Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 3, September, 2024

Transformer-Based Sequence Modeling Short Answer Assessment Framework

P. Sharmila [1], Kalaiarasi Sonai Muthu Anbananthen [2*], Deisy Chelliah [1], S. Parthasarathy [1], Baarathi Balasubramaniam [2], Saravanan Nathan Lurudusamy [3]

[1] *Thiagarajar College of Engineering, Madurai, Tamilnadu, 625015, India.*

[2] *Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[3] *Division Consulting & Technology Services, Telekom Malaysia, Kuala Lumpur 50672, Malaysia.*

Received 09 May 2024; Revised 11 August 2024; Accepted 18 August 2024; Published 01 September 2024

Abstract

Automated subjective assessment presents a significant challenge due to the complex nature of human language and reasoning characterized by semantic variability, subjectivity, language ambiguity, and judgment levels. Unlike objective exams, subjective assessments involve diverse answers, posing difficulties in automated scoring. The paper proposes a novel approach that integrates advanced natural language processing (NLP) techniques with principled grading methods to address this challenge. Combining Transformer-based Sequence Language Modeling with sophisticated grading mechanisms aims to develop more accurate and efficient automatic grading systems for subjective assessments in education. The proposed approach consists of three main phases: Content Summarization: Relevant sentences are extracted using self-attention mechanisms, enabling the system to effectively summarize the content of the responses. Key Term Identification and Comparison: Key terms are identified within the responses and treated as overt tags. These tags are then compared to reference keys using cross-attention mechanisms, allowing for a nuanced evaluation of the response content. Grading Process: Responses are graded using a weighted multi-criteria decision method, which assesses various quality aspects and assigns partial scores accordingly. Experimental results on the SQUAD dataset demonstrate the approach's effectiveness, achieving an impressive F-score of 86%. Furthermore, significant improvements in metrics like ROUGE, BLEU, and METEOR scores were observed, validating the efficacy of the proposed approach in automating subjective assessment tasks.

Keywords: Attention Model; Sequence Language Modeling; Subjective Assessment; Transformer.

1. Introduction

E-learning enables students to learn via the Internet. Online learning requires two prerequisites: learning resources and automatic assessments. Liu et al. [1] developed a semi-automated method for generating grammatical assessment tasks using NLP techniques. Automatic assessments for online examinations are challenging since subjective questions require human judgment and often involve complex reasoning, creativity, and language understanding. Researchers such as Ateeq et al. [2], Paiva et al. [3], and Ramesh & Sanampudi [4] have addressed this challenge by employing NLP and machine learning techniques to evaluate essay quality automatically.

* Corresponding author: kalaiarasi@mmu.edu.my

http://dx.doi.org/10.28991/HIJ-2024-05-03-06

In general, there are two categories of question types: objective questions and subjective questions. Figure 1 shows the different modalities of assessment methods. Objective questions, such as multiple-choice, yes-no, matching, and fill-in-the-blank, are easily graded by automated systems and encourage rote learning. Subjective questions, like short-answer and long-answer, require open-ended responses. Short answer questions applicable for assessing factual knowledge, definitions, key concepts, and basic application of knowledge. Long answer essays for promoting critical thinking, analysis, and creative expression. While objective questions are common in computerized tests for their quick and uniform evaluation, subjective assessments allow learners to explain concepts in their own words. However, they can be challenging to evaluate due to potential lexical or semantic similarities.

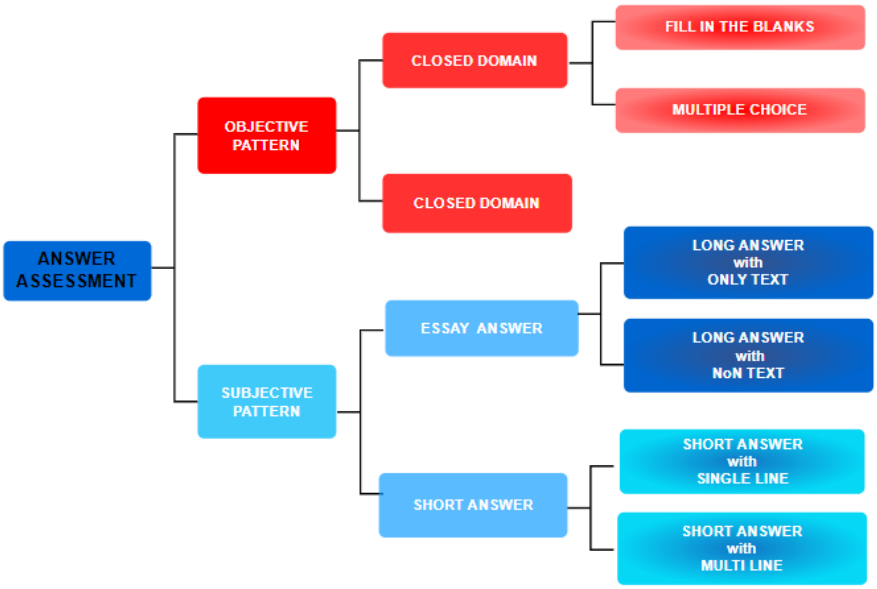

Figure 1. Different Modalities of Answer Evaluation

Short-answer subjective assessments are prevalent in education, language proficiency tests, content creation, and research, necessitating qualitative evaluation of responses. Objective questions are prevalent in computerized tests due to their quick, reliable, and uniform evaluation, typically focusing on correctness. Conversely, subjective assessments involve open-ended responses, which enable the learners to conceive and write an explanation in their own words that is challenging to evaluate, and subjective answers may share lexical or semantic similarities. Short-answer subjective assessments are common in education, language proficiency tests, content creation, and research, requiring qualitative evaluation of open-ended responses.

1.1. Motivation and Objectives

The motivation behind Short Answer Assessment stems from the inefficiency, inconsistency, and limited feedback provided by manual grading, particularly in large classes. Traditional methods, such as keyword matching or rule-based systems, fail to understand complex language and offer detailed feedback. The Transformer approach utilizes Transformer models to delve into the deeper meaning of responses, offering flexibility and the ability to provide targeted feedback on strengths and weaknesses. Hence, the ultimate goal is to develop an automated assessment system that is efficient, consistent, and informative for short-answer responses.

