Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques

Maram Meccawy, Afnan Ali Bayazed, Bashayer Al-Abdullah, Hind Algamdi
Information Systems Department, Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Automated Essay Scoring (AES) systems involve using a specially designed computing program to mark students' essays. It is a form of online assessment supported by natural language processing (NLP). These systems seek to exploit advanced technologies to reduce the time and effort spent on the exam scoring process. These systems have been applied in several languages, including Arabic. Nevertheless, the applicable NLP techniques in Arabic AES are still limited, and further investigation is needed to make NLP suitable for Arabic to achieve human-like scoring accuracy. Therefore, this comparative empirical experimental study tested two word-embedding deep learning approaches, namely BERT and Word2vec, along with a knowledge-based similarity approach; Arabic WordNet. The study used the Cosine similarity measure to provide optimal student answer scores. Several experiments were conducted for each of the proposed approaches on two available Arabic short answer question datasets to explore the effect of the stemming level. The quantitative results of this study indicated that advanced models of contextual embedding can improve the efficiency of Arabic AES as the meaning of words can differ in the different contexts. Therefore, serve as a catalyst for future research based on contextual embedding models, as the BERT approach achieved the best Pearson Correlation (.84) and RMSE (1.003). However, this research area needs further investigation to increase the accuracy of Arabic AES to become a practical online scoring system.**

Keywords*—Arabic language; Automated Essay Scoring (AES); Automated Scoring (AS); Educational Technologies; NLP

I. INTRODUCTION

Online learning has become an integral part of the educational system in the wake of the COVID-19 pandemic, during which most countries had to close their educational institutions as a precautionary measure to preserve the safety of the public from the spread of infection. Shifting to online education was an alternative solution to cope with the restrictions imposed by the lockdown of educational institutions; however, it imposes many social and educational challenges [1], particularly when it comes to online assessment. A study which looked into assessment during the COVID-19 lockdown has suggested the need for a multilevel approach to the problems of cheating and plagiarism [2]. Even prior to the pandemic, assessment was a well-known challenge in education encountered by both traditional and online education [3]. It continues to be a dominant issue in the online learning arena even in the post-pandemic world.

Assessment in education describes the "processes of evaluating the effectiveness of sequences of instructional activities when the sequence was completed" [4], and it has been divided into formative and summative assessments [5]. Formative assessment is part of the instructional process in the classroom; it provides the feedback needed to adjust the teaching and learning activities to suit the learners while they are engaged. On the other hand, summative assessment is given periodically in order to assess the learners' level of knowledge or achievement at a certain point in time. The AES in the context of this work is considered as a form of summative assessment.

Despite the diversity of methods for assessing students' progress, the examination method has been used predominately to measure students' performance and knowledge. Namely, examinations are held at the end of the course in addition to course assignments [6]. The academic examination is a considerable undertaking in the education process due to the significant number of students who take the exams. A massive overhead of time and effort is involved, with teachers having to score exams instead of focusing on other important aspects of the educational process [7].

At this point, automated scoring (AS) systems appear to be one of the best solutions to overcome these challenges. AS systems offer a collection of different grading approaches based on measuring the similarity between the answer posed and the expected answer [8]. AS systems introduce an effective alternative scoring mechanism for several types of questions such as true/false (T/F) questions, multiple choice questions (MCQs), and fill-in the blank questions. Nevertheless, the grading of essay questions and short answer questions is a complex task in AS systems, as such systems need deep knowledge and understanding of the nature of texts in a language process [9]. Hence, automated essay scoring (AES) [10] has emerged as a way to grade student essays.

There are different approaches used in the context of AES to measure the similarity between model answers (MAs) and student answers (SAs). One approach is string-based similarity [11], which involves scaling the string's sequence and the composition of the letter; in comparison, corpus-based similarity [12] measures similarity depending on the information obtained from large copra. Moreover, the knowledge-based similarity approach [13], which is one of the most popular approaches for measuring text similarity, utilizes information derived from the semantic network Arabic WordNet.

According to research presented in [14], many research efforts have focused on developing and studying automated

scoring for short and essay questions written in English. In comparison, a limited number of studies in this area have been conducted to address automated Arabic short answer questions. The Arabic language, spoken by approximately 400 million people [15], is a complex language, with many synonyms for one word, differences in the meaning of a word according to its different formation, and richness of morphology. Accordingly, there is a lack of a practical system in Arabic to conduct automated scoring of short or essay questions due to the accuracy of the proposed frameworks being insufficient. Some studies have proposed a framework that translates the Arabic answer into English to measure the similarity between the model answer and student answer [16]. In contrast, other studies have proposed solutions that are centred around processing the Arabic language. For example, the work presented in [17] showed that using synonyms and finding the root of words can close the difference between a model answer and student answer. Another study has suggested that using deep learning can enhance the accuracy of Arabic AES [14].

