# A Trustworthy Automated Short-Answer Scoring System Using a New Dataset and Hybrid Transfer Learning Method

Martinus Maslim[1,2], Hei-Chia Wang[1,3]*, Cendra Devayana Putra[1], Yulius Denny Prabowo[4]

[1] National Cheng Kung University (Taiwan)
[2] Universitas Atma Jaya Yogyakarta (Indonesia)
[3] Center for Innovative FinTech Business Models, National Cheng Kung University (Taiwan)
[4] Bina Nusantara University (Indonesia)

Received 1 September 2023 | Accepted 22 January 2024 | Published 5 February 2024



## Abstract

To measure the quality of student learning, teachers must conduct evaluations. One of the most efficient modes of evaluation is the short answer question. However, there can be inconsistencies in teacher-performed manual evaluations due to an excessive number of students, time demands, fatigue, etc. Consequently, teachers require a trustworthy system capable of autonomously and accurately evaluating student answers. Using hybrid transfer learning and student answer dataset, we aim to create a reliable automated short answer scoring system called Hybrid Transfer Learning for Automated Short Answer Scoring (HTL-ASAS). HTL-ASAS combines multiple tokenizers from a pretrained model with the bidirectional encoder representations from transformers. Based on our evaluation of the training model, we determined that HTL-ASAS has a higher evaluation accuracy than models used in previous studies. The accuracy of HTL-ASAS for datasets containing responses to questions pertaining to introductory information technology courses reaches 99.6%. With an accuracy close to one hundred percent, the developed model can undoubtedly serve as the foundation for a trustworthy ASAS system.

## Keywords

Automated Short Answer Scoring, Hybrid Transfer Learning, Student Answer Dataset, Trustworthy System.

DOI: 10.9781/ijimai.2024.02.003

## I. Introduction

The objective of schools is to educate students through the teaching of academic subjects. To determine the quality of schools and students, it is crucial to measure student competencies [1]. Student competencies can be evaluated by analyzing the outcomes of student learning. The quality of learning is established through the assessment and test of outcomes [2]-[4]. Assessments and evaluations measure students' knowledge and proficiency in each subject [5], [6]. A reliable assessment tool reveals not only the students performing inadequately but also the areas where they will succeed in the future [7]. The assessment process helps teachers analyze patterns in student errors. Teachers can use information from assessments to correct students and advise them about their errors in future classes, and students can subsequently learn from their mistakes [8]. Assessments are supported by various inquiry-based grading approaches and diverse question forms [2].

Some question formats, such as essay, multiple-choice, and short-answer, can be employed to assess the level of student comprehension [7], [9], [10]. Essay writing assessments are critical in gauging the logical reasoning, critical thinking, and foundational writing proficiencies of students [11]. While multiple-choice questions do prove to be an effective approach for assessing a considerable quantity of students, they are most suitable for evaluating knowledge and skills that are specific, well-defined, and often discrete [12], [13]. On the other hand, short-answer questions are a highly effective evaluative tool; they enable teachers to gauge students' comprehension of a subject matter through the provision of concise textual responses [14], [15]. Short-answer questions require students to provide responses ranging in length from three words to two paragraphs [7].

Although short answers are an effective evaluation method, teachers still struggle to use them, particularly in manual grading. Manual answer grading can be inconsistent since human graders must infer meaning from the student's answer [16]. Human graders may become fatigued after reviewing many responses, and the way they correct remaining responses may also vary [17]. This situation may be caused by fatigue, prejudice, or ordering effects [2], [8], [18]. Another reason for the discrepancy is that manual grading is subjective [19, 20] and highly dependent on the moods of the graders [21]. Moreover, the number of students [1], [5], [7] and the time-consuming [22]-[24] aspect of manually scoring short-answer questions pose difficulties. Approximately thirty percent of a teacher's time is spent evaluating students [25]. This problem is genuinely concerning since it means teachers cannot concentrate on their primary task, teaching. This condition will negatively affect teachers' and students' teaching and learning processes.

* Corresponding author.
E-mail address: hcwang@mail.ncku.edu.tw

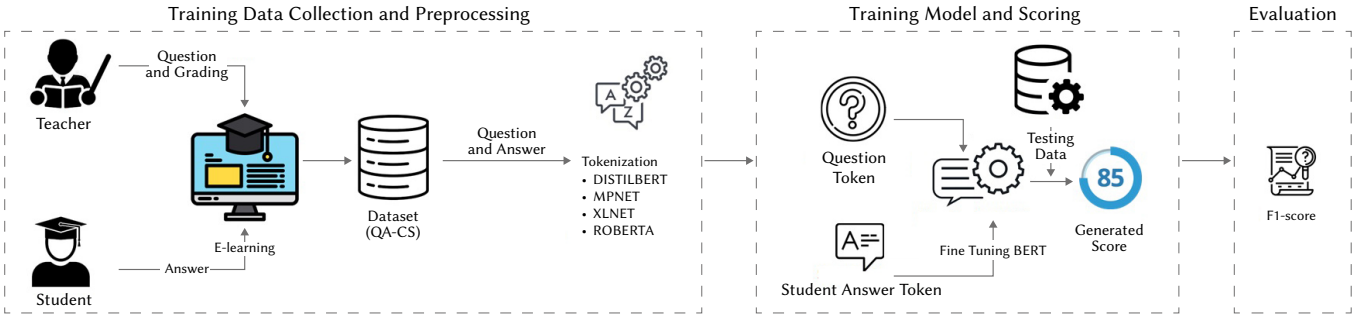


Fig. 1. Phases of the Proposed Research Framework.

Implementing artificial intelligence (AI) is the answer to this issue. AI's capability to produce innovative and real outcomes has elevated it to the forefront of attention in numerous industries, especially education [67]. Natural language processing (NLP) is an AI technology that has the potential to solve the issue of manual grading by enabling the development of a system that can grade responses to short-answer queries automatically; this is referred to as automated short-answer scoring system. The automatic scoring of short answers is one of the most important applications of NLP [26], [27]. In education, automated scoring of short answers has become increasingly popular, allowing for efficient and objective evaluation of student responses. Automatic short answer scoring (ASAS) assigns an output score to a given input answer [28]. The objective of ASAS is to develop a predictive model that takes as input a text response to a specific prompt (e.g., a question about a reading passage) and generates a score expressing the correctness of that response [10], [29], [30]. ASAS systems have garnered much interest because of their capacity to deliver fair and inexpensive grading of large-scale examinations and enhance learning in educational environments [31]. Many studies have focused on the creation of automatic short-answer grading systems, such as C-Rater [32], AutoSAS [25], and AutoMark [33]. However, the accuracy and reliability of these systems can be problematic, particularly as grading becomes more subjective and complex.