1.2. Overview of Subjective Automatic Assessment

Neshan & Akbari [5] proposed a hybrid approach combining lexicon-based and machine-learning techniques for short answer assessment, which showed significant improvements but faced increased noise levels with large datasets. Zhu et al. [6] discussed Transformer-based language models as promising tools for automatically scoring short written responses, offering a potential solution to the challenges of computerized evaluation of subjective questions. However, the accuracy of computer-based evaluation remains insufficient, and grading subjective questions manually is time-consuming and expensive. Despite this, multiple-choice questions have replaced subjective questions in many computerized exams due to their consistent evaluation but inability to assess writing abilities and critical reasoning. Automatic evaluation aims to provide timely and accurate feedback to students and instructors while reducing grading time and resources. Various techniques, such as those for short answer and essay questions, contribute to automated assessment methods.

Essay Questions: The computer program utilizes machine learning methods to assess the quality of the response to these longer-form written questions. Das & Majumder [7] proposed a method for extracting factual open-ended questions from text by identifying informative sentences.

***Short Answer Questions**:* The computer program employs natural language processing to assess the answers to these questions, which call for a brief written response. However, short-answer assessment, like any other form, has its own challenges. Here are some common challenges faced in short answer assessment: Ambiguity, Lack of context, Grading consistency, Limited response options, Redundant content (Cheating), Time constraints, and Subjectivity. Overall, short answer assessment can be an effective method for evaluating knowledge and understanding, but it requires careful consideration of the challenges and limitations of this form of assessment.

Automatic subjective assessment systems offer numerous advantages, including:

- Time and Resource Savings: These systems free up instructors' time and resources, allowing them to concentrate on other teaching and learning tasks.
- Immediate and Consistent Feedback: Students receive prompt and consistent feedback, aiding them in enhancing their performance and understanding.
- Objective and Unbiased Evaluation: By minimizing human involvement, these systems reduce the risk of grading errors or bias, ensuring fairness in assessment.
- Data-Driven Decision-Making: Automatic subjective assessment systems generate valuable insights into student performance and learning outcomes, enabling educators to make informed decisions based on data.

However, automatic assessment systems also have some limitations, such as:

- Difficulty in evaluating complex or creative responses that require human judgment.
- Limited understanding of the context and nuances of language may result in potential inaccuracies during evaluation.
- Developing and maintaining accurate and reliable assessment algorithms poses significant challenges.
- Language Modeling Automatic Assessment.

The rest of the paper is organized as follows. Section 2 describes the related works. Section 3 discusses the proposed framework for assessing answers, including an overview of the similarity measure and fusion technique. Section 4 explains the dataset and experimental setup. Section 5 shows the results and discussions. Finally, in Section 6, we conclude our paper by highlighting.

2. Literature Reviews

Every university has a unique evaluation procedure built on a reflective analysis. Hence, it is imperative to consider the assessment and evaluation conducted by computer-assisted appraisal systems as ICT-based teaching-learning approaches continue to expand. Automatic subjective assessment is a broad area of NLP and Artificial Intelligence (AI) research that focuses on developing methods to automatically evaluate subjective content, such as user-generated reviews, opinions, essays, and more [8]. In recent years, Transformer has advanced in resolving difficulties related to automatic subjective assessment.

2.1. Short Answer Assessment

Various features such as sentence length, word placement in a phrase, a chunk of the sentence, the verb, parts of speech, named entities, recognized words, unknown words, acronyms, and other linguistic elements were leveraged to train the SVM classifier, as demonstrated by Leacock et al. [9]. Meanwhile, Matsumori et al. [10] used a summarizer (MEAD) to directly select informative sentences for automatic CQ generation. Additionally, Feng et al. [11] use statistical measures, such as a vector space model, to calculate semantic relatedness between words or sentences.

2.2. Text Similarity

Measurement of text similarity plays a key role in assessment tasks by comparing two or more texts to determine how similar or related they are to each other. Sahu & Bhowmick [12] demonstrate that grading student responses is improved by combining various graph alignment criteria with lexical semantic similarity metrics. There are several methods and techniques used to measure text similarity, including Cosine similarity, as discussed by Rosnelly et al. [13], Jaccard similarity, and the Dice Coefficient, outlined by Wahyuningsih et al. [14]. Additionally, edit distance, as studied by Anbananthen et al. [15], and Latent Semantic Analysis (LSA), as described by Kaur & Sasi [16], are utilized for this purpose. Furthermore, Word embedding, a method represented by Mardini et al. [17], involves representing words as vectors in a high-dimensional space based on their context and co-occurrence in a large corpus of text.

2.3. Attentions – Transformer

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. Transformer models apply an evolving set of mathematical techniques, called Attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other. Hence, in our proposed model, cross-attention is incorporated along with self-attention for efficient computation and better performance.

The utilization of BERT in grading brief answers is implemented, as demonstrated by Bexte et al. [18]. Bonthu & Sree et al. [19] lays the foundation for the self-attention mechanism, a crucial component in Transformer-based models used for subjective assessment. ALBERT is a variation of BERT that achieves state-of-the-art results in various NLP tasks, making it also relevant for subjective assessment tasks, as highlighted by Lan et al. [20]. Klyuchnikov et al. [21] evaluated the performance of various NLP models, including Transformer-based models, on various tasks using the GLUE benchmark. Moreover, Khodeir [22] integrated BERT with a multi-layer bi-GRU to enhance the MOOC question-answer forum. Large transformer-based neural networks such as T5 by Raffel et al. [23], GPT by Alec et al. [24], and OPT by Frantar et al. [25] have been applied in question-answering tasks. BART, a pre-trained transformer model, has been applied in machine translation and question answering, as demonstrated by La Quatra & Cagliero [26]. Lastly, Zhu et al. [27] focused on feature engineering and demonstrated that ensemble-based models significantly outperform individual regression models.

2.4. Research Gaps

Automatically evaluating their responses remains an intriguing problem. Specifically, scoring assessments for open-ended, short-answer responses provide several difficulties.

- The response is written naturally, and students are free to react in a way that involves complex reasoning, creativity, and language understanding.
- Even for short answers, they can write up to two pages of an answer with redundant content, which increases training time.
- Freestyle writing requires semantic similarity.
- Extracting relevant content without redundancies and Choosing text similarity measures to maintain semantics are also bottlenecks in short answer assessment.
- Assigning partial marks is also considered a major problem.

