Leveraging Distance-Optimized Transformers for High-Performance Arabic Short Answers Grading

Hatem M. Noaman[1], Mohsen Rashwan[2], Hazem Raafat[3]*
Department of Computer Science-Faculty of Computer Science and Artificial Intelligence,
Beni-Suef University, Beni-Suef 62511, Egypt[1]
Department of Electronics and Electrical Communications Engineering-Faculty of Engineering,
Cairo University, Giza, 12613, Egypt[2]
Computer Science Department, Kuwait University, Kuwait[3]

Abstract—**This study presents comprehensive distance-optimized transformer architecture for Automated Arabic Short Answers Grading (AASAG) that systematically evaluates multiple semantic similarity measures. Short answer grading—assessment of responses typically 1-3 sentences long requiring conceptual understanding rather than factual recall—poses significant challenges in Arabic due to morphological complexity and limited computational resources. Our approach integrates pre-trained Arabic transformer models (AraBERT v02) with four distinct distance algorithms: cosine similarity, Manhattan distance, Euclidean distance, and dot-product calculations within a Siamese network architecture. Through systematic evaluation across three progressively enhanced datasets (original AR-ASAG, SemEval-augmented, and reference-integrated versions), our distance-optimized approach achieves state-of-the-art performance with correlation coefficients of 0.7998, representing a 5.5% improvement over existing methods. This advancement significantly outperforms traditional vector space models (0.7037 correlation), BERT-based approaches (0.7616), and hybrid semantic analysis methods (0.745), establishing new benchmarks for Arabic educational assessment technology.**

Keywords*—*Automatic Arabic Short Answers Grading; Arabic language processing; educational technology; pre-trained language models; semantic similarity

I. Introduction

A. Problem Definition and Scope

In the rapidly evolving domain of educational technology, the assessment of student responses, particularly short answers has emerged as both a compelling challenge and an area of rigorous research. While multiple choice questions offer ease of automated grading, they often fail to accurately assess a student's depth of understanding. In contrast, short answers questions provide valuable insights into students' cognitive processes, critical thinking abilities, and grasp of subject matter. Short answer questions, defined as assessment items requiring responses of one to three sentences that demonstrate conceptual understanding rather than mere factual recall. However, manually grading these responses is not only labor-intensive but also introduces elements of subjectivity and inconsistency. This reality has driven the development of automated grading systems that promise both efficiency and fairness. These systems utilize natural language processing, machine learning, and semantic analysis to evaluate the quality and correctness of student responses.

B. Arabic Language Challenges in Automated Grading

Despite significant advancements in English language assessment, Arabic language rich in nuances and with distinctive script characteristics remains relatively underexplored in this domain. The Arabic language, with its complex morphological structures and semantic intricacies, presents unique challenges for automated assessment systems. Despite serving millions of learners worldwide, Arabic has not received proportionate attention in educational technology research, creating a pressing need to advance Automatic Arabic Short Answers Grading (AASAG) capabilities. Transformer-based models, particularly sentence transformers, have revolutionized natural language processing due to their remarkable ability to capture semantic meaning [1]. These models generate dense vector representations of sentences that effectively encode contextual information and nuanced understanding capabilities essential for accurate short answer assessment. When integrated into grading systems, these vectors can be algorithmically compared to benchmark responses, providing reliable metrics of semantic similarity. This paper presents the development and evaluation of an Automatic Arabic Short Answers Grading model that leverages semantic similarity techniques. By capturing the semantic essence of Arabic language responses and measuring their similarity to model answers, our approach addresses a significant gap in the current literature.

Proposed model is validated using the AR-ASAG dataset [2], a comprehensive repository of Arabic student responses. Furthermore, we conduct an in-depth comparative analysis between our approach and existing models [3], [4], highlighting the effectiveness of integrating local and global weighting schemas, and outline promising future directions in this emerging field. In the subsequent sections, we shall elucidate related works, detail our proposed methodology, present experimental results, and conclude with discussions and implications of our findings.

C. Research Contributions and Objectives

This study addresses critical gaps in Arabic educational technology by developing a distance-optimized transformer architecture specifically designed for short answer grading. Our key contributions include: Comprehensive comparison of

*Corresponding authors. hnoaman@fcis.bsu.edu.eg; mrashwan@rdi-eg.com; mrashwan@rdi-eg.com four semantic similarity measures within transformer architectures for Arabic educational assessment, progressive augmentation of existing Arabic grading datasets with semantic textual similarity data and reference benchmarks, achievement of 0.7998 correlation coefficient, representing significant improvement over existing approaches and demonstrated effectiveness across five distinct question categories (definition, explanation, comparison, analysis, application)

This paper is structured as follows: Section II reviews existing literature on automated grading with emphasis on Arabic language challenges; Section III describes the three datasets used for evaluation; Section IV details our proposed distance-optimized transformer architecture; Section V presents experimental results; Section VI provides detailed discussion of findings; and Section VII concludes with implications and future directions.

II. Literature Review

The field of Automatic Short Answers Grading (ASAG) has gained significant progress in recent years, with researchers exploring various methodologies to enhance assessment accuracy and efficiency. This literature review examines key contributions to Automatic Arabic Short Answers Grading (AASAG), highlighting methodological approaches, datasets, evaluation metrics, and comparative performance.