The focus of this study was automated short text answer scoring presented in Arabic using a text mining technique with deep learning algorithms for natural language processing (NLP). The study aimed to investigate the effects of stemming level on measuring similarity between student answer (SAs) and model answer (MAs) It contributes the following to the following to this research area:

- Provide a comparative empirical study by comparing different word-embedding approaches to examine the word's surrounding context.
- Investigate mechanisms that depend on raising the percentage of similarity between the SA and MA by increasing the number of correct words in the SA.
- Evaluate the proposed models in the literature using two different available Arabic corpuses.

The rest of this paper is organized as follows. Section II presents the related work, while Section III explains the methodology used in this study. Section IV presents and discusses the study results. Finally, the conclusion in Section Section V and recommendations for future work are given in Section VI.

II. Related Works

This section presents works related to the processing of Arabic short answer scoring and the present state-of-the-art approaches for short answer scoring for the English and Arabic languages.

Previous students have introduced models for Arabic AES that use string-based techniques of text similarity; for example, research in [18] presents a system for online exams in Arabic that is based on the Stemming and Levenshtein algorithms. The system reduces the words that have the same stem to a common phrasing. The results of this study showed that the proposed system is effective as a classification tool for Arabic essay questions.

Several other studies dealing with AS in Arabic have employed corpus-based algorithms such as Latent Semantic Analysis (LSA), which is an NLP technique that evaluates the similarity between two documents. This method relies on generating vectors-presentation for semantic terms, words, or even the concepts [19]. Research in [20] presented a system for scoring Arabic short answers by embedding LSA with the main three important syntactic features: lemmatization, the mistake of words, and the number of common words. They employed bag-of-words (BOW) to present feature vectors that mapped into the Cosine algorithms to measure the similarity between student answers (SAs) and model answers (MAs). To evaluate their approach, an Arabic short answer corpus was generated, and the best result was 96.72%. Similarly, research in [21] applied a similar approach, though their approach was centred on a semantic perspective as it combined LSA with linguistic features. After performing the normalization process, the authors generated feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF) as the input to the Cosine algorithm. To improve the accuracy of LSA, the study leveraged part-of-speech (POS) with Term Frequency (TF) to take into account the syntactic of words. Moreover, the work in [22] utilized LSA with pre-processing of answers, for example, by removing stop-words, applying the replacing synonyms technique, and applying a stemming process. Meanwhile, the study in [23] proposed a new automated essay scoring approach focused primarily on measuring the similarity based on the root extraction and the synonyms of the keywords in addition to using the Cosine similarity. Moreover, they used the ROUGE metric to evaluate the obtained results, which gave a high accuracy rate of 84.5%, which indicated that the model's scoring could approximate human scoring.

Furthermore, another related work rendered an automated grading model for Arabic essays with the aim of gaining and achieving better efficiency [17]. In this case, features were extracted from the SA and MA by utilizing the F_score tools, and Arabic WordNet was used as a helpful method for semantic similarity, which is considered as a knowledge-based algorithm. The proposed model recorded better accuracy when Arabic WordNet was used than without it.

Moreover, other research efforts presented frameworks using hybrid approaches of a string-based corpus and knowledge-based corpus. Research presented by [24] compared different algorithms to inspect the optimal solution of Arabic automatic grading. They employed two string-based text similarities methods: the Damerau-Levenshtein algorithm and the N-gram algorithm. Further, the LSA and DISCO were used as corpus-based text similarity algorithms. The authors applied four testing methods, namely Stop, Raw, Stop-Stem, and Stem, to investigate the accuracy of string-based algorithms. However, they only used the stop method to test corpus-based algorithms, since the semantic similarity between the stop words does not need to be measured. In addition, they calculated the correlation constant between the manual grading and the automatic system grading. Thus, the results showed that the N-gram with stop method resulted in 0.820 as the best correlation. Generally, the character-based N-gram algorithm achieved better results than the other type.

On the other hand, for the corpus-based algorithms, the DISCO algorithm achieved a higher correlation than the LSA algorithm, since it is built on words that have a common distribution.

Similarly, research work in [8] introduced a comparative study by investigating 14 string-based algorithms and two corpus-based algorithms. These algorithms were evaluated across two main models. The first was the holistic model, which compares the full form of an SA to an MA without splitting the SA and ignoring the MA's partition scheme. The second was the partitioning model, which divides the answer of the student into a group of sentences based on the sentence boundary detection templates and then maps each sentence to the most similar element on the MA. The r correlation and the RMSE were used to measure the correlation, while the MaxSim and the AvgSim were used to calculate the similarity. Accordingly, the experiment showed that corpus-based algorithms produce lower error rate values, and the N-gram (Bi-gram, Tri-gram) approach recorded the best r correlation of 0.826, with the partitioning model achieving results better than the holistic model in all cases.

Additionally, the same authors have also addressed this issue by translating Arabic text into English language [16]. They developed a framework that evaluates the similarity between SAs and MAs by fundamentally translating Arabic answers into the English language. The reason behind this choice was the limitation of available Arabic text processing resources and tools. Their proposed system was composed of five main stages as they applied different measuring techniques of text similarity in a separate and hybrid way; thus, the obtained scoring was scaled using the K-mean cluster.