The objective is to assess variations in cognitive abilities and linguistic patterns among students through the quantitative analysis of lexical features present in their written compositions. Our proposed model employs neural networks with transformers. The Major Contributions are

- Construct short responses by extracting key sentences and avoiding redundancies using the self-attention encoder model.
- Extract keywords from the extracted responses using cross attention encoder and compare them with external reference or standard key answers to compute similarity.
- Grade the answer based on the similarity score.
- A decision-making system that utilizes Multiple Criteria Decision Making (MCDM) is carried out, encompassing a range of weights for partial correctness in responses.

3. Research Methodology

Subjective short-answer assessment involves evaluating NLP algorithms. The student's answer is first subjected to the extraction phase to extract the relevant and informative sentences, thereby avoiding redundancies. Next, the extracted sentences are subjected to the Attention-based similarity computation phase model. Keywords (open tags) are extracted using POS tags since most of the keys are nouns, verbs, adverbs, and adjectives. Then, the extracted words are dot products with external or standard reference keywords or phrases to compute the similarity. Multiple criteria are employed to assess answers, thereby reducing human manual work. Figure 2 shows the workflow of our proposed model. A multi-criteria decision-making approach evaluates student responses using model answers from textbooks and subject specialists.

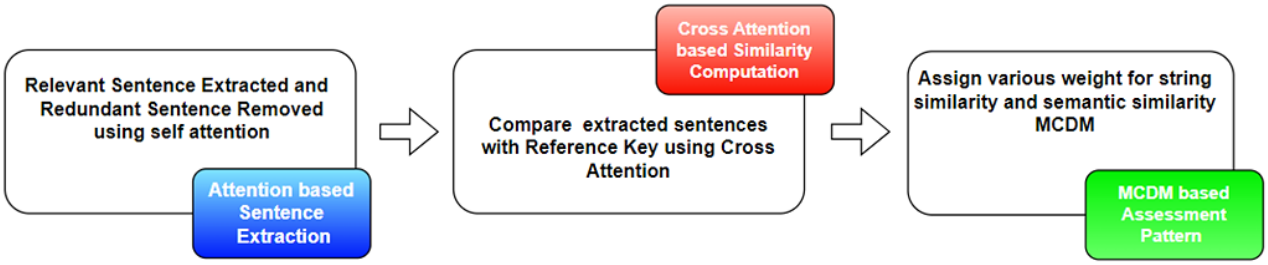

Figure 2. Workflow of Proposed Assessment Model

The proposed approach consists of three main phases:

Content Summarization: Relevant sentences are extracted using self-attention mechanisms, enabling the system to effectively summarize the content of the responses. By weighing one token's relevance in relation to others, self-attention enables the formation of a weighted representation that captures contextual dependencies. This considers long-range dependencies and contextual subtleties, facilitating a comprehensive interpretation of input sequences.

Key Term Identification and Comparison: Key terms are identified within the responses and treated as overt tags. These tags are then compared to reference keys using cross-attention mechanisms, allowing for a nuanced evaluation of the response content. Cross-attention mechanisms may also be employed in this phase, where the model attends to both the input sequence and additional context information (e.g., grading rubrics or teacher feedback) to generate the final output.

Grading Process: Responses are graded using a weighted multi-criteria decision method, which assesses various aspects of quality and assigns partial or sub-scores accordingly.

Figure 3 shows the First phase of the proposed subjective answer evaluation model. In the second phase, a cross-attention encoder is employed to compute the similarity between student answers and references. Most keywords come under "nouns, verbs, adverbs, and adjectives". The extracted sentences use the cross-attention encoder to filter the "set of nouns, verbs, adverbs, and adjectives" words. Then, compute the similarity between these answers. Figure 4 shows the second phase of the proposed similarity computation model. In the third phase, the answer was graded using MCDM, which considered the previous phase text or keyword matching and content or semantic matching as criteria to grade the answer.