A. Foundational Datasets and Vector Space Approaches

Ouahrani and Bennouar [2] made a significant contribution to the field by introducing AR-ASAG, a comprehensive Arabic dataset specifically designed for automatic short answers grading. Their dataset encompasses responses from three different exams across three student classes, with each test consisting of 16 short answers questions spanning five question types. For evaluation, they employed the COALS algorithm [5] to create a semantic space for answer representation. Their system achieved a Pearson Correlation of 0.7037 and RMSE of 1.0240, outperforming the SEMEVAL baseline by 11.75%. However, their approach fell slightly short compared to more advanced methodologies such as LIM-LIG's vectorized Word Embedding approach [3] and Huang and Su's topological approach [4].

B. Embedding-Based and Knowledge-Based Approaches

Meccawy et al. [6] expanded the empirical foundation by utilizing both the AR-ASAG dataset [2] and Rabbah & Al-Taani's dataset [7] to evaluate various linguistic processing techniques. Their research specifically examined the impact of stemming level (light stem versus base stem) on scoring precision for Arabic, a language characterized by extensive inflections. Their methodology incorporated modern natural language processing techniques including BERT [8], Word2vec [9], and knowledge-based similarity approaches utilizing Arabic WordNet [10]. Their comprehensive evaluation revealed impressive results, achieving a Pearson Correlation of 0.7758 for the AR-ASAG dataset and 0.8419 for the Rabbah & Al-Taani dataset, with corresponding RMSE values of 1.0439 and 1.0031, respectively. Notably, their findings demonstrated that light stemming produced more effective results than base stemming for Arabic text processing.

C. Semantic Analysis Techniques

Badry et al. [11] proposed an AASAG model leveraging semantic similarity approaches to measure the conceptual proximity between student responses and model answers. Using the AR-ASAG dataset [2], they demonstrated that a hybrid approach combining local and global weight-based Latent Semantic Analysis (LSA) significantly outperformed models using only local weight-based techniques. Their hybrid methodology achieved an F1-score of 82.82% and an RMSE of 0.798, establishing the effectiveness of multi-faceted semantic analysis for Arabic text assessment. In a related study, Abbas and Al-Qazaz [12] developed an Automatic Arabic Essay Scoring (AAES) system that integrated both Vector Space Model (VSM) and Latent Semantic Indexing (LSI). Their two-step methodology first extracted salient information from electronic essays using information retrieval techniques, then employed VSM and LSI to measure similarity between student essays and instructor-provided model answers.

Although limited to a single question with four model answers and 30 student responses, their work demonstrated the adaptability of information retrieval techniques for Arabic essay assessment.

D. Sentence Embedding and Deep Learning Approaches

El-Naka et al. [13] investigated sentence embedding techniques for evaluating short Arabic texts across multiple datasets, including AraScore, AR-ASAG, and two datasets with translated answers. Their comparative analysis indicated superior performance on the AraScore dataset, providing valuable insights into the contextual factors influencing the performance of automatic scoring systems in Arabic. Abdul Salam et al. [14] further advanced the field by implementing deep learning approaches for Arabic Short Answers Grading. Their innovative hybrid model, termed the LSTM-GWO model, combined Long Short-Term Memory (LSTM) networks [15] with the Grey Wolf Optimizer (GWO) [16]. When tested on science subject curricula data from schools in Egypt's Qalyubia-Governorate, their approach outperformed standalone models, including LSTM, Support Vector Machine (SVM), SVM-GWO, N-gram, Word2vec, Arabic WordNet, and MaLSTM across all evaluation metrics, demonstrating the significant potential of optimized deep learning architectures for Arabic text assessment.

E. Linguistic and Algorithm-Based Approaches

Jaber [17] introduced a distinctive approach that combines the Longest Common Subsequence (LCS) with Arabic WordNet (AWAN). This methodology first enriched student answers with synonyms using AWAN, then applied LCS to adjust the proximity between student and model answers. When tested on 330 student responses, this approach demonstrated remarkable accuracy with an RMSE of 0.81 and a Pearson correlation coefficient of 0.94, underscoring the potential of integrating linguistic resources with algorithmic matching techniques.

F. Educational Assessment and Feedback

Süzen et al. [18] examined text mining applications for educational assessment in the UK context, focusing on both automatic scoring and the provision of constructive feedback.

Their dual-focused approach aimed not only to evaluate response accuracy but also to offer insights that help students understand the rationale behind their scores, thereby facilitating deeper subject comprehension.

G. Comparative Analysis

The literature reveals a progressive evolution in AASAG methodologies, from basic vector space models to sophisticated deep learning architectures, as comprehensively summarized in Table I. Recent approaches incorporating transformer-based models, optimization techniques, and hybrid methodologies have demonstrated superior performance compared to earlier systems. The AR-ASAG dataset has emerged as a standard benchmark for evaluating Arabic Short Answers Grading Systems, allowing for meaningful comparisons across different approaches.