Many strategies are utilized to attain high accuracy in the automated scoring of short answer questions. Deep learning approaches have shown promise in enhancing the accuracy of automated scoring systems by enabling them to learn from large datasets and recognize patterns that conventional algorithms may overlook. This study explores constructing a reliable, efficient, and accurate ASAS system via deep learning. Our ultimate objective is to prove that deep learning can be utilized to improve automated scoring systems. While prior research has demonstrated that deep learning techniques can increase the accuracy of ASAS, our study uses a novel approach that focuses on constructing a reliable and more accurate system. Our model incorporates hybrid transfer learning for automated short answer scoring (HTL-ASAS). HTL-ASAS uses various pre-trained tokenizers in combination with the bidirectional encoder representations from transformers (BERT) to increase the accuracy of predictions. We also created a novel student-collected answer dataset for this study. This dataset was acquired without regard to gender or name to eliminate subjectivity and improve system reliability. By emphasizing accuracy and reliability, we seek to contribute to developing more dependable and trustworthy automated short-response scoring systems and enhance the educational experience for all students.

The following is the structure of this document: In section II, prior research concerning ASAS is discussed. The proposed development of the framework is illustrated in Section III. The findings and experimental context are detailed in Section IV. A discussion of the results of the findings is provided in Section V. In section VI, the concluding remarks are provided.

## II. Literature Review

ASAS is a challenging task that requires the capacity to evaluate the semantic content of a student's response accurately. In recent years, this topic has been the subject of numerous studies, and many techniques have emerged as potentially effective methods for enhancing the accuracy of ASAS systems. The fundamental concept of ASAS is to compare student responses to teacher responses, sometimes known as the "gold standard." Studies have utilized various approaches to calculate text similarity. One type of approach involves calculating text similarity based on semantic [34] or grammatical characteristics [13] or with the word overlapping approach [35]. Many advancements have been made to this fundamental idea, including using a semantic similarity measuring approach based on word embedding techniques and syntactic analysis to evaluate the learner's accuracy [5]. Combining semantic analysis with orthography and syntax analysis [36] or with graph-based lexico-semantic text matching is a further advancement that can be implemented [37].

Machine learning is another topic that can be applied to automatically scoring short answers. Term frequency inverse-document frequency (TF-IDF) [26], [38], long short-term memory (LSTM) [39, 40], support vector machines (SVMs) [7], [9], [41], latent semantic analysis [42], Gini [7], k-Nearest Neighbors (KNN) [7], finite state machine [18], and bagging and boosting [7] have all been employed. In addition to applying machine learning, various studies have employed deep learning to increase the accuracy of ASAS. Earlier research employed the concept of deep learning by utilizing transformers for data training. Transformers can be converted into graph transformers, which generate relation-specific token embeddings within each subgraph, which are subsequently aggregated to produce a subgraph representation [43]. Other studies have utilized pre-trained models such as BERT [22], [26], [31], [44]-[48], XLNET [49], [50], MPNET [51], [52], RoBERTa [50], [53], and Distil BERT [53].

In conclusion, many deep learning techniques can be applied to the ASAS system. These techniques have demonstrated potential for enhancing the accuracy of ASAS systems. Current research shows that deep learning models are excellent at enhancing the reliability of these systems.

## III. Proposed Framework

In the proposed framework, there are three different procedures. The first step is data collection and preprocessing. The second procedure consists of training and testing the model that has been trained using the created dataset. Evaluation of the trained model is the final phase. The results of this evaluation are then compared to those of other studies to show the strengths of the framework proposed in this study. Fig. 1 displays the phases of the proposed research framework.

### A. Data Collection Module

First, we consider the data collection process. Teachers and students of an introductory information technology course participated in this phase. The teacher administered questions via an e-learning platform. The students then responded to the teacher's questions. Students responded to ten questions related to the course. Within two weeks, the answers of 229 students who had responded to the ten questions posed by the teacher were gathered. After the collection of student responses was complete, the teacher manually evaluated the results of the students' work. The teacher conducted the evaluation based on a previously prepared answer key. The teacher scored 50 for incorrect responses and 100 for correct responses. To recognize the students' effort in responding to the question, teachers award 50 points to those who provide incorrect answers. For student responses like the teacher's, values between 50 and 100 were awarded in multiples of 10, namely, 60, 70, 80, and 90. The teacher also assigned a score of zero to students who did not respond to the given questions.

### B. Data Preprocessing Module

After data collection, the next stage is data preprocessing. Tokenization is performed during this phase. In natural language processing, tokenization is often used to extract needed abstract information of paragraphs or sentences into smaller units that can be assigned meaning more readily by machine. Tokenizers are typically either carefully constructed systems of language-specific rules, which are expensive and require both manual feature engineering and linguistic expertise, or data-driven algorithms that split strings based on frequencies within a corpus, which are more flexible and easier to scale but are ultimately too simplistic to handle the wide range of linguistic phenomena that are not captured by their string-splitting [54]. In deep learning, the fundamental model for extracting contextualized word embeddings is called a Transformer [64]. The central concept of the Transformer architecture is to employ multi-head attention for concurrent data processing while preserving the temporal sequence characteristic of time-series data by the inclusion of positional embedding within the embedding layer [66].

Table I illustrates the tokenization of the phrase "An artificial intelligence robot." This table displays the transformation of the sentence into word embedding and attention mask embedding, following the Transformer architecture. This tokenization process distinguishes our study from previous research. We employ four distinct tokenizers within the proposed hybrid transfer learning framework.