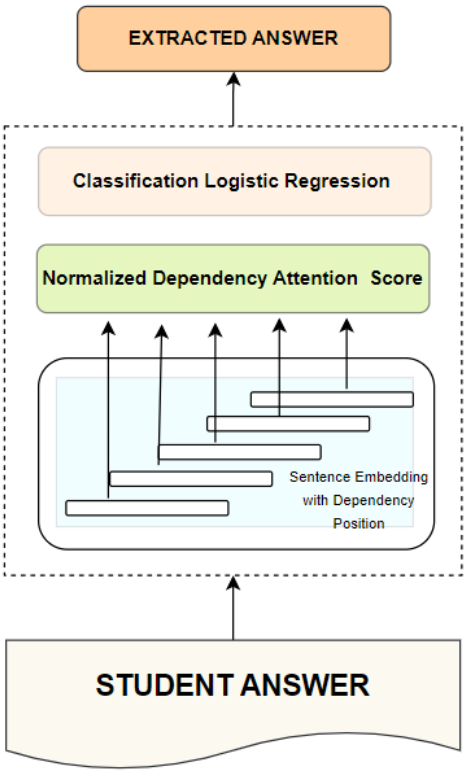

Figure 3. Subjective Answer Evaluation

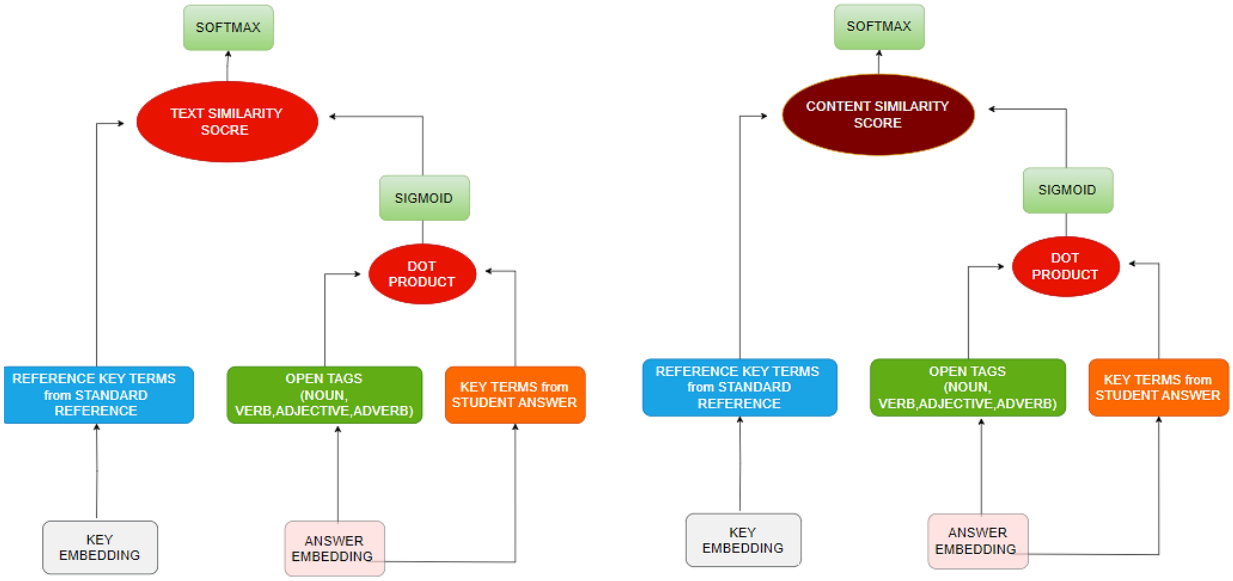

Figure 4. The second phase of the proposed similarity computation model: a) Algorithm 1: employs a Text similarity attention score for answer matching; b) Algorithm 2: employs Content Similarity Score for content matching

3.1. Attention-based Sentence Extraction

In this approach, the model selects important sentences or phrases from the original response to form the summary. It may employ techniques like sentence ranking based on features or graph-based algorithms. For lengthy sequences, self-attention excels ahead of CNN at feature extraction. Self-attention is applied to each embedding to identify the interdependencies and associations between the sentences. Self-attention has computational complexity bounds for lengthy sequences.

In contrast to the CNN and RNN frameworks, Attention is parallelizable. Hence, self-attention is incorporated in the first phase to extract interdependent sentences. But our proposed model is a transformer-based one. In the first phase, doc2vec, as introduced by Mikolov et al. [28], is used to create a sentence embedding from a text excerpt.

3.2. Cross Attention-based Similarity Computation

Text Similarity and Context Similarity are the two criteria considered in our proposed work. Cross-attention is employed to compute the similarity between the student and reference answers. Mostly, all the keywords are - nouns, verbs, adverbs, and adjectives [29]. From the extracted sentences, using a cross-attention encoder at the token level and filter, only the open tags assigned tokens as a set of nouns, verbs, adverbs, and adjectives. Then, compute the similarity between these answers.

The complexity of this work is that the solution can be written in several ways, such as utilizing active or passive phrases, synonyms, or word forms. Thus, we must analyze many forms of similarity to evaluate an answer. Here, we present two techniques, one for computing textual similarity (Algorithm 1) and the other for calculating semantic or contextual similarity (Algorithm 2).

Algorithm 1:

The algorithm can be used to evaluate how similar a student's answer is to the provided model answers. The higher the Max value, the more similar the student's answer is to at least one of the model answers based on the chosen text similarity metric. Depending on the specific text similarity metric used (Jaccard similarity), the algorithm will yield different results, so the choice of similarity metric should be based on the specific context and requirements of the task.

Algorithm 2:

After looping through all model answers and comparing them to the student's answer, return the value of Max. This represents the maximum semantic similarity between the student's answer and any model answer. This algorithm determines how similar a student's answer is to the provided model answers. It can be used to assess the quality of the student's response by finding the closest match among the model answers in terms of semantic content. The higher the Max value, the more similar the student's answer is to at least one of the model answers.

Algorithm 1 Text Similarity (TS)"

```
Input:
MA: A set of model answers (ma1, ma2, ..., map).
SA: The student's answer for comparison.
Output:
Max: The maximum text similarity score between sa and any model answer.
Initialization:
Set Max to 0, initially.
For each ma in MA, do the following:
Calculate the TSim(ma, sa) text similarity score between the model answer ma and the
student's answer sa.
// Computes text similarity using Jaccard similarity.
Update Maximum Similarity:
Check if TSim(ma, sa) is greater than the current value of Max.
 If it is, update Max with the new similarity score.
Return Maximum Similarity:
```

Algorithm 2 Context Similarity (CS)

```
Input:
MA: A set of model answers (MA1, MA2, ..., MAp).
SA: The student's answer for comparison.
Output:
Max: The maximum semantic similarity score between SA and any model answer.
Initialization:
Set Max to 0, initially.
For each ma in MA, do
Calculate the content similarity score CSim(MA, SA) between the model answer ma and
the student's answer SA.
// computes the semantic similarity between the text.
Update Maximum Similarity
Check if CSim(MA, SA) is greater than the current value of Max.
If it is, update Max with the new similarity score.
Return Maximum Similarity:
```

3.3. MCDM-Based Assessment Pattern

Evaluation and scoring is the last phase. Assessment depends on the learning domain, question type, complexity, scoring techniques, and total score. Hence, it is important to note that subjective answer evaluation based on relevant information may involve some degree of subjectivity, as it requires interpreting the depth of understanding, creativity, and original thought exhibited in the response. Therefore, clear and well-defined rubrics and consistent evaluation practices are essential to ensure fairness and accuracy in the assessment process. We have employed a variant of MCDM-based fusion.

The first step is to assign weights to each criterion, which may depend on the evaluators. Here, for partial answer assessment, the weight for text similarity is 0.3 and 0.7 for context similarity. Determining the weights of criteria poses a key problem in multi-criteria decision-making, as highlighted by Žižovic & Pamucar [30], Bhole & Deshmukh [31], Odu [32]. The weighted sum is then computed with criteria to combine the normalized scores as given.

This will result in a single value that represents the overall performance. Then, rank the alternatives based on the aggregated scores. The higher the score, the better the performance, and the moderate score for partial correctness. By changing the criteria, weights impact the final ranking or assessment. The final score for each short answer is calculated as:

$$Final\ Marks = [Text\ Similarity\ (\alpha) \times TSWF] + Context\ Similarity\ (\beta) \times CSWF \quad (1)$$

Where, α is the text similarity score, β is the content similarity score, $TSWF$ is the text similarity weight factor, and $CSWF$ is the content similarity weight factor.