On the other hand, work in [25] used supervised machine learning and classification algorithms to produce a new Arabic essay grading database that is compatible with machine learning. The study depended on leveraging machine learning algorithms to evaluate the database. Thus, the study used the several classifiers to build the training model of the database such as NB, Meta-classifier, and decision tree (J48). The results obtained from the third experiment using Meta-classifier showed a higher accuracy rate of 83%. Likewise, [26] provided a system in the context of web-based learning focusing on the Vector Space Model (VSM) and Latent Semantic Indexing (LSI). The system firstly extracted significant information from the essays by applying the VSM for information retrieval techniques. Then it applied the VSM and LSI to determine the degree of similarity between the student essay and the model essay after each essay has been converted to vector space. Finally, it used the Cosine similarity to measure the score of SA. The results showed that the proposed system provided scoring accuracy close to the traditional scoring by the professor.

Several state-of-the-art deep learning approaches have been used to process text similarity and conduct automatic scoring (AS). These approaches mainly rely on automated multi-layered feature-distributed representation and learning. Embedding models have emerged based on deep learning methods: word-embedding and paragraph-embedding have become cutting-edge models in the NLP field [27]. Several contributions have employed embedding models to address AS in the English language [28-30]. Moreover, the study in [31] provided two main approaches for grading short answer questions. In the first approach, the study used four different methods based on word-embedding models: Word2vec, GloVe, and Fasttext3 by summation pre-trained word vectors. In the second approach, it used trained three deep learning models to extract the paragraph vector. Finally, Cosine similarity was used to measure the similarity between the vector of the MAs and the vector of the SAs. The best value the study produced for RMSE was 0.797.

In addition, in its comparative empirical work, research provided in [32] developed a model for AS by configuration of three presentation feature vectors: a manually extracted feature, word2vec, and a contextual embedding feature using the BERTmodel. The best-recorded accuracy was by the configuration of three feature vectors of 75.2 ± 1.0 Quantized Classification. Table I summarises the approaches to handling automated short answer scoring as introduced in this section.

TABLE I. SUMMARY OF APPLIED APPROACHES IN RELATED WORK

Approach	***Published Work***	***Area***
String-based	[25].	Arabic automated online exam scoring
Corpus-based	[17], [20], [21], [23]	Arabic automated short answer scoring
Hybrid approaches (String- based, Corpus based)	[8], [24].	Arabic automated short answer scoring
Hybrid approaches (string based, corpus based, knowledge based – WordNet)	[16]	Translate Arabic short answers into English for automated scoring
Word embedding & paraphrase embedding with cosine	[31]	English automated short answer scoring
Word embedding (Word2vec), contextual embedding (Bert)	[27], [32],[33].	English automated short answer scoring

To conclude, more research on how to leverage deep learning approaches, or use contextual embedding, for Arabic automatic short answer scoring is needed. Thus, this comparative empirical study attempted to implement three different approaches for the following feature presentation vectors: 1) word-embedding using Word2vec, 2) contextual-embedding using Bert, and 3) WordNet as the knowledge-based algorithm, with the Cosine algorithm to measure text similarity between student answers (SAs) and model answers (MAs). Furthermore, it also investigated the effect of stemming levels on the performance of the proposed approaches.

III. METHODOLOGY

This section outlines the proposed process for evaluating two different models for measuring text similarity for Arabic short answer questions: (1) knowledge-based similarity and (2) word and contextual embedding similarity. The two models were implemented by some suggested libraries of NLP

and the Python programming language. The research methodology is illustrated in Fig. 1.

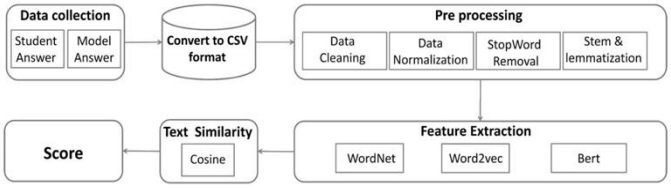

Fig. 1. Methodology.

In the first stage, data were collected, which include both SAs and MAs. In the second stage, data were converted to comma-separated value (CSV) format. In the third stage, the pre-processing took place, which consisted of four phases in the following order: data cleaning, data normalization, stop-word removal, and finally steam and lemmatization. In the fourth stage, the three different approaches (WordNet, Word2vec, and Bert) were tested to find the highest accuracy in AS. In the fifth stage, the similarity between the SAs and MAs was measured utilising the Cosine similarity. Finally, the scores were calculated.

A. Data Collection

In this study, data collection was based on two Arabic datasets provided by Ouahrani and Bennouar [14] and Rababah and Al-Taani [23].

The dataset in AR-ASAG [14] consists of three different exams with the MAs and SAs of three different classes collected from a cybercrime course exam. Each exam consists of 16 short answer questions, and each question on each exam has a different number of student answers. The dataset thus contains 2,133 SAs with a total of 48 questions.

The Arabic dataset produced in [23] consists of 11 questions from the official Jordanian History course exam. Each question includes the MA created by the teacher and the answers of 50 students, with an average of 50 words per answer. The questions in both datasets include one or more of the question types shown in Table II.