TABLE I. Examples of Tokenization

| **Before tokenization** | **An artificial intelligence robot** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Token** | [cls] | 'An' | 'artificial' | 'intelligence' | 'robot' | [pad] | |
| **Word Embedding** | $WE_1$ | $WE_2$ | $WE_3$ | $WE_4$ | $WE_5$ | $WE_n$ | 0 |
| **Attention Mask Embedding** | $Att_1$ | $Att_2$ | $Att_3$ | $Att_4$ | $Att_5$ | $Att_n$ | 0 |

Several other types of tokenizers were subjected to experimentation before the identification of the four types that would be utilized in this study. The findings of this experiment indicate that the input format of the BERT model, which was utilized during the training phase, is compatible with the four tokenizers selected for this study: DistilBERT, MPNet, XLNet, and RoBERTa. In this study, various experiments were conducted in which the outcomes of the chosen tokenizer were combined with the BERT model's training data. Hybrid MPNet refers to the output of the MPNet tokenizer when combined with the BERT model. Hybrid DistilBERT is the name given to the combination of the DistilBert tokenizer and the BERT model. The combination of the XLNet tokenizer and the BERT model is called Hybrid XLNet. The hybrid name for the RoBERTa tokenizer and the BERT model is Hybrid RoBERTa. A challenge appeared during the procedure of identifying the optimal combination: the lengthy duration of one experiment. To circumvent this, we conducted experiments on two separate servers. This is greatly beneficial in establishing correspondence between the tokenizer and the BERT model that was employed during the data training phase.

DistilBERT [55] is derived from BERT [56] by employing knowledge distillation. To create a more compact version of BERT, the architects of DistilBERT eliminated token-type embeddings and the pooler from the architecture and reduced the number of layers by a factor of 2. DistilBERT is a lightweight variant of BERT that is pre-trained using only the masked language model (MLM) task but with the same corpus: BookCorpus, which contains 800 million words; English Wikipedia, which contains 2,500 million words, a 30,000 token vocabulary, and WordPiece tokenization. Given an evolving word definition, the WordPiece model is combined with a data-driven approach to maximize the language-model likelihood of the training data. Given a training corpus and D desired tokens, the optimization problem is to select D word pieces such that when they are segmented according to the selected WordPiece model, the resulting corpus contains the fewest number of word pieces [57].

The masked and permuted pretraining model (MPNet) tokenizer was developed in collaboration with researchers from Microsoft and the Nanjing University of Science and Technology in 2020 [58]. MPNet incorporates the benefits of MLMs, such as BERT, and Pre-trained Language Models (PLMs), such as XLNet, by incorporating additional positional information into the permutation-based loss function. The MPNet tokenizer employs a byte-level byte pair encoding (BPE) algorithm to generate a vocabulary of subwords with a fixed size. The BPE algorithm iteratively replaces the most frequent pairs of consecutive bytes in the input text with a single new byte. This procedure is repeated until the desired vocabulary size has been attained. It can, therefore, comprehend a text based on its positional and nonpositional information. The tokenizer utilized by MPNet is inherited from BERT. MPNet was trained on many corpora of text totaling over 160 GB in size and optimized for multiple downstream NLP tasks [59].

The XLNet tokenizer is comparable to the tokenizers used in other transformer-based models but has some distinctive characteristics. Like other tokenizers, it transforms unprocessed text into a sequence of tokens the model can process. The tokenizer employs a subword-based approach, which divides words into smaller subwords and assigns a unique token to each subword. The total size of XLNet using subword fragments for Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl is 32.89B [60]. XLNet uses the SentencePiece tokenization algorithm. SentencePiece consists of a natural language processing tokenizer and detokenizer. It performs subword segmentation, supports the BPE algorithm and unigram language model, and converts this text into an id sequence while ensuring perfect reproducibility of the normalization and subword segmentation [61]. BPE is an algorithm for subword segmentation that encodes uncommon and unknown terms as sequences of subword units. The assumption is that various word classes can be translated using units smaller than words, such as names (via character reproduction or transliteration), compounds (via compositional translation), and cognates and loanwords (via phonological and morphological transformations) [62].

The RoBERTa tokenizer generates subword tokens using a variant of the BPE algorithm. BPE functions by iteratively merging the most frequently occurring character or character sequence pairs in the

training corpus until the maximum vocabulary size is attained. This method can produce a vocabulary of variable-length subword units that more accurately represent the morphology and syntax of the language than traditional word-based tokenization. Additionally, the RoBERTa tokenizer employs a variety of optimizations to enhance the quality of the tokenization procedure. It employs, for instance, dynamic masking to prevent overfitting during pretraining and removes whitespace from the input text to increase efficiency. RoBERTa was trained on a combined dataset for the same number of steps as before (100K). RoBERTa preprocessed over 160 GB of text in total [63].

Table II displays the tokenizers utilized in this study. Each tokenizer uses a distinct corpus for recognizing terms. DistilBERT and MPNet utilize scholarly sources such as BookCorpus and Wikipedia, whereas XLNet adds Giga5 and ClueWeb. RoBERTa expands its corpus to include CC-News, OpenWebTest, and Stories, among others. This distinction results in distinct text representations. This study investigated the appropriate tokenizer for short-answer question tasks.

TABLE II. Summary of Each Tokenizer

| Tokenizer | Corpus | Embedding Technique | #Tokens | #Positions |
|---|---|---|---|---|
| DistilBERT | BookCorpus, English Wikipedia | WordPiece Embedding | 85% | 100% |
| MPNet | BookCorpus, English Wikipedia | WordPiece Embedding | 95% | 100% |
| XLNet | BookCorpus, Wikipedia, Giga5, ClueWeb, | WordPiece Embedding | 92.5% | 92.5% |
| RoBERTa | BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories + pretrain for even longer | Byte Pair Encoding | - | - |

## C. Question and Answer Embedding Modules

The preprocessing dataset is trained after the tokenization of teacher questions and student responses is performed. BERT accomplished natural language understanding by considering input consistency. BERT extracts a single token sequence from a single text sentence, and for the NSP objective, it extracts a single token sequence from two text sentences (adding a [SEP] token as a separator) [56]. Each sequence has the specific classification embedding [CLS] added before it, and it serves as the input of the classification-task layer. Combining the corresponding token, segment, and position embeddings puts the corresponding representation of the input together. Each provided input token receives this kind of approach [56]. Fig. 2 illustrates the architecture of BERT fine-tuning for this study. The tokenization process is initiated once the learners have provided answers to the teacher's questions and the encoder layer has become working, as represented in Fig. 2. The BERT encoder has two primary sublayers: the multihead self-attention layer and the positionwise fully connected feedforward network layer [56]. The output of the Question and Answer embedding of Hybrid MPNet is the latent embedding of the answer and question. The question-and-answer embedding sizes are $1 \times 768$. These two embeddings are concatenated to predict the potential score for students and produce an object of size $2 \times 768$. Finally, we connect the regression layer to predict the probability of the score and use the highest probability as the predicted score.