Performance metrics across studies indicate that embedding-based approaches, particularly when combined with optimization techniques or semantic analysis, consistently outperform traditional information retrieval methods. The linguistic complexity of Arabic presents unique challenges that researchers have addressed through specialized preprocessing techniques, with light stemming generally proving more effective than more aggressive morphological reduction.

III. Datasets

A. DS1: AR-ASAG Dataset

The AR-ASAG (Arabic Short Answers Grading) dataset is a significant contribution to Arabic language processing research as it represents the first publicly available dataset specifically designed for automatic grading of short answers in Arabic. The AR-ASAG dataset was created from a cybercrime teaching course with approximately 170 master's students who are native Arabic speakers.

The dataset includes five types of questions:

- Definition questions - requiring students to define cybercrime concepts.
- Explanation questions - asking students to explain processes or phenomena.
- Comparison questions - requesting students to compare different cybercrime concepts.
- Analysis questions - requiring deeper analysis of cybercrime scenarios.
- Application questions - asking students to apply concepts to specific situations.

Table II lists five different examples from AR-ASAG dataset with its English translations and students highest, medium and lowest grades answers.

B. DS2: Enhanced AR-ASAG with SemEval STS Integration

The AR-ASAG dataset has been significantly expanded beyond its original scope of short answers questions from authentic Arabic exams. This enhanced version incorporates sentence pairs from the SemEval Semantic Textual Similarity (STS) dataset [25], substantially broadening its utility for Arabic language processing tasks. By integrating the SemEval STS data, the enhanced AR-ASAG dataset now serves a dual purpose. It maintains its original value for automatic short answers grading while extending its applicability to semantic textual similarity assessment.

C. DS3: Reference Answer Integration

The enhanced AR-ASAG dataset has been further augmented with a strategic reference answer component, adding significant value to its existing foundation of original short-answer questions and SemEval STS data. This critical enhancement introduces a reference answer subset designed to establish clear quality benchmarks within the assessment framework. Within this new subset, answers that received full marks are specifically highlighted and serve as exemplary benchmark responses. These high quality reference answers create a sophisticated evaluation framework against which other student responses with similar original marks can be systematically compared and analyzed. This comparative structure enables more nuanced assessment of answer quality variations among similarly-graded responses. By integrating these top performing student answers as reference benchmarks, the dataset offers researchers and educators a unique, multidimensional perspective on answer quality assessment and grading standards. This approach allows for more precise calibration of automated scoring systems by providing concrete examples of ideal responses rather than relying solely on abstract scoring criteria. Table III outlines the characteristics and composition of three distinct datasets utilized for training Arabic Automated Short Answers Scoring (ASAS) systems.

IV. Proposed Model

Our model addresses the Arabic Short Answers Grading (ASAG) problem by leveraging Sentence Transformers for semantic similarity assessment. The proposed architecture effectively compares student answers with model answers through a two-phase approach.

A. Model Architecture

This research train Sentence Transformers using Siamese or triplet network structures, optimizing the model to minimize distance between semantically similar sentences while maximizing distance between dissimilar ones. This approach yields fixed-size embeddings that capture the semantic essence of text inputs with reduced computational overhead compared to traditional transformer models.

As illustrated in Fig. 1, our model processes both the question model answer and student answer through identical pathways:

1) Input processing: Each answer is represented as a sequence of tokens, with model answer tokens $w_1, w_2, \ldots, w_n$ and student answer tokens $w'_1, w'_2, \ldots, w'_m$. Each sequence is fed into separate instances of the same Pretrained Transformer.

TABLE I. Evolution of Reference-Dependent Arabic Automatic Short Answers Grading Methods

Dataset(s)	Key Methodology	Performance Metrics	Contributions	Ref
Information Retrieval Approaches				
Custom (1 question, 4 model answers, 30 responses)	Vector Space Model (VSM) [19], [20], [21] + Latent Semantic Indexing (LSI) [22]	N/A	Two-step process: information extraction followed by similarity measurement	[12]
AR-ASAG (original creation)	COALS algorithm for semantic space	Pearson: 0.7037 RMSE: 1.0240	Created the AR-ASAG dataset with 16 questions across 5 question types; outperformed SEMEVAL by 11.75%	[2]
UK educational context	Text mining for scoring and feedback	N/A	Dual focus on assessment accuracy and constructive feedback provision	[18]
Linguistic Resource Integration Approaches				
Custom (330 student responses)	Longest Common Subsequence (LCS) [24] + Arabic WordNet (AWAN) [23], [10]	Pearson: 0.94 RMSE: 0.81	Combined linguistic synonyms with algorithmic matching	[17]
Neural Networks & Embedding Models Approaches				
AraScore AR-ASAG Two translated datasets	Sentence embedding approach	Best performance on AraScore dataset (metrics not specified)	Comparative analysis across multiple datasets revealed context-dependent performance factors	[13]
Custom (science curricula from Egyptian schools)	Hybrid LSTM-GWO (Long Short-Term Memory [15] + Grey Wolf Optimizer [16])	Outperformed all baseline models (SVM, LSTM, Word2vec, etc.)	Demonstrated superiority of optimized deep learning over traditional approaches for Arabic assessment	[14]
Advanced Hybrid Approaches				
AR-ASAG [2] Rabbah & Al-Taani [7]	BERT [8], Word2vec [9], Arabic WordNet [10] with light/base stemming	AR-ASAG Pearson: 0.7758 RMSE: 1.0439 Rabbah & Al-Taani Pearson: 0.8419 RMSE: 1.0031	Demonstrated light stemming's superiority over base stemming for Arabic; integrated transformer models with linguistic processing	[6]
AR-ASAG [2]	Hybrid local and global weight-based Latent Semantic Analysis (LSA)	F1-score: 82.82% RMSE: 0.798	Proved hybrid LSA approach outperforms local weight-based techniques alone; combined traditional IR with advanced weighting schemes	[11]