The weight factor is a random value given by question setters, which depends on them to assign a partial score for each answer.

4. Dataset and Experimental Setup

4.1. Dataset

A sophisticated open QA system can be created using the dataset. This dataset can be expanded for our proposed assessment application by incorporating our generated dataset. Over 1000 K samples from Wikipedia articles make up the collection. Each sample consists of a passage and question-answer pairs. The SQuAD dataset is used as a benchmark to assess the proposed method's ability to replicate human expert assessments on short answer responses. Improvements in ranking accuracy and semantic similarity scores are key success indicators. These findings would demonstrate the Transformer-based framework's effectiveness in automating short response assessments[*].

4.2. Experiments

At the sentence extraction level, several pre-processing techniques are needed. Remove punctuation and symbols from student responses using materials to embed the sentences that were parsed out of the pre-processed responses. Applying the proposed dependency extraction model on a sentence-by-sentence basis can help choose which phrases need to be extracted; irrelevant and duplicated phrases can be avoided based on the normalized attention score.

At the second level of cross-attention-based similarity computation. Pre-processing is done as tokenization, and a part-of-speech (POS) tag is assigned for each token [28]. Dot product attention was computed (Algorithm,1) to extract only open tag words such as nouns, verbs, adverbs, and adjectives, and most of the key or reference answers are of only open tag words. Then again, a dot product with a standard answer to get the final score (Algorithm.2). At the end of this level, two similarity scores are obtained.

Lastly, the grading level, TSWF (Text Similarity Weight Factor) values for measuring text similarity, and SSWF (Semantic Similarity Weight Factor) for measuring semantic similarity are predetermined and dependent on the evaluation experts. Weight factor assigned by the evaluator. Different values are trained for better performance. Hence, the TSWF value of 0.3 and CSWF of 0.7 are fixed. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text.

4.3. Evaluation Metrics

Evaluation metrics such as ROUGE, BLEU, and METEOR score are used at the sentence extraction level to compare the generated summaries with the reference summaries. These metrics assess the quality of the summaries based on factors like content overlap, grammaticality, and coherence. The data was assessed using three metrics, specifically BLEU introduced by Fabbri et al. [33], which measures the similarity in terms of precision for n-grams, and ROUGE by Barbella & Tortora [34], which captures different aspects of text quality including overlap and longer contiguous sequences.

METEOR, as described by Saadany & Orasan [35], is considered a synonym and stemming, making it more robust than BLEU or ROUGE. No single metric is perfect, and it is often advisable to use multiple metrics to gain a more comprehensive understanding, especially for tasks that require high levels of fluency and semantic accuracy. The evaluation process involves comparing the responses generated by the algorithm and the gold standard replies annotated by human evaluators. Again, the aggregate performance will improve if multiple similarity metrics are combined.

5. Results and Discussion

Consider the evaluation query "what is "JVM?" If the student's response to this query exceeds 100 to 130 characters, it is assumed to be lengthy. This response is fed into our proposed answer extraction model to abbreviate the response without compromising its content and eliminate redundant content. By feeding our suggested model responses of varied lengths (minimum lengths of 40 and 70 characters, respectively), When the maximum length of the student's answer is 100 characters, and the extracted answer length is 40 characters, the F-Measure is 0.84, indicating a relatively good accuracy in extracting key terms. Similarly, for a maximum answer length of 130 characters and an extracted answer length of 60 characters, the F-Measure improves to 0.86, increasing accuracy. At this juncture, we use the ROUGE (R1, R2, and RL) scores against reference answers by subject experts to calculate the similarity score. However, it is interesting to note that in some cases, such as when the maximum answer length is 100 characters and the extracted answer length is 60 characters, the F-Measure drops slightly to 0.82, indicating a decrease in accuracy compared to shorter extracted answers. These results suggest that there may be an optimal balance between the length of the extracted answer and the accuracy of key term extraction. The observed ROUGE scores are summarized in Table 1. Table 2 also shows the result of the F-Measure using POS tags.

[*] https://www.kaggle.com/datasets/ananthu017/squad-csv-format

Table 1. Similarity scores of our proposed Sentence Extraction model

Max Length of Student Answer	Extracted Answer Length	Similarity Scores		
		R1	R2	RL
100	40	0.432	0.410	0.384
	60	0.449	0.399	0.391
130	40	0.489	0.399	0.391
	70	0.512	0.493	0.384

Table 2. Result of F-Measure using POS Tags

Maximum Length of Student Answer	Extracted Answer Length	Extracted Key Term	Similarity Scores
			F-Measure
100	40	0.81	0.84
	60	0.85	0.82
130	40	0.84	0.86
	60	0.91	0.84

The results of the sentence extraction model are shown in Figure 5 showcases the effects of adjusting the summary length and student answers with varying lengths. The analysis involves examining the extracted lengths of 40 and 70, as well as the maximum or actual length of 100 and 130, using R1, R2, and RL to achieve improved outcomes. The sentence length for the extracted text has been set to a fixed value of 70, resulting in a notable improvement in the evaluation metrics. Specifically, the achieved scores are 0.512 for R1, 0.493 for R2, and 0.384 for the RL score.