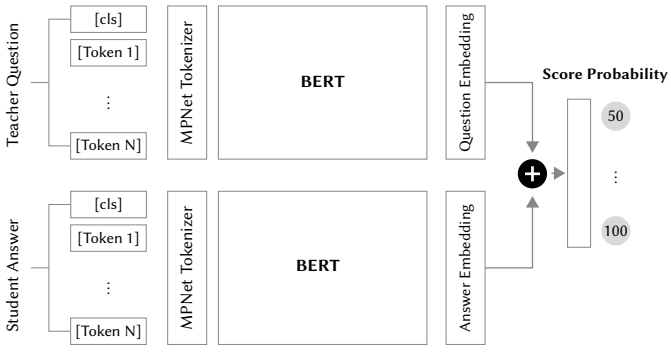


Fig. 2. BERT-Based Question-Answer Embedding Architecture.

## D. Evaluation Technique

In this research, k-fold cross-validation is used as the evaluation methodology. The dataset is first divided into k folds, with k-1 folds used for training and the remaining fold used for evaluation. The folds are then switched until all folds have been trained and evaluated against the remaining k-1 folds, and an average is then calculated. This study utilizes ten-fold cross-validation. The F1-score is utilized for the evaluation matrices in the study. Formula 1 defines the F1-score as the weighted average of precision and recall based on the weight function β. Formula 2 defines the F1-score as the harmonic mean of precision and recall. The F1 score is also referred to as the F-measure. Different F1-score indices can assign distinct weights to precision and recall. Precision is calculated by dividing the number of correct instances retrieved by the total number of instances retrieved, as in Formula 3. Recall is calculated by dividing the number of correct instances retrieved by the total number of correct instances, as in Formula 4.

$$F - score: \; F_\beta = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \tag{1}$$

When $\beta = 1$, the standard F1-score is obtained

$$F - score: \; F_1 = F = 2 * \frac{P * R}{P + R} \tag{2}$$

$$Precision: \; P = \frac{tp}{tp + fp} \tag{3}$$

$$Recall: \; R = \frac{tp}{tp + fn} \tag{4}$$

# IV. Results

In this section, the experimental results of the proposed Hybrid MPNet are presented. We collect and sign the score of each answer. Then, we propose a new deep-learning technique to predict the score. Finally, we evaluate the accuracy of our proposed method using the F1-score, precision, and recall.

## A. Dataset

This study employs a dataset compiled by the authors. The collected dataset comprises four columns: teacher-initiated questions, teacher-prepared responses, student responses, and students' grading in numerical form. The example of the collected dataset can be seen in Table III. The questions given to students were related to an introductory course in information technology. There were five categories and two questions per category for ten questions. The five categories were: 1) data and information, 2) the most recent technology, 3) software, 4) hardware, and 5) the development of computer networks. The scores assigned by the teacher have a value of 0, 50, 60, 70, 80, 90, or 100. 229 students responded to the questions, so the total data collected contained 2290 data. After the data were collected, they were cleaned. First, responses with a zero value were removed, indicating that the student did not answer the question. This

was done to reduce the homogeneity of the training data and ensure that the resulting model is highly accurate and creates a trustworthy system. Following the data cleaning procedure, 2023 data points remained. The distribution of word length in student responses is illustrated in Fig. 3. The majority of responses are typically between zero and two hundred words in length. The longest response exceeds 400 words in length.

TABLE III. Examples of Collected Dataset

| Question | Teacher Answer | Student Answer | Grade |
|---|---|---|---|
| Please define what a "computer network" is. | A computer network can be defined as a communication system that connects two or more computers and peripheral devices and allows data transfer between components. | A computer network is a link between one computer/device and another computer/device that uses network media as an intermediary. | 70 |
| | | A unit that causes a computer to run. | 50 |
| | | A communication network that allows computers to communicate with each other by exchanging data. | 100 |
| | | - | 0 |
| Please explain the definition of Artificial Intelligence (AI). | A field of computer science devoted to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving and pattern recognition. | A program with a neural network that can think like humans in carrying out tasks. | 70 |
| | | A field of computer science that tries to solve cognitive problems like learning, solving problems, and recognizing patterns, that are often associated with human intelligence. | 100 |
| | | A smart technology embedded in a device. | 60 |
| | | An artificial intelligence robot. | 50 |

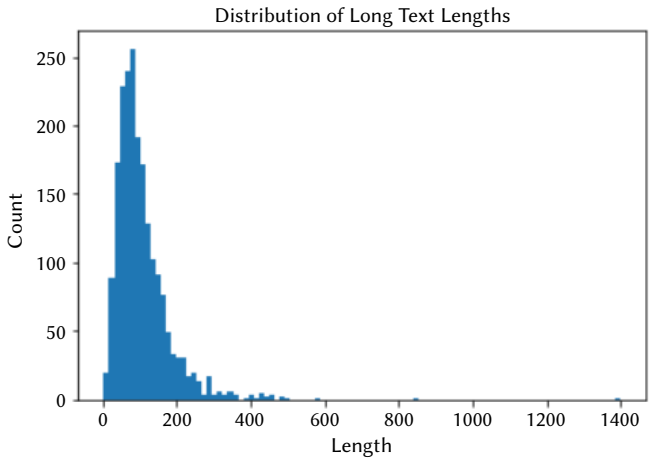


Fig. 3. Distribution of Word Length in Student Responses.

As a consequence of the data cleansing process, the number of responses varied for each question. After the data cleansing procedure, the number of student responses categorized by question type is presented in Table IV. There are three questions to which fewer than 200 students respond. One question pertains to the data and information category, one concerns software, and one investigates the development of computer networks. Because numerous students failed to respond to that question, the instructor therefore assigns a zero grade.

Table V displays the quantity of student responses used in this study based on assigned scores after the data cleaning process. As shown in Table V, the quantity of responses for each value is not distributed exactly equally. The answers provided by the majority of students yield scores ranging from 60 to 80. From the complete data for 2023, this is evident from the 1,419 answers that obtained a score within that range. At 50, 90, and 100, the remainder was balanced. It is possible to conclude from this distribution that a small number of students submitted responses attaining a perfect score of 100. Also, a small number of students submitted responses that received a minimum score of 50.