TABLE II. AR-ASAG Dataset Question Types with Examples

Question Type	Arabic Question	English Translation	High Grade	Medium Grade	Low Grade
Definition	ما هو الاختراق الإلكتروني؟	What is electronic hacking?	هو عملية الدخول غير المشروع إلى أنظمة الحاسوب	هو الدخول إلى أجهزة الكمبيوتر بدون إذن	هو عندما يخترق شخص_ما جهاز شخص آخر
Explanation	اشرح كيفية عمل هجمات حجب الخدمة؟	Explain how Denial of Service attacks work?	هجمات تهدف إلى منع المستخدمين الشرعيين من الوصول إلى الخدمات	هي إرسال طلبات كثيرة للسيرفر حتى يتوقف عن العمل	هي عندما يتم إيقاف موقع على الإنترنت
Comparison	قارن بين البرمجيات الخبيثة والبرمجيات الفدية	Compare between malware and ransomware	البرمجيات الخبيثة تشمل جميع أنواع البرامج الضارة	البرمجيات الخبيثة تضر الجهاز بينما البرمجيات الفدية تطلب المال	البرمجيات الخبيثة سيئة والفدية أسوأ
Analysis	حلل مخاطر استخدام شبكات الواي فاي العامة	Analyze the risks of using public Wi-Fi networks	المخاطر تشمل إمكانية اعتراض البيانات المرسلة والمستقبلة	يمكن للمخترقين رؤية بياناتك على الواي فاي العام	الواي فاي العام غير آمن ويجب تجنبه
Application	كيف يمكن لمؤسسة تعليمية حماية بيانات الطلاب من الاختراق؟	How can an educational institution protect student data from hacking?	يجب على المؤسسة تشفير قواعد البيانات واستخدام أنظمة تحكم بالوصول قوية	يجب استخدام كلمات مرور قوية وتثبيت برامج مكافحة الفيروسات	يجب عليهم استخدام برامج حماية جيدة

TABLE III. Comparative Analysis of Arabic Language Assessment Datasets

	DS1: Original AR-ASAG	DS2: SemEval STS Integration	DS3: Reference Answer Subset
Primary Purpose	Short answers grading from authentic Arabic exams	Semantic textual similarity assessment	Benchmark quality standards for comparative evaluation
Source	Cybercrime course exam responses from 170 master's students	AR-ASAG + Arabic sentence pairs from SemEval dataset [25]	Top-scoring student answers from original AR-ASAG
Question Types	Definition, explanation, comparison, analysis, application	Varied sentence relationships	Primarily focuses on exemplary answers across all question types
Key Contribution	First Arabic Short Answers Grading dataset	Broadens scope to include semantic similarity	Establishes concrete quality benchmarks within context