TABLE II. DATASET QUESTION TYPES

Arabic Question Type	*Translation*
عرف	Define
اشرح	Explain
علل	Justify
ما النتائج المترتبة على	What are the consequences
ما الفرق	What is the difference

B. Convert to CSV Format

A comma-separated value (CSV) file is a set text file that uses a comma to separate values. Each line of the file is a data record that consists of one or more fields separated by commas. After the datasets were obtained, the data were converted to the CSV format.

C. Pre-process

1) Cleaning data: cleaning the data is an essential process in text mining that removes the noise from the data and prepares the data for processing. Therefore, all the punctuation marks were removed, including full stops, commas, and parentheses, in order to make the data more understandable in the comparison with the correct answer in the MA. The difference in data before and after cleaning is shown in Table III.

TABLE III. DATA CLEANING

Student Answer before Data Cleaning	**Student Answer after Data Cleaning**	**Translation**
هي كل سلوك غير أخلاقي يتم باستخدام الوسائل الالكترونية (الهاتف، الكمبيوتر..)، يتمثل في حصول مرتكب الجريمة على ما يريد لتحقيق أهدافه الشخصية بينما يتحمل الضحية وهو المستخدم العقوبة، تتمثل في سرقة المعلومات	هي كل سلوك غير أخلاقي يتم باستخدام الوسائل الالكترونية الهاتف الكمبيوتر يتمثل في حصول مرتكب الجريمة على ما يريد لتحقيق أهدافه الشخصية بينما يتحمل الضحية وهو المستخدم العقوبة تتمثل في سرقة المعلومات	It is every immoral behaviour that takes place using electronic means (telephone, computer...), represented by the perpetrator obtaining what he wants to achieve his personal goals, while the victim, who is the user, bears the penalty, represented by stolen information.
هي سلوك غير قانوني عبر أجهزة إلكترونية، لأهداف مادية أو معنوية غالبا لإتلاف أو سرقة المعلومات وهي مثلا: النصب والاحتيال، التعدي الإلكتروني، التجسس وانتهاك الخصوصية	هي سلوك غير قانوني عبر أجهزة إلكترونية لأهداف مادية أو معنوية غالبا لإتلاف أو سرقة المعلومات وهي مثلا النصب والاحتيال التعدي الإلكتروني التجسس وانتهاك الخصوصية	It is illegal behaviour through electronic devices, often for material or moral purposes, to destroy or steal information, for example: fraud, electronic infringement, espionage and violation of privacy.
هي سلوك غير قانوني يتم باستخدام الأجهزة الإلكترونية، يتم تحميل المجرم منه على فوائد مادية ومعنوية، يتحمل الضحية خسارة مقابل ذلك الهدف من الجريمة إتلاف أو سرقة المعلومات	هي سلوك غير قانوني يتم باستخدام الأجهزة الإلكترونية يتم تحميل المجرم منه على فوائد مادية ومعنوية يتحمل الضحية خسارة مقابل ذلك الهدف من الجريمة إتلاف أو سرقة المعلومات	It is an illegal behaviour that takes place using electronic devices, for which the criminal is charged with material and moral benefits, and the victim bears a loss due to the goal of the crime, destroying or stealing information.

2) Normalization: At this stage, data were processed using advanced techniques by developing the normalization functions for some specific letters, as in Arabic, some letters are written in various forms. Thus, the Tashaphyne Library [34] was used for normalizing the following letters: Alef (إ،أ،آ،ا), Hamza (ؤ،ئ،ء), Ya'a (ي،ى), and Ha'a (ة،ه). Furthermore, other methods were used for removing diacritics format (known in as Tashkeel) as shown in Table IV.

TABLE IV. NORMALIZATION

Letters form	*Normalized into*	*Function name*
إ،أ،آ،ا	ا	Alef normalization
ى،ي	ي	Ya,a Normalization
ة،ه	ه	Ha'a Normalization
ؤ، و، ئ	ء	Hamza Normalization

3) StopWord: At this stage, words that have no meaning in NLP were removed from both the MAs and SAs, as they were not used as index terms and were not useful in AS. For example, from each MA and SA, all the following stop-words were removed: (' هيهات', 'إما', 'إذن', 'سوف', 'إليكم', 'ذاك', 'ذلكم', 'بيد', ' أنتن', 'هذا', 'حيث', 'ممن', 'بكن', 'هو', 'بما', 'كلا', 'لكم', 'بي', ''…).

4) Stem & Lemmatization: the aim of this stage was to employ stemming techniques to extract the root-base of each word. Stemming is a crucial method to process complex morphological words such as those in the Arabic language. This technique refers to the task of stripping prefixes, suffixes, and infixes from all words. This process also includes lemmatization, which extracts a relevant root-base called a lemma that refers to the dictionary of words [35].

In the research presented in this paper, two available NLP and morphological tools that provide a stemmer for the Arabic language have been utilised. The first tool was the FARASA library proposed by [36], which is an accurate stemmer based on SVM ranking for manipulating Arabic text. The second tool was the Arabic ISRI Stemmer, which is available in NLTK packages and designed to retrieve low-forms of words.