TABLE IV. Number of Student Answers Based on Question Type

| Category | Question | Number of Answers |
|---|---|---|
| Data and information | What are data and information? Please compare the differences. | 215 |
| | What is the information processing cycle? | 198 |
| Recent technology | What is Augmented Reality (AR)? | 203 |
| | Please explain the definition of Artificial Intelligence (AI). | 203 |
| Software | Please define what "freeware" is. | 164 |
| | Please define what an "operating system" is and explain its function. | 219 |
| Hardware | Explain the function of a router. | 213 |
| | What is a Central Processing Unit (CPU)? | 213 |
| Development of computer networks | Please define what a "computer network" is. | 183 |
| | Please give the definition of the Internet of Things (IoT). | 212 |

TABLE V. Number of Student Answers Based on Teacher Grading

| Grading | Number of Answers |
|---|---|
| 50 | 247 |
| 60 | 403 |
| 70 | 586 |
| 80 | 430 |
| 90 | 174 |
| 100 | 183 |

## B. Parameter Setting

We propose hybrid transfer learning as a model for ASAS. Before conducting model training experiment, we set our parameters. Table VI summarizes some of the study's parameters.

TABLE VI. Parameter Settings of the Automated Short Answer Grading Model

| Parameter | Value |
|---|---|
| Batch size | 10 |
| Optimizer | Adam |
| Learning rate | 0.00001 |
| Embedding size | 300 |
| Activation function | ReLU |
| The final layer activation function | Sigmoid |

### C. Result

**Experiment 1**: In the first experiment, the authors trained the proposed model (HTL-ASAS) for various epochs to determine which provided the most accurate model. The epochs tested were 10, 20, 30, and 40. Training used ten-fold cross-validation to validate the model. Comparison between the evaluation generated by the computer and the evaluation conducted by the teacher yields the F1 score accuracy. When both the machine and the teacher arrive at the same evaluation, this represents a true positive. False positives happen when the assessments of the machine and the instructor differ. Fig. 4 displays each model's F1 score after 10, 20, 30, and 40 epochs for each tokenizer. The average F1 score, as shown in Fig. 4, is the result of the evaluation that was performed.

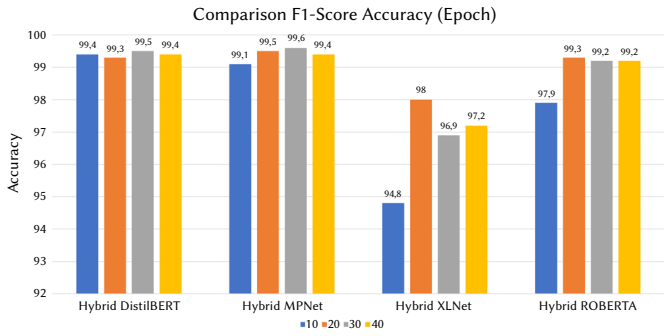


Fig. 4. Results of F1-Score by Epoch.

**Experiment 2:** In the second experiment, the highest F1-score obtained by the various hybrid transfer learning algorithms in the proposed framework, namely, Hybrid DistilBERT, Hybrid MPNet, Hybrid XLNet, and Hybrid RoBERTa, were compared. Fig. 5 displays the comparison of F1 scores.

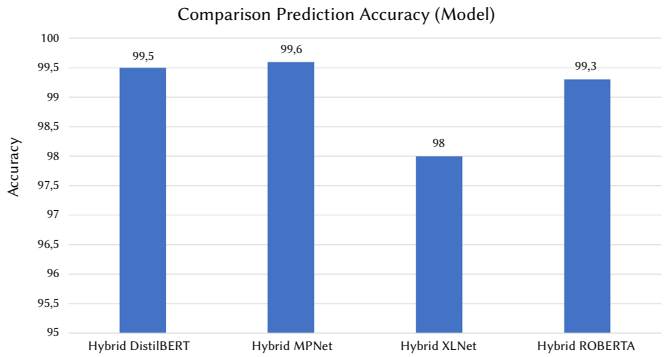


Fig. 5. Results of F1-Score by Proposed Framework Model.

**Experiment 3**: The objective of the third experiment was to compare the F1-scores of the proposed framework with the F1-scores models based on prior research. The prior research models were trained using the dataset from this study. The F1 scores were then obtained from the results of the training model. The following models were used for comparison:

1. BERT architecture to grade short answers [47], [53], [65]. A pre-trained version of the BERT base model was utilized in these experiments.
2. MPNet, specifically mpnet-base-v2 model, which has been used to determine the similarity of short texts [52].
3. DistilBERT, which has been used to grade short answer responses [53]. This study utilized a pre-trained DistilBERT model, namely, the DistilBERT base model.
4. XLNet, which is a pre-trained model used to assess short-answer responses [50].
5. Pre-trained RoBERTa and RoBERTa base architectures used in previous studies [50], [53], [65].

Each model from previous research was trained on this study's dataset and then compared to the framework proposed in this study. For previous BERT research, ten epochs were used to train the model. For MPNet, 30 epochs were used, the same number used for the MPNet hybrid proposed in this study. For DistilBERT, the number of epochs was set to 30, the optimal number for the DistilBERT hybrid. The final two models, XLNet and RoBERTa, were trained in 20 epochs, the optimal number of epochs for hybrid XLNet and RoBERTa. Table VII compares the F1 scores of the proposed framework and models from previous studies.

TABLE VII. Comparison of F1-Score Accuracy

| Research | Model | F1-Score Accuracy |
|---|---|---|
| [47], [53], and [65] | BERT | 0.992 |
| [52] | MPNet | 0.952 |
| [53] | DistilBERT | 0.961 |
| [50] | XLNet | 0.899 |
| [50], [53], and [65] | RoBERTa | 0.963 |
| Our Proposed Framework (*) | Hybrid DistilBERT | 0.995 |
| | Hybrid MPNet | **0.996** |
| | Hybrid XLNet | 0.98 |
| | Hybrid RoBERTa | 0.993 |

## V. Discussion

The first experiment's results, depicted in Fig. 4, indicate that increasing the number of epochs has no effect on the accuracy of predictions made for the framework proposed in this study. Each tokenizer employed by the proposed framework requires a distinct number of epochs to attain the highest level of accuracy. Except for the DistilBERT tokenizer, the accuracy of each tokenizer is the lowest for epoch 10. In this investigation, the sample size at epoch 10 was insufficient for each model to achieve maximum accuracy. Upon entering epoch 20, the F1-score of several models increased. Only the DistilBERT tokenizer experienced a reduction in score. The XLNet and RoBERTa tokenizers reached their maximal F1-scores of 98% and 99.3%, respectively, in the 20th epoch; thus, the F1-scores of the corresponding hybrid models at epochs 30 and 40 were lower than at epoch 20. At epoch 30, the MPNet tokenizer attained its highest F1-score of 99.6%, and the DistilBERT tokenizer reached its maximum accuracy rate of 99.5%. The first experiment's results were used for comparison in the second experiment. The MPNet tokenizer paired with the BERT layer (Hybrid MPNet) achieved the greatest accuracy of 99.6%, as shown in Fig. 5. The DistilBERT tokenizer paired with the BERT layer (Hybrid DistilBERT) achieved the next-best accuracy of 99.5%. The accuracy of Hybrid RoBERTa was 99.3%, while the accuracy of Hybrid XLNet was only 98%. This comparison demonstrates that by employing hybrid transfer learning, accuracy increases, and the resulting data can enable the development of a more reliable ASAS system.