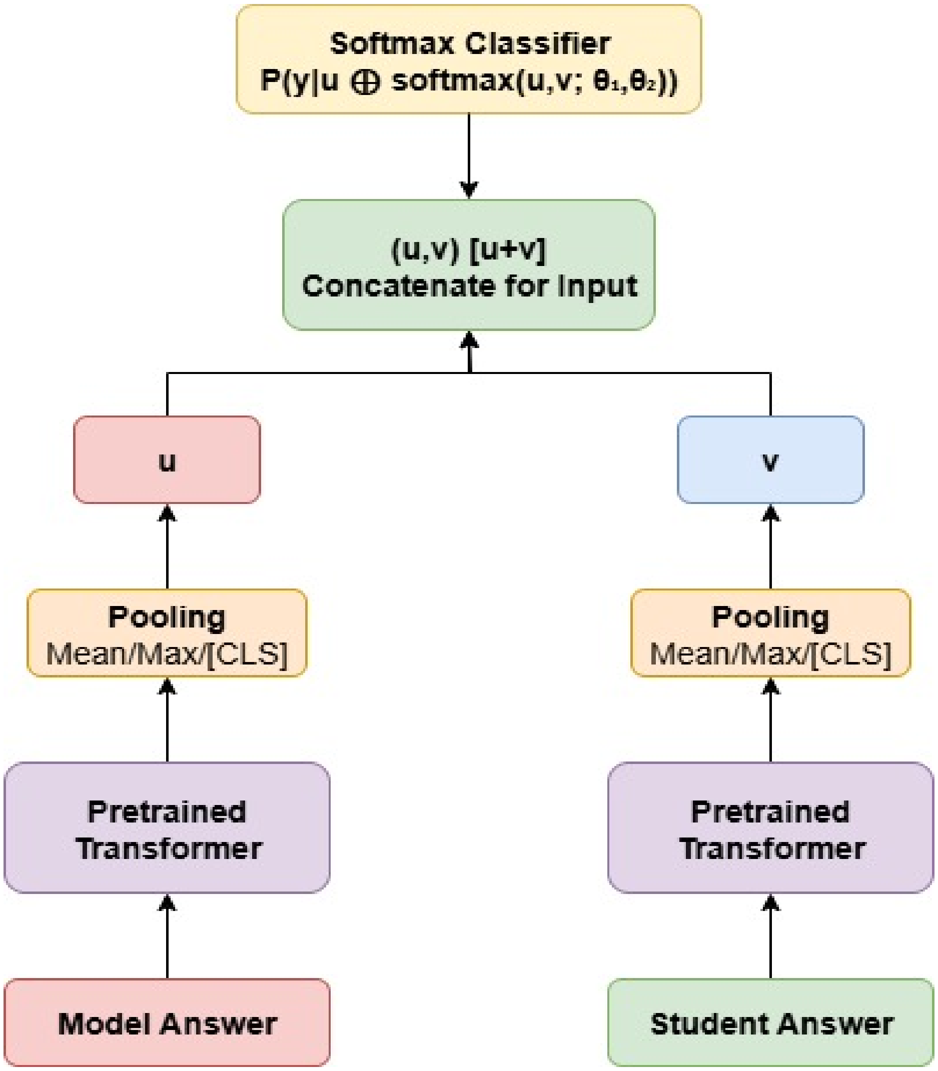

Fig. 1. Sentence transformer-based model for Arabic Short Answers Grading.

2) Contextualized embeddings: The transformer generates rich, contextualized representations for each token according to:

$$h_i = \text{Transformer}(w_i) \tag{1}$$

where $H \in \mathbb{R}^{n \times d}$ for model answers and $H' \in \mathbb{R}^{m \times d}$ for student answers.

3) Pooling layer: A pooling operation summarizes the sequence embeddings into fixed-size vectors: u for the model answer and v for the student answer.

$$u = \text{Pooling}(H) \in \mathbb{R}^d \tag{2}$$

$$v = \text{Pooling}(H') \in \mathbb{R}^d \tag{3}$$

4) Vector combination: The model combines vectors u and v along with their absolute differences:

$$c = [u; v; |u - v|] \in \mathbb{R}^{3d} \tag{4}$$

5) Classification: A softmax classifier processes these combined embeddings to determine the semantic similarity:

$$P(y|u, v) = \text{softmax}(W \cdot c + b) \tag{5}$$

where, y represents the similarity score or classification.

Fig. 2 visualizes the process to develop a specialized Sentence Transformer model for Arabic Short Answers Grading (ASAG). The development workflow consists of several interconnected stages designed to optimize embedding generation for Arabic shorts assessment.

The process begins with loading the ASAG dataset, which serves as the foundation for both training and evaluation. This dataset is split into training and testing portions through a

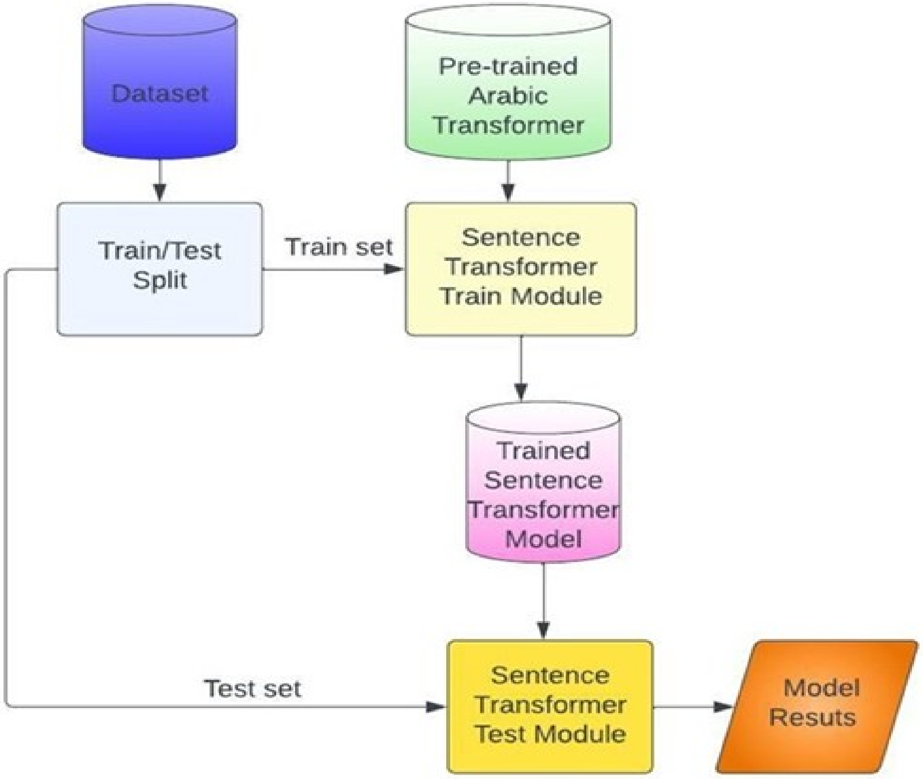
Dataset
Pre-trained Arabic Transformer
Train/Test Split
Train set
Sentence Transformer Train Module
Trained Sentence Transformer Model
Test set
Sentence Transformer Test Module
Model Resuts