FARASA provided the light stem by removing prefixes, suffixes, and infixes, while ISRI conducted the base stem. Hence, they were used together to investigate the effect of stemming level on similarity accuracy. Table V presents the stemming process for the dataset of FARASA and ISRI, and Table VI shows examples of the resulting text from each pre-processing step.

TABLE V. THE STEMMING PROCESS FOR FARAS AND ISRI

Word	***FARASA***	***Arabic ISRI***
السلوك	سلوك	سلك
يتحدثها	تحدث	حدث
تأمين	تأمين	تأم
صحفيون	صحفي	صحف

TABLE VI. PRE-PROCESSING RESULTS

Answers	***After Cleaning and Normalization***	***After stop-word***	***Stem using FARASA***	***Stem using ISRI***
هي كل سلوك غير أخلاقي يتم باستخدام الوسائل الالكترونية (الهاتف، الكمبيوتر..) ، يتمثل في حصول مرتكب الجريمة على ما يريد لتحقيق أهدافه الشخصية بينما يتحمل الضحية و هو المستخدم العقوبة ، تتمثل في سرقة المعلومات مثلا	هي كل سلوك غير اخلاقي يتم باستخدام الوسائل الالكترونيه الهاتف الكمبيوتر يتمثل في حصول مرتكب الجريمه علي ما يريد لتحقيق اهدافه الشخصيه بينما يتحمل الضحيه و هو المستخدم العقوبه تمثل في سرقه المعلومات مثلا	سلوك اخلاقي يتم باستخدام الوساءل الالكترونيه الهاتف الكمبيوتر يتمثل حصول مرتكب الجريمه يريد لتحقيق اهدافه الشخصيه بينما يتحمل الضحيه و المستخدم العقوبه تمثل سرقه المعلومات مثلا	سلوك اخلاقي تم استخدام وساءل الكترونيه هاتف كمبيوتر تمثل حصول مرتكب جريمه أراد تحقيق اهداف شخصيه تحمل ضحيه و مستخدم عقوبه مثل سرق معلومة مثل	سلك خلق يتم باستخدام سءل الكترونيه هتف كمبيوتر مثل حصل ركب جرم يرد حقق هدف شخص بين حمل ضحه و خدم عقب مثل سرق علم ثلا
المعلومات مثلا				
علم احصائي لسلوك الانسان يتكون من علم (**bio** وقياس **metric**)	علم احصاءي لسلوك الانسان يتكون من علم و قياس	علم احصاءي لسلوك الانسان يتكون علم و قياس	علم احصاءي سلوك انسان تكون علم وقياس	علم حصء سلو انس يتك علم و قيس

D. Feature Extractions and Text Similarity Measure

1) WordNet: WordNet is a knowledge-based tool used to measure semantic similarity. It is a lexical database that places synonyms that have the same meaning, and which are not based on the form or linguistic similarity of the words, in groups called synsets [17]. The Arabic WordNet tool was created in 2006 and expanded in 2016 to include more synonyms. This technique is used to find similarly meaningful synonyms in SAs to increase accuracy in AS [24]; after pre-processing for SAs, Arabic WordNet tool is used to consider all the synonyms then measure semantic similarity using the Cosine similarity to find the similarity of both sentences. The result of normalizing the sentences is from 0 to 5.

2) Word2ve: Word2vec is a word-embedding technique that represents words as vectors of numbers in a vector space and trains the word vector with the aim of facilitating the process of measuring the similarity between these words, wherein the vectors that represent similar words are placed close to each other; the less similarity between words, the greater the distance between their vectors [37]. This method generates for each distinct word in the dataset a numerical representation referred to as a vector; after defining all the words that it can identify as having a key relationship with the vector, it calculates the angles between these vectors by using similarity measures. Word2vec performs its function through two basic models. The first is Continuous Bag-of-Words (CBOW), which works by predicting a word by looking at and combining the surrounding words that the word falls between. The second is the Skip-grams model, which performs the opposite process to the previous model, as it relies on a word to predict the surrounding words.

In this paper, the CBOW model was applied as it is faster than the other noted models and represented the frequent words in a more efficient way. AraVec was used to set up the Word2vec model, which is an open-source project that provides a massive set of pre-trained word-embedding models for Arabic NLP investigations. It has been created based on three fields of Arabic content: Wikipedia Arabic articles, Twitter tweets, and WWW pages. Furthermore, the Gensim Python library was used to load this model to extract embedding vector representation for each SA and MA by calculating average word embedding for each answer.

3) BERT: For addressing the contextual embedding between words, the study employed the Bidirectional Encoder Representations from Transformers (BERT) model [38]. Bert is a bidirectional model that is pre-trained in a deep sense regarding context and flow of language. Hence, this unlabelled data model can be fine-tuned throughout, adding further

output layers to support ultra-modern approaches that process different enormous jobs [33].

This work employed a pre-trained BERT model from the AraBert models' list. The bert-base-arabertv2 was predicted to extract layers where the external output layer was selected to draw out all remaining embedding layers. Thus, the proposed model utilized only the external node of the last embedding layer as it perfectly defined the sentences in a few dimensions. These embeddings will be closer to each other if they are more similar.