The results of Experiment 3, presented in Table VII, indicate that our proposed framework that combines the MPNet tokenizer with the BERT layer, also known as Hybrid MPNet, has the highest F1 score among the other models. Hybrid MPNet achieves an F1-score of 99.6%. In addition, Hybrid DistilBERT and Hybrid RoBERTa are among the models proposed in this study that have the highest value relative to models used in previous research. The F1 scores produced by Hybrid DistilBERT and Hybrid RoBERTa were 99.5% and 99.3%, respectively.

The BERT model [47], [53], [65] produced the highest values among previous models. The F1-score for this BERT model was 99.2%. This F1-score is greater than that of one of the proposed models in this investigation, namely, Hybrid XLNet (98%), and is also greater than the F1-scores obtained by several models used in previous studies, including MPNet (95.2%) [52], DistilBERT (96.1%) [53], XLNet [50], and RoBERTa (96.3%) [50], [53], [65].

The results of this study indicate that Hybrid MPNet is a more accurate method than those used in previous research. This is because MPNet utilizes the dependencies between predicted tokens through permuted language modeling and enables the model to see supplementary position information to overcome the difference between pretraining and fine-tuning. In addition, Hybrid MPNet's better results compared to alternative methods can be attributed to the specific correspondence between the corpus trained in the MPNet tokenizer and the collected dataset. The corpus utilized by the MPNet tokenizer is comprised of words extracted from English Wikipedia and BookCorpus, as shown in Table II. DistilBERT tokenizer operations utilize the identical corpus. An important distinction is found in the fact that DistilBERT trains a lower percentage of tokens (85%) than MPNet Tokenizer. In comparison to alternative tokenizers, the MPNet tokenizer utilizes a greater quantity of training tokens. Experiments on various tasks demonstrate that MPNet outperforms MLM and PLM, as well as previously robust pre-trained models, including BERT, XLNet, and RoBERTa, by a substantial margin [59].

The findings derived from this study will influence the area of education. This system will improve the performance of teachers when evaluating student work. It will not be long before the students are informed of the assessment results. As a result, teachers can dedicate more time to planning and refining the learning process within the classroom. Instead of having to wait for the instructor to evaluate their work manually, this method enables students to obtain immediate feedback. Students will receive more objective grades as a consequence of the reduced subjectivity of the teacher caused by the implementation of this system. By establishing confidence among teachers and students, the experimental results indicate that utilizing AI to assess short-answer assessments produces reliable and objective outcomes. Aside from that, the implementation of this system's results demonstrates that artificial intelligence can be applied to the field of education. The opportunities for both educators and students to utilize AI are described by the Sustainable Development Goals (SDG4) of the UNESCO 2030 Agenda as they pertain to the impact of AI in education [68].

## VI. Conclusion

Teachers can select from various effective assessment methods for students, one of which is short answer questions. However, one of the most challenging aspects of teaching is evaluating student work in a limited amount of time. Consequently, the results of an assessment can be inconsistent if the teacher is pressured. Our study assists teachers in overcoming these inconsistencies by developing a system that automatically assigns grades to students' short answers. The goal is to construct a trustworthy system, so students believe the assessments are accurate. A method that can generate near-perfect system accuracy is required to achieve this objective. In addition, the system must be objective about student work. For the method proposed in this study, both objectives are met. We implement hybrid transfer learning as a novel technique for achieving high accuracy and generate a new training dataset containing students' short responses and feedback. We anticipate that the constructed system will be capable of objective evaluation with this dataset. Based on the results of the conducted experiments, the hybrid transfer learning method proposed in this study has the highest accuracy of 99.6%. Despite focusing solely on the F1-score to assess accuracy, the test results for this system indicate a 99.6% accuracy rate, which signifies a highly optimistic implementation potential. Nevertheless, additional assessment utilizing additional matrices is necessary. There is no doubt that a more comprehensive assessment of the system's capability to evaluate student exams can be obtained by administering tests utilizing a broader variety of comprehensive matrices. The F1-score matrix, nevertheless, is considered satisfactory from the perspective of this study.

Future research may concentrate on implementing the proposed framework in disciplines other than information technology. In addition, other evaluation matrices can be applied to evaluate this mode. In the future, automated scoring will hopefully make administering assessments easier for teachers to concentrate on enhancing the quality of learning.

## References


[1] Q. Aini, A. E. Julianto, and D. Purbohadi, "Development of a Scoring Application for Indonesian Language Essay Questions," in Proceedings of the 2018 2nd International Conference on Education and E-Learning, 2018, pp. 6-10.

[2] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," International Journal of Artificial Intelligence in Education, vol. 25, no. 1, pp. 60–117, Oct. 2014, doi: https://doi.org/10.1007/s40593-014-0026-8.

[3] B. S. J. Kapoor et al., "An analysis of automated answer evaluation systems based on machine learning," in 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2020, pp. 439-443.

[4] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," Artificial Intelligence Review, Sep. 2021, doi: https://doi.org/10.1007/s10462-021-10068-2.

[5] F. F. Lubis et al., "Automated Short-Answer Grading using Semantic Similarity based on Word Embedding," International Journal of Technology, vol. 12, no. 3, p. 571, Jul. 2021, doi: https://doi.org/10.14716/ijtech.v12i3.4651.

[6] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 752-762.

[7] A. Çınar, E. Ince, M. Gezer, and Ö. Yılmaz, "Machine learning algorithm for grading open-ended physics questions in Turkish," Education and Information Technologies, Mar. 2020, doi: https://doi.org/10.1007/s10639-020-10128-0.