Fig. 2. The proposed model train/test flowchart.

standard train/test partition. Simultaneously, we integrate a pre-trained Arabic Transformer with the Sentence Transformer architecture. This integration ensures the model leverages the linguistic knowledge captured by the pre-trained transformer while optimizing for sentence-level embeddings. The Sentence Transformer is trained using the training portion of the ASAG dataset. This fine-tuning process adapts the model to capture the semantic nuances specific to Arabic short answers in academic contexts. After training, the resulting "ASAG Sentence Transformer Model" becomes a specialized tool for generating embeddings optimized for grading tasks. Performance evaluation uses the held-out test portion of the dataset, applying the trained model to generate embeddings and compute similarity measures. These embeddings capture rich semantic information, allowing for effective comparison between model answers and student responses.

Our approach employs four distinct similarity metrics to quantify the distance between embedding vectors:

$$d_{\text{euclidean}}(u, v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{6}$$

$$d_{\text{cosine}}(u, v) = \frac{u \cdot v}{||u||_2 ||v||_2} \tag{7}$$

$$d_{\text{manhattan}}(u, v) = \sum_{i=1}^{n} |u_i - v_i| \tag{8}$$

$$d_{\text{dot\_product}}(u, v) = \sum_{i=1}^{n} u_i . v_i \tag{9}$$

V. Experiments and Results

A. Model Performance Comparison

The experimental results demonstrate that model architecture and size substantially impact performance across all datasets. As shown in Fig. 4, the 'aubmindlab-bert-large-arabertv02' model consistently achieves superior performance on DS2, with exceptional scores (cosine: 0.7985, manhattan: 0.7931) that represent the highest values in our evaluation. Similarly, Fig. 5 demonstrates this model's continued dominance on DS3 (cosine: 0.7955, manhattan: 0.7937). Interestingly, for DS1, Fig. 3 illustrates that the 'asafaya-bert-large-arabic' model outperforms all others, achieving the highest scores across all metrics (cosine: 0.7755, manhattan: 0.7701).

We observe a clear dataset difficulty gradient, with DS1 presenting more significant challenges for all models compared to DS2 and DS3. This pattern is evident when comparing Fig. 3, 4, and 5, where the overall height of the bars increases progressively from DS1 to DS3. This suggests inherent differences in linguistic patterns or semantic relationships across these datasets that affect model performance.

Regarding similarity metrics, cosine similarity consistently produces the most favorable results across all models and datasets, as visible in all three figures. This is followed closely by manhattan and euclidean distances. The dot product metric (represented by orange bars in Fig. 3 to 5) consistently yields the lowest performance scores, particularly struggling with the UBC-NLP-AraT5v2-base model on DS1 (0.5395), as evident in Fig. 3.

The correlation coefficient choice also impacts results, with Spearman generally yielding higher values than Pearson, particularly for DS3, where the best model reaches 0.7998 for manhattan distance using Spearman correlation. While our figures display Pearson correlation results, comparative analysis between correlation types reveals this important distinction. These findings provide valuable insights for selecting appropriate Arabic language models and evaluation metrics for specific applications, highlighting the critical importance of considering both model architecture and evaluation dataset characteristics when assessing Arabic NLP model performance.

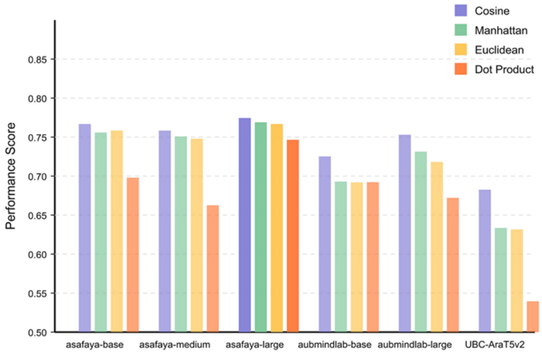

Fig. 3. DS1 performance scores for cosine, Manhattan, Euclidean and dot-product distance measurements.

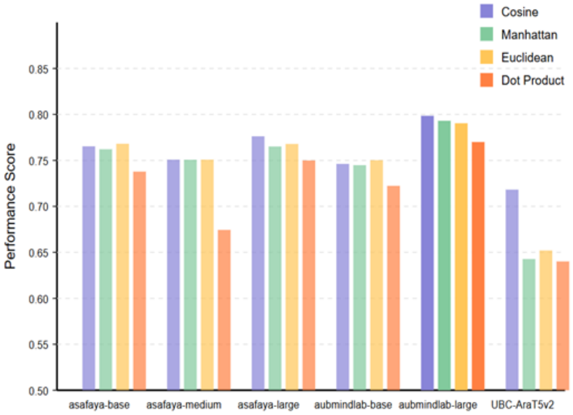

Fig. 4. DS2 performance scores for cosine, Manhattan, Euclidean and dot-product distance measurements.