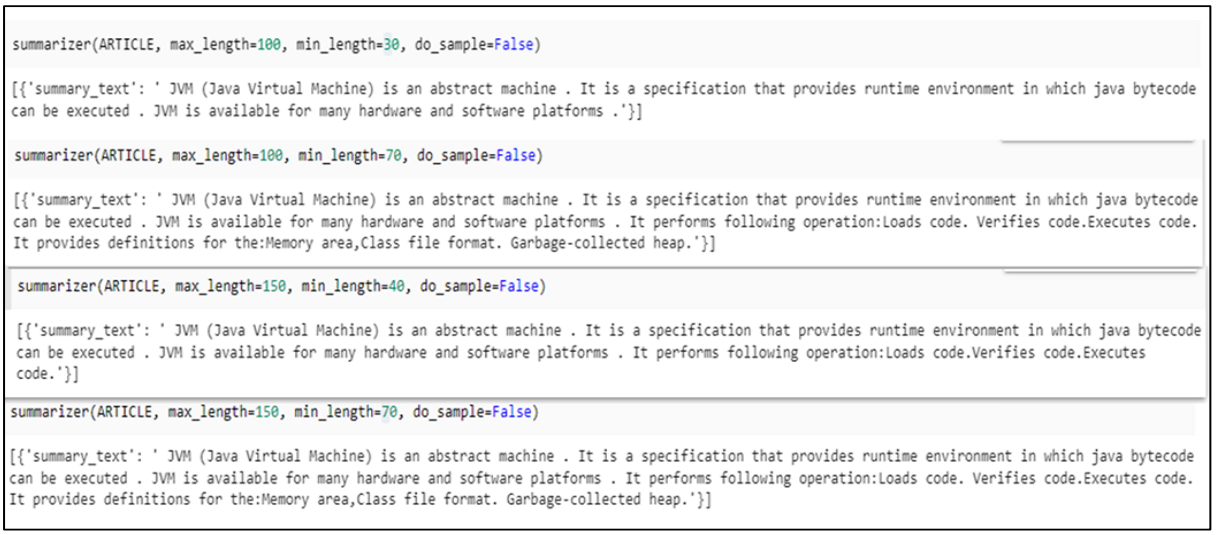

Figure 5. Result of Extractive Summary with Varying Summary Length and Student Answer

The maximum length of the student's answer is less than the threshold of 100 or 130. There is no need to process the first level. Similarity scores are computed by varying the extracted summary length as 40 and 70. The following code ("summarizer (ARTICLE, max_length=130, min_length=60)") is to adjust the summary length.

The proposed model results in minimum training time and better F1-Score than the BERT model. Table 3 Compare the result of our proposed model on the SQuAD Dataset in terms of BLEU, METEOR and ROUGE. A small increment in all these above values shows that our proposed model is better. Table 4 shows Test Accuracy– epochs and 10 epochs with Training Time for DNN models.

Table 3. Similarity Compared Result of our proposed model on the SQuAD Dataset

Model	BLEU	METEOR	ROUGE
BERT	0.641	0.243	0.486
Proposed Assessment Model	0.699	0.259	0.553

Table 4. Test Accuracy– epochs and 10 epochs with Training Time he SQuAD Dataset

Models	Test Accuracy (5Epochs)	Test Accuracy (10Epochs)	Time to Train Epoch (seconds)
BERT	80.121	81.005	29.501
Proposed Assessment Model	81.880	81.920	28.475

The subsequent phase involves identifying open (POS) tags for nouns, verbs, adverbs, and adjectives that align with the keywords discovered in the preceding step and shown in Figure 6. Figure 4(a) illustrates the utilization of text similarity cross-attention to extract keywords from the preceding level. Conversely, Figure 4(b) demonstrates the application of content similarity cross-attention for keyword extraction. The similarity score is calculated using the dot product to measure the focus between the extracted terms and the reference keywords or responses. The sigmoid activation function is afterwards employed to ascertain the relevance of the student's response. In the computation of a final score, only questions that have received suitable responses are considered. TSWF and CSWF are introduced in our evaluation approach; however, they are only for partial scores.

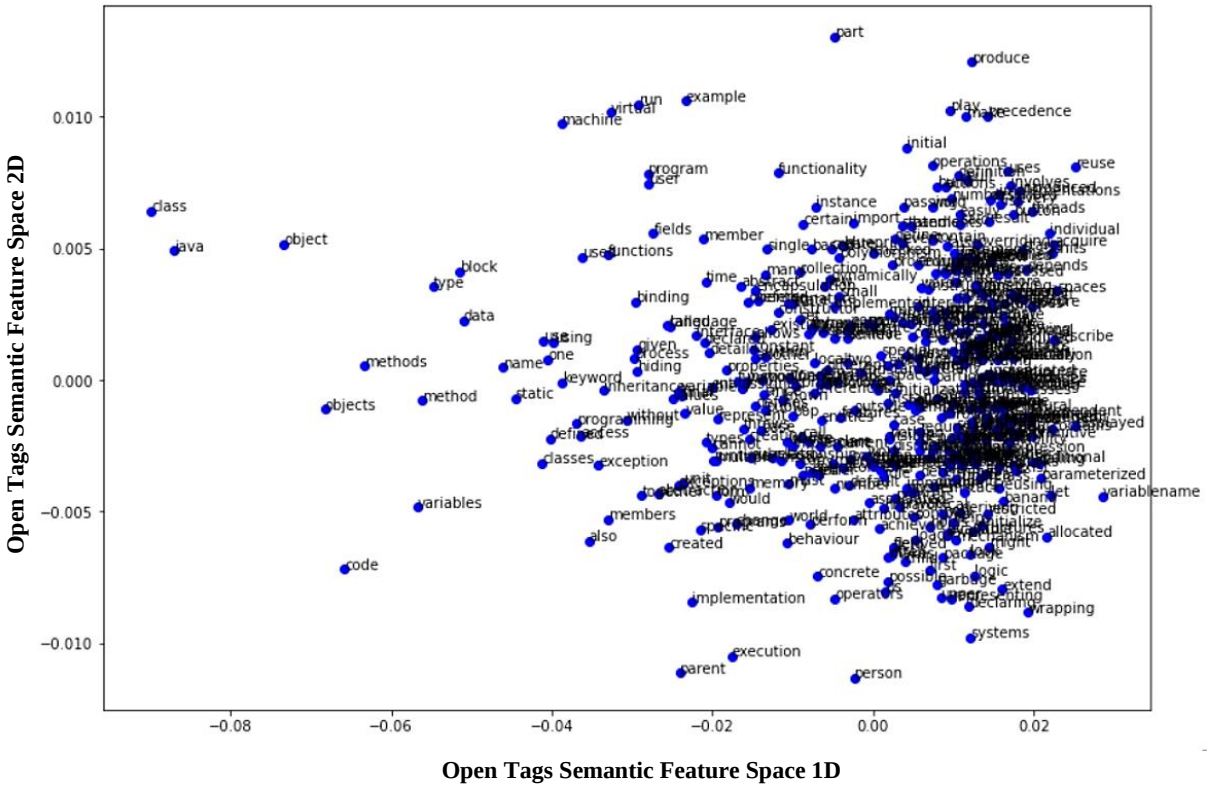

Figure 6. Extracted keywords using open (POS) tags

5.1. Comparative Analysis for Short Answers

Evaluating freestyle short answer assessments requires more sophisticated metrics than Exact Match (EM) due to the variability in student responses. Three effective alternatives are ROUGE, BLEU, and METEOR, considered with the F-measure to provide a robust evaluation. Table 5 analyses the various metrics for answer evaluation.