E. Text Similarity Measures

To measure the similarity between the character space vectors of an SA and an MA, the Cosine method was employed. Cosine works mathematically by calculating the Cosine of the angle between two vectors dropped down in a multi-dimensional space. The resulting similarity value falls in the range from -1 to 1. The -1 indicates strong non-similarity, while 1 refers to perfect similarity [39] and [27]. Thus, in this study, to align a predicted score with a human score, the data were normalized to 0-5.

IV. Results (Score) and Discussion

This section discusses the results of the comparative experiments that tested the three proposed approaches aimed at addressing Arabic automated short answer scoring. These experiments were conducted using data from two datasets: AR-ASAG and another dataset in [23]. Each dataset and approach were tested using the two mentioned stemming tools to examine the influence of the stemming level, light stem and base stem, on scoring precision for Arabic, which includes massive inflections. The proposed approaches were evaluated by comparing human scoring with the automated model scoring using the two most frequently mentioned measurement methods in related works for this area. All the experiments reported using both the Pearson correlation coefficient and Root Mean Squared Error (RMSE). The Pearson correlation coefficient is a precise measurement used to assess the linear relationship across two variables, represented in the range of -1, 1 to indicate the weakness or strength of the relationship, where the higher value is preferable. RMSE is the ideal method for measuring the variance between a predicted score and a true score. This method gives a non-negative value where, generally, a low RMSE is best. Table VII and Table VIII display the sample of grades to compare human scoring with model scoring.

A. Proposed Approaches on AR-ASAG

Table IX reports the results acquired from all three approaches that were applied to the dataset AR-ASAG. The first approach, WordNet with Cosine, achieved a relatively better Pearson correlation with the light stem (.75) while recording an RMSE with a lower value with the base stem. In Word2vec with Cosine, the light stem again produced an approximately higher Pearson correlation (0.7758) and lower RMSE (1.3577) than the base stem. Similar results can be observed in BERT with Cosine, with a light stem producing a Pearson correlation of 0.7616 and an RMSE value of 1.0439.

Overall, the Word2vec with Cosine resulted in the best Pearson correlation at .77 compared with the other approaches, while BERT with Cosine achieved the lowest RMSE with light stem on AR-ASAG.

B. Proposed Approaches on Dataset (Rababah & Al-Taani, 2017)

The same experiments were performed on the dataset [23] as that shown in Table X. WordNet, Word2vec, and BERT resulted in the lower RMSE with the base stem as 1.06, 1.12, and 1.003, respectively. The higher Pearson Correlation with base stem was achieved by Word2vec of .83 and BERT of 0.84. For this dataset, the BERT + Cosine approach recorded the best Pearson Correlation and RMSE.

TABLE VII. Sample of Human & Model Scoring on AR-ASAG

Approaches	*Human Scoring*	*Model Scoring*
WordNet+Cosine	3.5	3
	4	3
	5	2.5
	3.75	3
Word2vec+Cosine	4	3.79
	2	2.5
	4.5	4.33
	5	4.67
BERT+Cosine	2.5	2.5
	5	3
	2.25	3
	3	3

TABLE VIII. Sample of Human & Model Scoring on (Rababah & Al-Taani, 2017) Dataset

Approaches	*Human Scoring*	*Model Scoring*
WordNet+Cosine	2	3
	2	2.5
	2	1.5
	2	2
Word2vec+Cosine	1	1.5
	1	1
	1	2
	4.43	4.5
BERT+Cosine	1	1
	1	1.5
	2.5	3
	2	3

TABLE IX. The Result of Proposed Approaches on AR-ASAG

Approaches	*Stem*	*Pearson Correlation*	*RMSE*
WordNet+Cosine	Base	0.7469	1.4977
	Light	0.7553	1.4646
Word2vec+Cosine	Base	0.7693	1.3879
	Light	**0.7758**	1.3577
BERT+Cosine	Base	0.7536	1.4516
	Light	0.7616	**1.0439**

TABLE X. THE RESULT OF PROPOSED APPROACHES ON DATASET OF (RABABAH & AL-TAANI)

Approaches	*Stem*	*Pearson Correlation*	*RMSE*
WordNet+Cosine	Base	0.806195	1.1220652
	Light	0.820854	1.153378
Word2vec+Cosine	Base	0.837094	1.0655779
	Light	0.828779	1.1118528
BERT+Cosine	Base	**0.841902**	**1.00308459**
	Light	0.837253	1.0439487

The following observations are introduced based on the results of the above experiments. First, the light root and base root are approximately equivalent as they achieved close results to each other, as another study [14] also reported. The best among them cannot be determined here, as this work recorded that the optimal performed stemming level can differ with different datasets and various feature representation approaches. The light stemming was better when performed on the AR-ASAG dataset, while the other dataset had a higher Pearson correlation and lower REMS with the base stem.