B. Benchmarking with Previous Results

To assess the performance of our model in relation to established benchmarks, we conducted a comprehensive comparison using Pearson correlation coefficients. Table IV presents the comparative analysis between our best-performing model and previous benchmark results in the field of Automatic Short Answer Grading (ASAG).

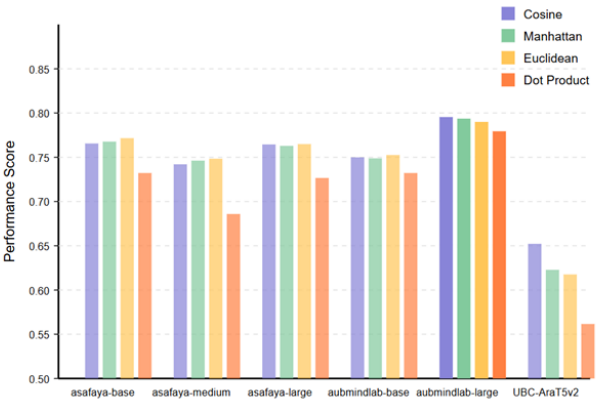

Fig. 5. DS3 performance scores for cosine, Manhattan, Euclidean and dot-product distance measurements.

Our proposed system demonstrates superior performance across all distance measures, with the arabertv02 + Cosine distance configuration achieving the highest Pearson coefficient of 0.7998. This outperforms all previous methodologies, including the AASAG models (Stemming model: 0.723, local weight: 0.731, and hybrid local global weight: 0.745) and exceeds the performance of Word2vec+Cosine + Light Stem (0.7758) from the Automatic Essay Scoring approach. The consistent high performance across different distance measures (Manhattan: 0.7937, Euclidean: 0.7904, and Dot-Product: 0.7794) further validates the robustness of our arabertv02-based approach.

The significant performance improvement (approximately 5.5% increase over the best previous method) can be attributed to several key factors:

1) Pre-training advantages: The arabertv02 model was pre-trained specifically on Arabic text, allowing it to capture nuanced semantic relationships particular to the language.

2) Contextual embeddings: Unlike static embeddings used in previous approaches, our model generates contextual embeddings that better capture the meaning of words based on their surrounding context.

3) Distance measure optimization: Our systematic evaluation of multiple distance measures allowed us to identify the optimal configuration for this specific task.

4) Large-scale transformer architecture: The arabertv02 architecture provides a more sophisticated representation of language compared to traditional methods, particularly beneficial for understanding complex answer structures.

These results highlight the effectiveness of our methodology and demonstrate its potential as a state-of-the-art solution for ASAG applications, particularly in Arabic language contexts.

VI. Discussion

Our proposed system demonstrates substantial improvements over existing Arabic grading approaches, achieving a 5.5% increase in correlation coefficients compared to previous best methods. This improvement stems from several key

TABLE IV. Comparison and Evaluation of Study Results Against Previous Models

Approach	Methodology	Pearson	Year
Proposed System	arabertv02 + Cosine	**0.7998**	2025
	arabertv02 + Manhattan	0.7937	
	arabertv02 + Euclidean	0.7904	
	arabertv02 + Dot-Product	0.7794	
AASAG	Stemming model	0.723	2023 [11]
	Local weight	0.731	
	Hybrid local & global	0.745	
Automatic Essay Scoring	WordNet+Cosine + Base	0.7469	2023 [6]
	WordNet+Cosine + Light	0.7553	
	Word2vec+Cosine + Base	0.7693	
	Word2vec+Cosine + Light	0.7758	
	BERT+Cosine + Base	0.7536	
	BERT+Cosine + Light	0.7616	
ASAG-Combined	Root Stem	0.7010	2020 [2]
	Light Stem	0.7037	
ASAG W-SM System	Root Stem	0.6830	2020 [2]
	Light Stem	0.6818	
ASAG-Basic System	Root Stem	0.6550	2020 [2]
	Light Stem	0.6340	

factors: AraBERT v02's pre-training on extensive Arabic corpora enables capture of nuanced semantic relationships specific to Arabic language patterns, providing significant advantages over generic multilingual models or traditional static embeddings.Unlike static representations used in previous approaches, our transformer-based system generates contextual embeddings that adapt meaning based on surrounding context, crucial for educational assessment where context determines answer correctness. Systematic evaluation of multiple distance measures reveals that cosine similarity provides optimal performance for Arabic semantic assessment, outperforming traditional Euclidean distance approaches commonly used in earlier systems. Progressive dataset enhancement through SemEval integration and reference answer inclusion provides more robust training signals, enabling better generalization across diverse question types and response patterns.