- ***ROUGE:*** Useful for identifying common sequences and n-grams between student and reference answers.

Example: Captures common sequences like "The cat sat on the mat" and "The cat is sitting on the mat."

- ***BLEU:*** Adapted for assessing the similarity in wording and structure between student and reference answers.

Example: Measures similarity for phrases like "The quick brown fox jumps over the lazy dog" and "A fast brown fox leaps over the lazy dog."

- ***METEOR:*** Handles paraphrased content effectively by recognizing semantic similarities.

Example: Recognizes "The cat sat on the mat" and "The feline rested on the rug" as similar.

Table 5. Comparative Analysis for Short Answers on SQuAD Dataset

Author/Year	Model	EM (%)	ROUGE	BLUE	METEOR
Muludi et al. (2024) [36]	RAG	-	-	0.568	-
Liu et al. (2019) [37]	RoBERTa	88.9	-	-	-
Yang et al. (2019) [38]	XLNet	89	0.4820	-	-
Chen et al. (2019) [39]	BERT	-	-	0.617	0.752

Table 6 displays the subjective assessment pattern, wherein the values for TSWF and CSWF are 0.3 and 0.7, respectively. The computation of the Sum score and Product score involves utilizing the variables α and β. A product will be awarded if it achieves a score greater than 80% of the maximum possible score. If the product's score falls within the range of 50% to 80%, a score equivalent to half of the total will be assigned. Conversely, if the product's score falls from 0% to 50%, a mark of 25% will be allocated. The decision criteria for the supplied question were text similarity and content similarity, both of which were simple to grant partial marks for.

Table 6. Subjective Assessment Pattern

Sample Answers	Similarities Weight factor				Sum Score	Proposed Assessment Model	
	Text Similarity (α)	TSWF = 0.3 (α×0.3)	Semantic Similarity (β)	SSWF = 0.7 (β×0.7)		Product score	Accurate Marks
1	0.6	0.18	0.9	0.63	0.15	0.81	2
2	0.2	0.06	0.5	0.35	0.7	0.021	0.5
3	0.1	0.3	0.3	0.21	0.4	0.063	1
4	0.7	0.21	0.5	0.35	0.12	0.94	2

6. Conclusion

Our proposed method, which integrates attention-based transformer encoding with collaborative decision-making mechanisms, significantly advances automated subjective assessment for short-answer responses. This approach addresses the challenges of accurately ranking student responses by considering both semantic meaning and textual similarity. Content summarizing extracts vital content, reducing duplicate or redundant information, significantly boosting training speed and reducing computational resources. In linguistic-based keyword extraction, the system analyzes context and word relationships to extract essential elements, allowing for deeper understanding. Equitable assessment and grading are achieved by focusing on key elements and mitigating the influence of irrelevant information, resulting in fairer and more consistent grading practices. This approach removes bias based on writing style or superfluous details, ensuring all students compete equally. The strong performance on the SQUAD dataset demonstrates its effectiveness.

However, there are areas for further refinement. The success of our method hinges on careful keyword selection, which requires additional research to ensure its efficacy across diverse datasets. Specifically, the ideal weighting of keywords within queries plays a crucial role in optimizing performance. Therefore, our future efforts will focus on tackling these challenges. We aim to develop solutions that enhance the robustness and generalizability of our method, allowing it to be effectively applied to a wider range of assessment tasks. This includes refining keyword selection strategies and exploring weight distribution techniques that adapt to different datasets. By addressing these areas, we believe our method has the potential to become an even more powerful tool for improving automated subjective assessment.

7. Declarations

7.1. Author Contributions

Conceptualization, K.S.M.A. and P.S.; methodology, K.S.M.A., B.B., and S.P.; validation, B.B. and S.P.; formal analysis, S.P., D.C., and S.N.L.; investigation, S.P. and S.N.L.; resources, B.B., P.S., and S.P.; writing—original draft preparation, B.B., D.C., and S.P.; writing—review and editing, P.S. and K.S.M.A.; visualization, S.N.L.; supervision, P.S. and D.C.; project administration, K.S.M.A. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: *https://www.kaggle.com/datasets/ananthu017/squad-csv-format.*

7.3. Funding

This research is supported by Multimedia University, Malaysia.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Liu, T., Hu, Y., Wang, B., Sun, Y., Gao, J., & Yin, B. (2023). Hierarchical Graph Convolutional Networks for Structured Long Document Classification. IEEE Transactions on Neural Networks and Learning Systems, 34(10), 8071–8085. doi:10.1109/TNNLS.2022.3185295.

[2] Ateeq, M. A., Tiun, S., Abdelhaq, H., & Rahhal, N. (2024). Arabic Narrative Question Answering (QA) Using Transformer Models. IEEE Access, 12, 2760 - 2777. doi:10.1109/ACCESS.2023.3348410.

[3] Paiva, J. C., Leal, J. P., & Figueira, Á. (2022). Automated Assessment in Computer Science Education: A State-of-the-Art Review. ACM Transactions on Computing Education, 22(3), 1–40. doi:10.1145/3513140.

[4] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: a systematic literature review. Artificial Intelligence Review, 55(3), 2495–2527. doi:10.1007/s10462-021-10068-2.

[5] Neshan, S. A. S., & Akbari, R. (2020). A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis. 2020 6th International Conference on Web Research, ICWR 2020, 8–14. doi:10.1109/ICWR49608.2020.9122298.

[6] Zhu, X., Wu, H., & Zhang, L. (2022). Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. IEEE Transactions on Learning Technologies, 15(3), 364–375. doi:10.1109/TLT.2022.3175537.

[7] Das, B., & Majumder, M. (2017). Factual open cloze question generation for assessment of learner's knowledge. International Journal of Educational Technology in Higher Education, 14(1), 1–12. doi:10.1186/s41239-017-0060-3.