[8] A. Olowolayemo, S. D. Nawi, and T. Mantoro, "Short answer scoring in English grammar using text similarity measurement," in 2018 International Conference on Computing, Engineering, and Design (ICCED), IEEE, 2018, pp. 131-136.

[9] G. De Gasperis et al., "Automated grading of short text answers: preliminary results in a course of health informatics," in Advances in Web-Based Learning–ICWL 2019: 18th International Conference, Magdeburg, Germany, September 23–25, 2019, Proceedings, Springer International Publishing, 2019, pp. 190-200.

[10] S. Patil and K. P. Adhiya, "Automated Evaluation of Short Answers: a Systematic Review," in Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021, 2022, pp. 953-963.

[11] Y.-H. Park, Y.-S. Choi, C.-Y. Park, and K.-J. Lee, "EssayGAN: Essay Data Augmentation Based on Generative Adversarial Networks for Automated Essay Scoring," Applied Sciences, vol. 12, no. 12, p. 5803, Jun. 2022, doi: https://doi.org/10.3390/app12125803.

[12] M. J. Gierl, S. Latifi, H. Lai, A.-P. Boulais, and A. De Champlain, "Automated essay scoring and the future of educational assessment in medical education," Medical Education, vol. 48, no. 10, pp. 950–962, Sep. 2014, doi: https://doi.org/10.1111/medu.12517.

[13] S. H. Mijbel and A. T. Sadiq, "Short Answers Assessment Approach based on Semantic Network," Iraqi Journal of Science, pp. 2702–2711, Jun. 2022, doi: https://doi.org/10.24996/ijs.2022.63.6.35.

[14] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling

short-answer grading by combining peer assessment with algorithmic scoring," in Proceedings of the first ACM conference on Learning@Scale Conference, 2014, pp. 99-108.

[15] A. K. F. Lui, S. C. Ng, and S. W. N. Cheung, "A framework for effectively utilising human grading input in automated short answer grading," International Journal of Mobile Learning and Organisation, vol. 16, no. 3, p. 266, 2022, doi: https://doi.org/10.1504/ijmlo.2022.124160.

[16] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," Procedia Computer Science, vol. 169, pp. 726-743, 2020.

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

[18] K. Anekboon, "Automated scoring for short answering subjective test in Thai's language," in 2018 International Conference on Image and Video Processing, and Artificial Intelligence, vol. 10836, SPIE, 2018, pp. 324-329.

[19] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," Journal of Computing in Higher Education, Jun. 2021, doi: https://doi.org/10.1007/s12528-021-09283-1.

[20] J. Xiong, J. M. Wheeler, H. Choi, J. Lee, and A. S. Cohen, "An empirical study of developing automated scoring engine using supervised latent dirichlet allocation," in Quantitative Psychology: The 85th Annual Meeting of the Psychometric Society, Virtual, Springer International Publishing, 2021, pp. 429-438.

[21] P. Kudi, A. Manekar, K. Daware and T. Dhatrak, "Online Examination with short text matching," 2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN), 2014, pp. 56-60, doi: 10.1109/GCWCN.2014.6998787.

[22] A. Condor, "Exploring automatic short answer grading as a tool to assist in human rating," in Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II, vol. 12164, Springer International Publishing, 2020, pp. 74-79.

[23] S. Roy, Y. Narahari, and O. D. Deshmukh, "A perspective on computer assisted assessment techniques for short free-text answers," in Computer Assisted Assessment. Research into E-Assessment: 18th International Conference, CAA 2015, Zeist, The Netherlands, June 22–23, 2015. Proceedings, vol. 18, Springer International Publishing, 2015, pp. 96-109.

[24] X. Ye and S. Manoharan, "Machine Learning Techniques to Automate Scoring of Constructed-Response Type Assessments," in 2018 28th EAEEIE Annual Conference (EAEEIE), IEEE, 2018, pp. 1-6.

[25] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, and R. Zimmermann, "Get it scored using autosas—an automated system for scoring short answers," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 9662-9669.

[26] P. Shweta and K. Adhiya, "Comparative Study of Feature Engineering for Automated Short Answer Grading," in 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), IEEE, 2022, pp. 594-597.

[27] C. N. Tulu, O. Ozkaya, and U. Orhan, "Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM," IEEE Access, vol. 9, pp. 19270–19280, 2021, doi: https://doi.org/10.1109/access.2021.3054346.

[28] T. Sato, H. Funayama, K. Hanawa, and K. Inui, "Plausibility and Faithfulness of Feature Attribution-Based Explanations in Automated Short Answer Scoring," presented at the International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science, vol 13355. Springer, Cham.

[29] M. Heilman and N. Madnani, "The impact of training data on automated short answer scoring performance," presented at the Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Denver, Colorado, 2015, pp. 81-85.

[30] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee, "Investigating neural architectures for short answer scoring," presented at the Proceedings of the 12th workshop on innovative use of NLP for building educational applications, Copenhagen, Denmark, 2017, pp. 159-168.

[31] H. Funayama, T. Sato, Y. Matsubayashi, T. Mizumoto, J. Suzuki, and K. Inui, "Balancing Cost and Quality: An Exploration of Human-in-the-Loop Frameworks for Automated Short Answer Scoring," presented at the International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science, vol 13355. Springer, Cham.

[32] C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," Computers and the Humanities, vol. 37, no. 4, pp. 389-405, 2003.

[33] R. Siddiqi, C. J. Harrison, and R. Siddiqi, "Improving Teaching and Learning through Automated Short-Answer Marking," IEEE Transactions on Learning Technologies, vol. 3, no. 3, pp. 237–249, Jul. 2010, doi: https://doi.org/10.1109/tlt.2010.4.

[34] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," presented at the Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 2009, pp. 567-575.

[35] F. S. Pribadi, A. B. Utomo, and A. Mulwinda, "Automated short essay scoring system using normalized Simpson methods," in Proceedings of the 6th International Conference on Education, Concept, and Application of Green Technology, Semarang, Indonesia, 2018.

[36] L. dela-Fuente-Valentín, E. Verdú, N. Padilla-Zea, C. Villalonga, X. P. Blanco Valencia, and S. M. Baldiris Navarro, "Semiautomatic Grading of Short Texts for Open Answers in Higher Education," in Higher Education Learning Methodologies and Technologies Online, 2022, pp. 49-62.