VII. Conclusion and Future Work

This study represents a significant advancement in the field of Automatic Short Answer Grading (ASAG) for Arabic language assessment. By leveraging the arabertv02 pre-trained language model with various distance measures, we have developed a system that substantially outperforms previous approaches. The experimental results demonstrate that our proposed methodology achieves state-of-the-art performance with a Pearson correlation coefficient of 0.7998 using the Cosine distance measure, representing a meaningful improvement over existing solutions. Our comprehensive evaluation has shown that transformer-based architectures specifically trained on Arabic text provide superior semantic understanding compared to traditional NLP approaches. The consistent performance across different distance metrics further validates the robustness of our approach. The integration of contextual word embeddings from arabertv02 has proven particularly effective at capturing the nuanced semantic relationships in Arabic short answers, enabling more accurate assessment that better aligns with human grading. This research contributes to the growing body of evidence supporting the efficacy of transformer-based models in educational technology applications, particularly for languages beyond English. Our findings have important implications for the development of automated assessment tools that can reduce the workload of educators while maintaining high standards of evaluation accuracy.

Future work will focus on expanding the evaluation to additional domains beyond cybersecurity, investigating cross-domain generalization capabilities, and exploring the integration of automated feedback generation systems. Additionally, we plan to conduct longitudinal studies on real classroom deployments to assess the practical impact of our approach in educational settings.

Data Availability

The AR-ASAG dataset used in this paper is available at: `https://data.mendeley.com/datasets/dj95jh332j/1`

References

[1] Reimers, Nils and Gurevych, Iryna, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.

[2] Ouahrani, Leila and Bennouar, Djamal, "AR-ASAG an Arabic dataset for automatic short answer grading evaluation," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

[3] El Moatez Billah Nagoudi and Ferrero, Jérémy and Schwab, D. S., "LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 134–138, 2017, Association for Computational Linguistics.

[4] Wu, H. and Huang, H. and Jian, P. and Guo, Y. and Su, C., "BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, vol. L, pp. 77–84, 2017, Association for Computational Linguistics.

[5] Rohde, Douglas L. T. and Gonnerman, Laura M. and Plaut, David C., "An Improved Method for Deriving Word Meaning from Lexical," *Cognitive Psychology*, vol. 7, pp. 573–605, 2004.

[6] Meccawy, Maram and others, “Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[7] Rababah, H. and Al-Taani, A. T., “An automated scoring approach for Arabic short answers essay questions,” in *Proceedings of 8th International Conference on Information Technology (ICIT)*, pp. 697–702, 2017, Amman, Jordan, doi: 10.1109/ICITECH.2017.8079930.

[8] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[9] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey, “Efficient Estimation of Word Representations in Vector Space,” *arXiv preprint arXiv:1301.3781*, 2013.

[10] Shehab, A. and Faroun, M. and Rashad, M., “An Automatic Arabic Essay Grading System based on Text Similarity Algorithms,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, pp. 263–268, 2018, doi: 10.14569/IJACSA.2018.090337.

[11] Badry, Rasha M. and others, “Automatic Arabic Grading System for Short Answer Questions,” *IEEE Access*, 2023.

[12] Abbas, Ayad R. and Al-qazaz, Ahmed S., “Automated Arabic essay scoring (aaes) using vectors space model (VSM) and latent semantics indexing (LSI),” *Engineering and Technology Journal*, vol. 33, no. 3, pp. 410–426, 2015.

[13] ElNaka, Abdelrahman and others, “AraScore: Investigating Response-Based Arabic Short Answer Scoring,” *Procedia Computer Science*, vol. 189, pp. 282–291, 2021.

[14] Abdul Salam, Mustafa and Abd El-Fatah, Mohamed and Hassan, Naglaa Fathy, “Automatic grading for Arabic short answer questions using optimized deep learning model,” *Plos One*, vol. 17, no. 8, pp. e0272269, 2022.

[15] Hochreiter, Sepp and Schmidhuber, Jürgen, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Mirjalili, Seyedali, “Grey wolf optimizer,” *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[17] Abdeljaber, Hikmat A., “Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet,” *IEEE Access*, vol. 9, pp. 76433–76445, 2021.

[18] Süzen, Neslihan and others, “Automatic short answer grading and feedback using text mining methods,” *Procedia Computer Science*, vol. 169, pp. 726–743, 2020.

[19] Salton, G. and Wong, A. and Yang, C.-S., “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[20] Robertson, S. E. and Jones, K. Sparck, “Relevance weighting of search terms,” *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129–146, 1976.

[21] Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier, *Modern Information Retrieval*, ACM Press, New York, NY, USA, 1999.

[22] Deerwester, Scott and Dumais, Susan T. and Furnas, George W. and Landauer, Thomas K. and Harshman, Richard, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[23] Elkateb, Sabri and Black, William and Rodríguez, Horacio and Alkhalifa, Musa and Vossen, Piek and Pease, Adam and Fellbaum, Christiane, “Building a wordnet for Arabic,” in *Proceedings of The fifth international conference on Language Resources and Evaluation*, 2006.

[24] Hunt, James W. and McIlroy, M. Douglas, “An algorithm for differential file comparison,” *Computing Science Technical Report*, Bell Laboratories, 1976.

[25] Cer, Daniel and others, “Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation,” *arXiv preprint arXiv:1708.00055*, 2017.