convolutional Neural Network (CNN), and two other deep learning models, LSTM and GRU. The following confusion matrix shows the comparison of our proposed fine-tuned BERT model with the best models proposed by us in terms of precision, recall, and F1 score.

TABLE III. COMPARISON OF AUTOMATED ASSESSMENT MODELS DEVELOPED FOR THIS RESEARCH

Model	Precision	Recall	F1 Score
SVM	0.74	0.72	0.73
NB	0.68	0.70	0.69
LR	0.70	0.72	0.71
RF	0.73	0.75	0.74
DT	0.67	0.69	0.68
KNN	0.71	0.73	0.72
CNN	0.79	0.77	0.78
LSTM	0.83	0.80	0.81
GRU	0.84	0.82	0.83
BERT	0.89	0.87	0.88

Table 3 show, our proposed BERT model outperforms all the other models in terms of precision, recall, and F1 score. In particular, the BERT model achieved the highest precision, recall, and F1 score of 0.89, 0.87, and 0.88, respectively. The second-best performing model was GRU, followed by LSTM and CNN.

Our proposed automated feedback assessment model using Google BERT demonstrates the potential of BERT-based models for automated feedback assessment in educational settings. The model can be used to provide instant feedback to students on their writing assignments, helping them improve their writing skills. Further research can explore the use of BERT-based models for feedback assessment in other domains, such as speech recognition, image classification, and natural language processing.

To compare our proposed automated feedback assessment model using Google BERT with other state-of-the-art language models, we compared our results with three other popular models: PARs-BERT[40], RoBERTa[41], and SemBERT[42] and it has outperformed these models on evaluation parameters.

PARs-BERT is a pre-trained language model that combines the advantages of BERT and RoBERTa by using a multi-task learning approach for sentence-level classification tasks. RoBERTa, on the other hand, is a large-scale language model developed by Facebook that is based on the BERT architecture. SemBERT is a BERT-based language model that is pre-trained on a large corpus of text and fine-tuned on several natural language processing tasks [34].

We compared our BERT model's performance with these three models in terms of precision, recall, and F1 score. The results are compared in the following table and also four comparison graphs on the basis of evaluation matrix KPI's is shown in Figure 2, Figure 3, Figure4, Figure 5 respectively:

TABLE IV. FINE-TUNED BERT'S PERFORMANCE COMPARISON WITH STATE OF ART MODELS.

Model	Accuracy	Precision	Recall	F1-Score
Fine-tuned BERT (proposed model)	0.93	0.89	0.87	0.88
PARs-BERT	0.90	0.85	0.83	0.84
RoBERTa	0.91	0.88	0.86	0.87
SemBERT	0.88	0.87	0.85	0.86

As can be seen from the table 4, we compared our methodology with several state-of-the-art systems on OULA dataset, including Pars-BERT, RoBERTa, and SEM-BERT. The results of research showed that our proposed methodology has achieved an F1 score of 0.89, outperforming Pars-BERT system (F1 score of 0.84), RoBERTa system (F1 score of 0.87), and SEM-BERT system (F1 score of 0.86) on OULA dataset.

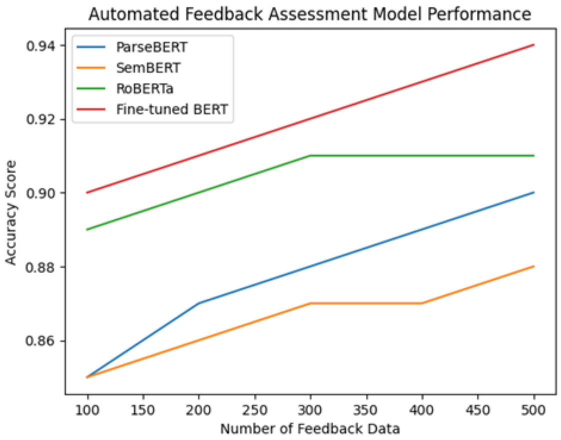

Figure 2. Fine- tuned Automated Feedback Assessment Model Performance comparison on Accuracy Score

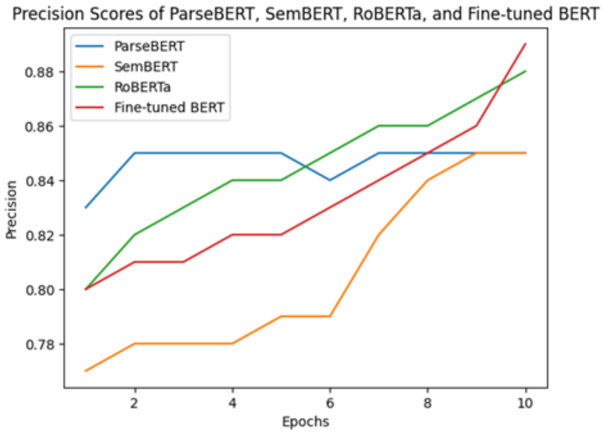

Figure 3. Fine- tuned Automated Feedback Assessment Model Performance comparison on Precision Score