[8] Sonai, K., Anbananthen, M., Mohamed, A., & Elyasir, H. (2013). Evolution of Opinion Mining. Australian Journal of Basic and Applied Sciences, 7(6), 359–370.

[9] Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). Automated Grammatical Error Detection for Language Learners, Second Edition. Synthesis Lectures on Human Language Technologies: Springer Nature, 7(1), 149-151. doi:10.2200/S00562ED1V01Y201401HLT025.

[10] Matsumori, S., Okuoka, K., Shibata, R., Inoue, M., Fukuchi, Y., & Imai, M. (2023). Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language Model. IEEE Access, 11, 9835–9850. doi:10.1109/ACCESS.2023.3239005.

[11] Feng, Y., Bagheri, E., Ensan, F., & Jovanovic, J. (2017). The state of the art in semantic relatedness: A framework for comparison. Knowledge Engineering Review, 32, 10. doi:10.1017/S0269888917000029.

[12] Sahu, A., & Bhowmick, P. K. (2020). Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. IEEE Transactions on Learning Technologies, 13(1), 77–90. doi:10.1109/TLT.2019.2897997.

[13] Rosnelly, R., Hartama, D., Sadikin, M., Lubis, C. P., Simanjuntak, M. S., & Kosasi, S. (2021). The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms. Turkish Journal of Computer and Mathematics Education, 12(3), 1415-1422. doi:10.17762/turcomat.v12i3.938.

[14] Wahyuningsih, T., Henderi, & Winarno. (2021). Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient. Journal of Applied Data Sciences, 2(2), 45–54. doi:10.47738/jads.v2i2.31.

[15] Anbananthen, K. S. M., Kannan, S., Busst, M. M. A., Muthaiyah, S., & Lurudusamy, S. N. (2022). Typographic Error Identification and Correction in Chatbot Using N-gram Overlapping Approach. Journal of System and Management Sciences, 12(5), 91–104. doi:10.33168/JSMS.2022.0506.

[16] Kaur, A., & Sasi Kumar, M. (2019). Performance Analysis of LSA for Descriptive Answer Assessment. Lecture Notes in Networks and Systems, 74, 57–63. doi:10.1007/978-981-13-7082-3_8.

[17] Mardini G, I. D., Quintero M, C. G., Viloria N, C. A., Percybrooks B, W. S., Robles N, H. S., & Villalba R, K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. Education and Information Technologies, 29(4), 4565–4590. doi:10.1007/s10639-023-11890-7.

[18] Bexte, M., Horbach, A., & Zesch, T. (2023). Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring?. In Findings of the association for computational linguistics: ACL 2023, 1892-1903.

[19] Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2023). Improving the performance of automatic short answer grading using transfer learning and augmentation. Engineering Applications of Artificial Intelligence, 123, 106292. doi:10.1016/j.engappai.2023.106292.

[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A Lite Bert for Self-Supervised Learning of Language Representations. 8th International Conference on Learning Representations, ICLR 2020, 344-350, Shenzhen, China. doi:10.1109/SLT48900.2021.9383575.

[21] Klyuchnikov, N., Trofimov, I., Artemova, E., Salnikov, M., Fedorov, M., Filippov, A., & Burnaev, E. (2022). NAS-Bench-NLP: Neural Architecture Search Benchmark for Natural Language Processing. IEEE Access, 10, 45736–45747. doi:10.1109/ACCESS.2022.3169897.

[22] Khodeir, N. A. (2021). Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT. IEEE Access, 9, 58243–58255. doi:10.1109/ACCESS.2021.3072734.

[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(1), 5485–5551.

[24] Alec, R., Jeffrey, W., Rewon, C., David, L., Dario, A., & Ilya, S. (2019). Language Models are Unsupervised Multitask Learners|Enhanced Reader. OpenAI Blog, 1(8), 9.

[25] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). OPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. 11th International Conference on Learning Representations, 1-16. doi:10.48550/arXiv.2210.17323.

[26] La Quatra, M., & Cagliero, L. (2023). BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization. Future Internet, 15(1), 15. doi:10.3390/fi15010015.

[27] Zhu, X., Wu, H., & Zhang, L. (2022). Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. IEEE Transactions on Learning Technologies, 15(3), 364–375. doi:10.1109/TLT.2022.3175537.

[28] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. ArXiv, 1-9. ArXiv:1310.4546. doi:10.48550/arXiv.1310.4546.

[29] Anbananthen, K. S. M., Krishnan, J. K., Sayeed, M. S., & Muniapan, P. (2017). Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text. American Journal of Applied Sciences, 14(9), 843–851. doi:10.3844/ajassp.2017.843.851.

[30] Žižovic, M., & Pamucar, D. (2019). New model for determining criteria weights: Level based weight assessment (LBWA) model. Decision Making: Applications in Management and Engineering, 2(2), 126–137. doi:10.31181/dmame1902102z.

[31] Bhole, G. P., & Deshmukh, T. (2018). Multi-criteria decision making (MCDM) methods and its applications. International Journal for Research in Applied Science & Engineering Technology, 6(5), 899-915.

[32] Odu, G. O. (2019). Weighting methods for multi-criteria decision-making technique. Journal of Applied Sciences and Environmental Management, 23(8), 1449. doi:10.4314/jasem.v23i8.7.

[33] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409. doi:10.1162/tacl_a_00373.

[34] Barbella, M., & Tortora, G. (2022). Rouge Metric Evaluation for Text Summarization Techniques. SSRN Electronic Journal, 1-31. doi:10.2139/ssrn.4120317.

[35] Saadany, H., & Orăsan, C. (2022). BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. Translation and Interpreting Technology Online, 48–56. doi:10.26615/978-954-452-071-7_006.

[36] Muludi, K., Fitria, K. M., Triloka, J., & Sutedi. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. International Journal of Advanced Computer Science and Applications, 15(3), 776–785. doi:10.14569/IJACSA.2024.0150379.

[37] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint, arXiv:1907.11692. doi:10.48550/arXiv.1907.11692.

[38] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32, 5754– 5764.

[39] Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2019). Evaluating question answering evaluation. MRQA@EMNLP 2019 - Proceedings of the 2nd Workshop on Machine Reading for Question Answering, 119–124. doi:10.18653/v1/d19-5817.