Moreover, processing the contextual embedding between words has improved the accuracy of Arabic AES compared with other similarity measurements. The BERT with Cosine achieved the best RMSE across the two used datasets as the lowest RMSE was 1.00308. In addition, the best Pearson Correlation among all performed experiments was 0.841902 for the BERT algorithm.

V. CONCLUSION

This comparative empirical study evaluated the efficiency of different word embedding approaches in the context of Arabic automatic essay scoring (AES). Two available datasets were acquired, and several pre-processing methods were employed for these datasets. For the feature presentation, three different approaches were proposed to examine their efficiency as feature extract models in this domain. Therefore, the WordNet, Word2vec, and BERT approaches have been applied individually to extract the features of the student answer (SA) and model answer (MA), and the Cosine similarity was used to identify the closest-scoring model to human scoring by measuring the similarity between the SAs and MAs.

Four experiments were conducted for each proposed approach to study the effect of stemming techniques on the performance of these approaches. After that, Pearson correlations and RMSEs were calculated to compare the scores produced by the experiments with human scores. The results indicated that advanced models of contextual embedding can improve the efficiency of Arabic AES as the meaning of words can differ in the different contexts. Nevertheless, it is worth mentioning that these experiments were conducted on only two available Arabic short answers datasets, and hence the results are tied to them. The same experiments should be repeated on more datasets to get more generic results. Therefore, this area needs more investigation to improve the accuracy of Arabic AES in order for it to be realised as a practical online scoring system.

VI. FUTURE WORK

Future work could endeavour to present a benchmark dataset for the Arabic language. Furthermore, proposing a hybrid approach, such as combining WordNet with word-embedding and contextual-embedding, could enhance the accuracy of the approach.

REFERENCES

[1] M. Meccawy, Z. Meccawy, and A. Alsobhi, "Teaching and Learning in Survival Mode: Students and Faculty Perceptions of Distance Education during the COVID-19 Lockdown," Sustainability, vol. 13, no. 14: 8053, July 2021. https://doi.org/10.3390/su13148053.

[2] Z. Meccawy, M. Meccawy, and A. Alsobhi, "Assessment in 'survival mode': student and faculty perceptions of online assessment practices in HE during Covid-19 pandemic," Int. J. for Edu. Integ., vol. 17, no. 16, pp. 1-24, August 2021. https://doi.org/10.1007/s40979-021-00083-9.

[3] G. Di Pietro, F. Biagi, P. Costa, Z . Karpiński, and J. Mazza, "The Likely Impact of COVID-19 on Education: Reflections based on the Existing Literature and Recent International Datasets," In Publications Office of the European Union, Luxembourg: vol. EUR 30275, no. JRC121071, 2020. https://doi.org/10.2760/126686.

[4] D. Wiliam,"What is assessment for learning?," Stud. Educ. Eval, vol: 37, no:1, pp. 3-14, March 2011. https://doi.org/10.1016/j.stueduc.2011.03.001.

[5] D. D. Dixson, and F. C. Worrell, "Formative and summative assessment in the classroom," Theory Pract., vol. 55, no .2, pp. 153-159, March 2016. https://doi.org/10.1080/00405841.2016.1148989.

[6] N. Suzen, A. Gorban, J. Levesley, and E. Mirkes, "Automatic short answer grading and feedback using text mining methods," Procedia Comput. Sci, vol. 169, pp.726–743, 2020. https://doi.org/10.1016/j.procs.2020.02.171.

[7] B. Clauser, M. Kane, and D. Swanson, "Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems," Appl. Meas.Educ,vol.15,pp.413–432,2002. https://doi.org/10.1207/S15324818AME1504_05.

[8] W. Gomaa, and A. Fahmy, "Arabic Short Answer Scoring with Effective Feedback for Students," Int. J. Comput. Appl., vol: 86, no.2, pp. 35-41, January 2014. https://doi.org/10.5120/14961-3177.

[9] A.E. Magooda, M. Zahran, M. Rashwan, H. Raafat, and M. Fayek, "Vector Based Techniques for Short Answer Grading," In Proceedings of the 29th International Flairs conference, May 2016.

[10] J. Wang, and M.S. Brown, "Automated essay scoring versus human scoring: A comparative study," JTLA, vol. 6, no. 2, October 2007.

[11] D. Prasetya, A. Wibawa, and T. Hirashima,"The performance of text similarity algorithms," Int. J. Adv. Intell. Informatics, vol. 4, no. 1, pp.63-69, March 2018. https://doi.org/10.26555/ijain.v4i1.152.

[12] W.H. Gomma, and A. A. Fahmy, "A Survey of Text Similarity Approaches," Int. J. Comput. Appl., vol. 68, no. 13, pp.13–18, April 2013.

[13] T. Veale, "Wordnet sits the sat a knowledge-based approach to lexical analogy," In Proceedings of the 16th European Conference on Artificial Intelligence, pp. 606-610, August 2004.

[14] L. Ouahrani, and D. Bennouar, "AR-ASAG an Arabic dataset for automatic short answer grading evaluation," In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp. 2634–2643, May 2020.