[37] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching," in Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Denver, Colorado, 2015, pp. 97-106.

[38] I. G. Ndukwe, B. K. Daniel, and C. E. Amadi, "A Machine Learning Grading System Using Chatbots," in Artificial Intelligence in Education, 2019, pp. 365-368.

[39] H. Qi, Y. Wang, J. Dai, J. Li, and X. Di, "Attention-based hybrid model for automatic short answer scoring," in Simulation Tools and Techniques: 11th International Conference, SIMUtools 2019, Chengdu, China, July 8–10, 2019, Proceedings 11, 2019.

[40] M. Uto and Y. Uchida, "Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory," in Artificial Intelligence in Education, 2020, pp. 334-339.

[41] K. Sakaguchi, M. Heilman, and N. Madnani, "Effective feature integration for automated short answer scoring," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, 2015, pp. 1049-1054.

[42] N. LaVoie, J. Parker, P. J. Legree, S. Ardison, and R. N. Kilcullen, "Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test," Educational and Psychological Measurement, vol. 80, no. 2, pp. 399–414, Jul. 2019, doi: https://doi.org/10.1177/0013164419860575.

[43] R. Agarwal, V. Khurana, K. Grover, M. Mohania and V. Goyal, "Multi-Relational Graph Transformer for Automatic Short Answer Grading," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022, pp. 2001-2012.

[44] H. Oka, H. T. Nguyen, C. T. Nguyen, M. Nakagawa and T. Ishioka, "Fully Automated Short Answer Scoring of the Trial Tests for Common Entrance Examinations for Japanese University," in International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science, vol 13355. Springer, Cham.

[45] J. Sawatzki, T. Schlippe and M. Benner-Wickner, "Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers," in Artificial Intelligence in Education: Emerging Technologies, Models and Applications: Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology, 2022, Lecture Notes on Data Engineering and Communications Technologies, vol 104. Springer, Singapore.

[46] K. Steimel and B. Riordan, "Toward instance-based content scoring with pretrained transformer models," in Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, vol. 34.

[47] C. Sung, T. I. Dhamecha and N. Mukhi, "Improving short answer grading using transformer-based pretraining," in Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I, 2019.

[48] S. Takano and O. Ichikawa, "Automatic scoring of short answers using justification cues estimated by BERT," in Proceedings of the

17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Seattle, Washington, 2022, pp. 8-13.

[49] H. A. Ghavidel, A. Zouaq and M. C. Desmarais, "Using BERT and XLNET for the Automatic Short Answer Grading Task," in Proceedings of the 12th International Conference on Computer Supported Education - Volume 1: CSEDU, 2020, pp 58-67.

[50] R. Somers, S. Cunningham-Nelson, and W. Boles, "Applying natural language processing to automatically assess student conceptual understanding from textual responses," Australasian Journal of Educational Technology, vol. 37, no. 5, pp. 98–115, Dec. 2021, doi: https://doi.org/10.14742/ajet.7121.

[51] J. Garg, J. Papreja, K. Apurva, and G. Jain, "Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading," presented at the 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6.

[52] V. Ramnarain-Seetohul, V. Bassoo, and Y. Rosunally, "Work-in-Progress: Computing Sentence Similarity for Short Texts using Transformer models," presented at the 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 2022, pp. 1765-1768.

[53] M. H. Haidir and A. Purwarianti, "Short answer grading using contextual word embedding and linear regression," Jurnal Linguistik Komputasional, vol. 3, no. 2, pp. 54-61, 2020.

[54] J. H. Clark, D. Garrette, I. Turc, and J. Wieting, "Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation," Transactions of the Association for Computational Linguistics, vol. 10, pp. 73–91, 2022, doi: https://doi.org/10.1162/tacl_a_00448.

[55] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[56] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining 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, Vol. 1 (Long and Short Papers), 2019, pp. 4171-4186.

[57] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.

[58] B. Rodrawangpai and W. Daungjaiboon, "Improving text classification with transformers and layer normalization," Machine Learning with Applications, vol. 10, p. 100403, Dec. 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100403.

[59] K. Song, X. Tan, T. Qin, J. Lu and T. Y. Liu, "Mpnet: Masked and permuted pretraining for language understanding," in Advances in Neural Information Processing Systems, vol. 33, pp. 16857-16867, 2020.

[60] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[61] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66-71, 2018.

[62] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1715-1725, 2016.

[63] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., "Attention is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[65] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II, vol. 21, pp. 43-48, Springer International Publishing, 2020.

[66] J. Seo, S. Lee, and L. Liu, "TA-SBERT: Token Attention Sentence-BERT for Improving Sentence Representation," IEEE Access, vol. 10, pp. 39119-39128, Apr. 2022, doi: https://doi.org/10.1109/ACCESS.2022.3164769.

[67] F. García-Peñalvo, & A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 4, pp. 7-16, 2023, https://doi.org/10.9781/ijimai.2023.07.006

[68] J. M. Flores-Vivar, & F. J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)". Comunicar, vol. 31, no. 74, pp. 37-47, 2023, https://doi.org/10.3916/C74-2023-03

Martinus Maslim

Martinus Maslim earned a master's in informatics engineering in 2012 from Universitas Atma Jaya Yogyakarta, Indonesia. His areas of interest are artificial intelligence, information systems, and data science. Since 2013, he has been a lecturer in the Department of Informatics at Universitas Atma Jaya Yogyakarta, and he is now studying for his PhD at National Cheng Kung University, Taiwan. He is part of the Web Knowledge Discovery Lab.

Hei-Chia Wang

Hei-Chia Wang is a Professor from the Institute of Information Management at National Cheng Kung University, Taiwan. He got a doctoral degree from The University of Manchester in 1999. His research interests in natural language processing, e-learning, bioinformatics, information retrieval, and knowledge discovery. Currently, he is the leader of the Knowledge Discovery Lab.

Cendra Devayana Putra

Cendra Devayana Putra earned a master's degree in the Institute of Information Management at National Cheng Kung University, Taiwan. His research areas include natural language processing, deep learning, and management science. He is currently a doctoral student at National Cheng Kung University, Taiwan. He is also part of the Web Knowledge Discovery Lab.

Yulius Denny Prabowo

Yulius Denny Prabowo is a lecturer at Computer Science Department, Bina Nusantara University Jakarta, Indonesia. He finishes his doctoral degree from Bina Nusantara University in 2022. His research interests are natural language processing, machine learning, and deep learning.