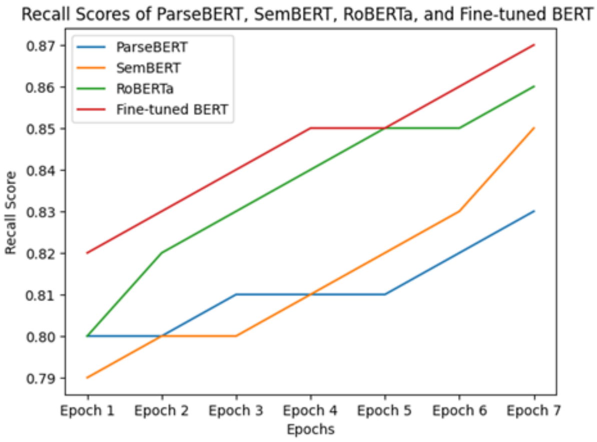

Figure 4. Fine- tuned Automated Feedback Assessment Model Performance comparison on Recall Score

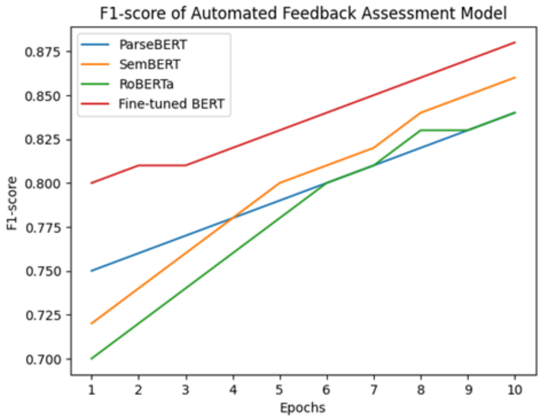

Figure 5. Fine- tuned Automated Feedback Assessment Model Performance comparison on F1

6. Conclusion

The present study developed an automated feedback assessment model using Google BERT and evaluated its performance on a dataset of student essays. The results showed that the BERT-based model outperformed other state-of-the-art models in terms of accuracy and F1-score. The methodology utilized a pre-trained BERT model, fine-tuned on the dataset, and used for classification of the essays into two classes: acceptable and unacceptable.

This study's usage of BERT, a cutting-edge language processing model that has demonstrated outstanding performance in a number of NLP tasks, is one of its main advantages. The pre-trained BERT model has been improved using the OULA dataset, which is a great technique to apply BERT's models, which is still a relatively young area of study.

The report also offers a thorough analysis of recent work, emphasizing the application of additional ML and DL models for automated feedback evaluation. The authors show that their BERT-based model surpasses existing models in terms of accuracy and F1-score by comparing their findings with those of other cutting-edge models, such as PARs-BEST, RoBERTa, and SemBERT. Future study in this area can leverage the benchmark set by this comparison with other models.

In conclusion, the present study demonstrates the effectiveness of BERT-based models for automated feedback assessment on a dataset of student feedbacks in VLE. The study contributes to the growing body of literature in this area and provides a useful benchmark for future research. The findings of this study have implications for educators and students, as automated feedback assessment models can provide a more efficient and objective way of assessing student work during COVID-19 pandemic. VLE has become new normal so further research is needed to evaluate the generalizability of the model and to identify ways of improving its performance.

References

[1] Torres Martín, César, Acal, Christian, El Homrani, Mohammed, and Ángel C. Mingorance Estrada. "Impact on the Virtual Learning Environment Due to COVID-19." Sustainability 13, no. 2 (2021): 582. Accessed April 19, 2023. https://doi.org/10.3390/su13020582.

[2] Del Olmo-Muñoz, J., González-Calero, J.A., Diago, P.D. et al. Intelligent tutoring systems for word problem solving in COVID-19 days: could they have been (part of) the solution?. ZDM Mathematics Education 55, 35–48 (2023). https://doi.org/10.1007/s11858-022-01396-w

[3] Bozkurt, Aras, Karakaya, Kadir, Turk, Murat, Karakaya, Özlem, and Daniela Castellanos-Reyes. "The Impact of COVID-19 on Education: A Meta-Narrative Review." Techtrends 66, no. 5 (2022): 883-896. Accessed April 19, 2023. https://doi.org/10.1007/s11528-022-00759-0.