[15] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," J. King Saud Univ., vol.33, no.5, pp. 497-507, June 2021.https://doi.org/10.1016/j.jksuci.2019.02.006.

[16] W.H. Gomma, and A. A. Fahmy, "Automatic scoring for answers to Arabic test questions," Comput. Speech. Lang., vol.28, no.4, pp. 833–857, July 2014. https://doi.org/10.1016/j.csl.2013.10.005.

[17] S. Awaida, B. Al-Shargabi, and T. Al-Rousan, "Automated Arabic Essays Grading System based on F-Score and Arabic WordNet,"JJCIT, vol.5, no.1, December 2019. https://doi.org/10.5455/jjcit.71-1559909066.

[18] E. F. Al-Shalabi, "An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations," IJCSI, vol. 13, no. 5, September 2016. https://doi.org/10.48550/arXiv.1611.02815.

[19] P. W. Foltz, "Latent semantic analysis for text-based research," Behav. res. meth. instrum. comput., vol. 28, no. 2, pp. 197–202, June 1996. https://doi.org/10.3758/BF03204765.

[20] M. Alghamdi, M. Alkanhal, M. Al-Badrashiny, A. Al-Qabbany, A. Areshey, and A. Alharbi, "A hybrid automatic scoring system for Arabic essays," AI Commun., vol. 27, no. 2, pp.103–111, 2014. https://doi.org/10.3233/AIC-130586.

[21] R. Mezher, and N. Omar, "A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring," J. Appl. Sci., vol. 16, no. 5, pp. 209–215, 2016. https://doi.org/10.3923/jas.2016.209.215.

[22] M. M. Refaat, A. Ewees, M. Eisa, and A. Sallam, "Automated Assessment Of Students' Arabic Free-Text Answers," Int. J. Cooperative. Inform. Syst. (IJICIS), vol. 12, no.1, pp. 213–222, January 2012.

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

[24] A. Shehab, M. Faroun, and M. Rashad, "An Automatic Arabic Essay Grading System based on Text Similarity Algorithms," IJACSA, vol.9, no.3, pp. 263-268, 2018. https://doi.org/10.14569/IJACSA.2018.090337.

[25] B. Al-shargabi, R. Alzyadat, and F. Hamad, "Aegd : Arabic Essay Grading Dataset For Machine Learning," J. Theor. Appl. Inf. Technol., vol. 99, no. 6, pp. 1329–1338, March 2021.

[26] A. R. Abbas, and A.S. Al-qazaz, "Automated Arabic Essay Scoring (AAES) using Vectors Space Model (VSM) and Latent Semantics Indexing (LSI)," Eng. Technol, vol. 33, no. 3, pp. 410–426, 2015.

[27] J. Lun, J. Zhu, Y. Tang, M. and Yang, "Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring," In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 09, pp. 13389-13396, 2020. https://doi.org/10.1609/aaai.v34i09.7062.

[28] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan, " A Memory-Augmented Neural Model for Automated Grading, " In Proceedings of the fourth ACM conference on learning @ scale, pp. 189-192, April 2017. https://doi.org/10.1145/3051457.3053982.

[29] T. Gong, and X. Yao, "An Attention-based Deep Model for Automatic Short Answer Score," IJCSSE, vol. 8, no.6, pp. 127–132, June 2019.

[30] M. Cozma, A.M. Butnaru, and R.T. Ionescu, "Automated essay scoring with string kernels and word embeddings," arXiv preprint arXiv:1804.07954, April 2018. https://doi.org/10.48550/arXiv.1804.07954.

[31] S. Hassan, A. A. Fahmy, and M. El-Ramly, "Automatic short answer scoring based on paragraph embeddings," IJACSA, vol. 9, no. 10, pp.397–402, 2018. https://doi.org/10.14569/IJACSA.2018.091048.

[32] M. Beseiso, and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring," IJACSA, vol. 11, no. 10, pp. 204–210, 2020. https://doi.org/10.14569/IJACSA.2020.0111027.

[33] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, R. and Arora, R. "Pre-Training BERT on Domain Resources for Short Answer Grading," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6071-6075, 2019. https://doi.org/10.18653/v1/D19-1628.

[34] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python: Analysing Text with the Natural Language Toolkit, O'Reilly Media Inc, 2009.

[35] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, 2018. https://doi.org/10.1109/MCI.2018.2840738.

[36] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations, pp. 11-16, June 2016. https://doi.org/10.18653/v1/N16-3003.

[37] L. Ma, and Y. Zhang, "Using Word2Vec to process big text data," In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), pp. 2895–2897, 2015. https://doi.org/10.1109/BigData.2015.7364114.

[38] V. Moshkin, A. Konstantinov, and N. Yarushkina, "Application of the BERT Language Model for Sentiment Analysis of Social Network Posts," In Artificial Intelligence: 18th Russian Conference, RCAI 2020, Moscow, Russia, October 10–16, 2020, Proceedings 18, pp. 274-283, Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-59535-7_20.

[39] A. Singhal, "Modern Information Retrieval: A Brief Overview," IEEE Data Eng. Bul., vol. 24, no. 4, pp. 35-43, 2001.