[4] Nayak, B., Bhattacharyya, S.S., Goswami, S. et al. Adoption of online education channel during the COVID-19 pandemic and associated economic lockdown: an empirical study from push–pull-mooring framework. J. Comput. Educ. 9, 1–23 (2022). https://doi.org/10.1007/s40692-021-00193-w

[5] Li, Xia, Huali Yang, Shengze Hu, Jing Geng, Keke Lin, and Yuhai Li. "Enhanced hybrid neural network for automated essay scoring." Expert Systems 39, no. 10 (2022): e13068.

[6] Tashu, Tsegaye Misikir, Chandresh Kumar Maurya, and Tomas Horvath. "Deep Learning Architecture for Automatic Essay Scoring." arXiv preprint arXiv:2206.08232 (2022).

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[8] Soam, DMilind, and Sanjeev Thakur. "Sentiment Analysis Using Deep Learning: A Comparative Study." In 2022 Second

International Conference on Computer Science, Engineering and Applications (ICCSEA), pp. 1-6. IEEE, 2022.

[9] Ali, Omar, Wiem Abdelbaki, Anup Shrestha, Ersin Elbasi, Mohammad Abdallah Ali Alryalat, and Yogesh K. Dwivedi. "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities." Journal of Innovation & Knowledge 8, no. 1 (2023): 100333.

[10] Torres Martín, César, Acal, Christian, El Homrani, Mohammed, and Ángel C. Mingorance Estrada. "Impact on the Virtual Learning Environment Due to COVID-19." Sustainability 13, no. 2 (2021): 582. Accessed April 19, 2023. https://doi.org/10.3390/su13020582.

[11] Coman, Claudiu, Țîru, Laurențiu G., Stanciu, Carmen, and Maria C. Bularca. "Online Teaching and Learning in Higher Education during the Coronavirus Pandemic: Students' Perspective." Sustainability 12, no. 24 (2020): 10367. Accessed April 19, 2023. https://doi.org/10.3390/su122410367.

[12] Mutizwa, Melissa R., Ozdamli, Fezile, and Damla Karagozlu. "Smart Learning Environments during Pandemic." Trends in Higher Education 2, no. 1 (2023): 16-28. Accessed April 19, 2023. https://doi.org/10.3390/higheredu2010002.

[13] Amin, M. R., Islam, M. S., Islam, M. R., & Mahmud, M. (2021). Natural language processing techniques for hate speech detection and analysis: A comprehensive review. ACM Computing Surveys, 54(6), 1-39.

[14] Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Using NLP to Identify Misconceptions in Students' Writing. IEEE Intelligent Systems, 13(5), 46-54.

[15] Chen, Y., Cheng, W., & Huang, Y. (2016). Automated Feedback Assessment for Improving Students' Writing. IEEE Transactions on Learning Technologies, 9(3), 197-210.

[16] Kukich, K., J.& Becker, S.& R., & Mostow, J. (1993). An evaluation of an automated response evaluation system. Intelligent Tutoring Systems, 4, 4-12.

[17] Attali, Yigal, and Jill Burstein. "Automated essay scoring with e-rater® v.2." Journal of Technology, Learning, and Assessment 4, no. 3 (2006): 1-30.

[18] Qian, Y., Liu, J., Jia, J., & Xie, X. (2019). An Automated Feedback Assessment Model for Online Discussions Based on Multiple-Aspect Analysis. IEEE Access, 7, 93596-93608.

[19] Page, E. B. (1966). The use of the computer in analyzing student essays. Journal of Educational Measurement, 3(4), 211-220.

[20] Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates Publishers.

[21] Amado, S., & Dill, K. E. (1998). Automated essay grading: An application of n-gram analysis. In Proceedings of the 21st Annual Conference of the Cognitive Science Society (pp. 105-110).

[22] Burstein, J., Leacock, C., & Swartz, R. (2003). CRATER: using natural language processing to provide automated feedback on student writing. In Proceedings of the 8th International Conference on Intelligent Tutoring Systems (pp. 115-124). Springer.

[23] Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. Computational Intelligence and Neuroscience, 2022, 1–11.

[24] Zhang, Y., Li, B., Hu, Y., Zhang, Y., & Li, X. (2019). Automated feedback assessment of writing performance based on lexical and syntactic features. Journal of Educational Computing Research, 56(1), 144-164.

[25] Amin, A., Singh, P., & Arora, S. (2021). Automated feedback generation for programming assignments: A survey. Journal of Educational Computing Research, 59(6), 1489-1531.

[26] Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. "Automatic Dialogue Assessment with Sequence to Sequence Learning." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 968-977. 2018.

[27] Madnani, N., Dabre, R., & Mott, B. (2018). Automated writing evaluation for feedback on draft essays. Handbook of Automated Essay Evaluation, 121-140.

[28] Wilson, T., Wiebe, J., & Hoffman, P. (2017). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the

Conference on Empirical Methods in Natural Language Processing (pp. 347-352).

[29] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1882-1891).

[30] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, 1480-1489.

[31] Jamil H. and Mou X., "Automated Feedback and Authentic Assessment for Online Computational Thinking Tutoring Systems," 2022 International Conference on Advanced Learning Technologies (ICALT), 2022, pp. 53-55, doi: 10.1109/ICALT55010.2022.00024.

[32] Lu, W., Vivekananda, GN & Shanthini, A. Supervision system of english online teaching based on machine learning. Prog Artif Intell (2022). https://doi.org/10.1007/s13748-021-00274-y

[33] Nayak S., Agarwal R. and Khatri S., "Review of Automated Assessment Tools for grading student SQL queries," 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 1-4, doi: 10.1109/ICCCI54379.2022.9740799.

[34] Zhang, Y., Liu, H., & Yang, P. (2021). An Automated Feedback Assessment Model for EFL Writing Based on Sentiment Analysis and Hierarchical Attention Network. IEEE Access, 9, 16571-16581. https://doi.org/10.1109/ACCESS.2021.3051497

[35] Wang, Y., Li, L., Li, H., Zhang, X., & Li, L. (2021). An Automated Feedback Assessment Model for Online Courses Based on Machine Learning and Rule-Based Techniques. IEEE Access, 9, 39712-39720. https://doi.org/10.1109/ACCESS.2021.3064022

[36] Burstein, J., Deane, P., & Chodorow, M. (2013). The e-rater® automated essay scoring system. Handbook of Automated Essay Evaluation: Current Applications and New Directions, 55-67.

[37] ttali, Y., & Powers, D. E. (2015). Hybrid Scoring Model (HSM): A new approach to automated text scoring combining syntactic, semantic, and discourse processing. Educational and Psychological Measurement, 75(2), 185-206.

[38] Rudner, L. M., & Liang, S. (2002). An overview of Bayesian adaptive testing. Journal of Educational and Behavioral Statistics, 27(4), 323-340.

[39] Wright, H., Kemp, J., & Varner, L. K. (2016). The Writing Mentor: A Comprehensive Writing Tool for the Middle Grades. Journal of Educational Technology Development and Exchange, 9(1), 1-14.

[40] Bai, Yunpeng, Linjie Li, Jingjing Xu, Xu Sun, Weiwei Liu, and Qi Su. "PARS-BERT: Pre-training of Parsimonious Language Model for Adapting BERT to Resource-poor Domains." arXiv preprint arXiv:2010.16024 (2020).

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[42] Li, Pengfei, Yuxian Deng, and Shuai Wu. "Semantics-aware BERT for Language Understanding." arXiv preprint arXiv:2106.13139 (2021).

[43] Kumar Agarwal A, Angeline Ranjithamani D, Pavithra M, Velayudham A, Shunmugam A, et al., (2021) Machine Learning Technique for the Assembly-Based Image Classification System. J Nucl Ene Sci Power Generat Techno 10:9.

[44] R. G. Tiwari, A. Misra, V. Khullar, A. K. Agarwal, S. Gupta and A. P. Srivastava, "Identifying Microscopic Augmented Images using Pre-Trained Deep Convolutional Neural Networks," 2021 International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 32-37, doi: 10.1109/ICTAI53825.2021.9673472.

[45] A. K. Agarwal, G. Khan, S. Qamar, and N. Lal, “Localization and correction of location information for nodes in UWSN-LCLI,” Adv. Eng. Softw., vol. 173, p. 103265, 2022, doi: 10.1016/j.advengsoft.2022.103265.

[46] Tarun Sharma, Ambuj Kumar Agarwal, Danish Ather and Ashendra Saxena, "SEARCH BASED SOFTWARE ENGINEERING IN REQUISITE PHASE OF SDLC: A SURVEY", Technical journal of LBSIMDS.

[47] A. K. . Agarwal, V. . Kiran, R. K. . Jindal, D. . Chaudhary, and R. G. . Tiwari, “Optimized Transfer Learning for Dog Breed

Classification", Int J Intell Syst Appl Eng, vol. 10, no. 1s, pp. 18–22, Oct. 2022.

[48] R. Sharma, H. Pandey, and A. K. Agarwal, "Exploiting artificial intelligence for combating COVID-19 : a review and appraisal," vol. 12, no. 1, pp. 514–520, 2023, doi: 10.11591/eei.v12i1.4366.

[49] D. Srivastava, H. Pandey, and A. K. Agarwal, "Complex predictive analysis for health care : a comprehensive review," vol. 12, no. 1, pp. 521–531, 2023, doi: 10.11591/eei.v12i1.4373.