IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 March 2023, accepted 7 April 2023, date of publication 14 April 2023, date of current version 25 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267407

RESEARCH ARTICLE

Automatic Arabic Grading System for Short Answer Questions

RASHA M. BADRY, MOSTAFA ALI, ESRAA RSLAN, AND MOSTAFA R. KASEB
Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum 63514, Egypt

Corresponding author: Mostafa R. Kaseb (mrk00@fayoum.edu.eg)

This work was supported in part by the Egypt Newton-Mosharafa Fund 30812 Partnership under Newton Institutional Grant 347762518, and in part by the 'U.K. Department for Business, Energy and Industrial Strategy' and 'Science and Technology Development Fund (STDF)' and delivered by the British Council.

ABSTRACT The era of technology and digitalization has been advantageous to the educational sector. The examination system is one of the most important educational pillars that have been affected. As automatic exam grading is a revolution in the history of exam development, and therefore the automatic grading system has started to replace the traditional assessment system. The automatic grading system allows the examiners to automatically assign grades for students' answers compared to the model answers. And, generate results based on the examiners' answers. In this paper, we especially address the short answer questions. Most research has been done on the English language. On the other side, few research works have been conducted on Arabic. Moreover, Arabic is considered one of the rare resource languages. This paper is aimed to build an Automatic Arabic Short Answer Grading (AASAG) model using semantic similarity approaches. It is used to measure the semantic similarity between the student and model answer. The proposed model is applied to one of the Arabic scarce publicly available datasets which is called (AR-ASAG). It contains 2133 pairs of models and student answers in several versions such as txt, xml, and db. The efficiency of the proposed model was evaluated through two conducted experiments using two weighting schemas local, and hybrid local and global weighting schema. The developed approach with hybrid local and global weight-based LSA achieved better results than using local weight-based LSA with (82.82%) as F1-score value, and 0.798 as an RMSE (Root-Mean-Square Error) value using hybrid local and global weight-based LSA.

INDEX TERMS Short answer grading system, Arabic language, global weight-based LSA.

I. INTRODUCTION

In the teaching and learning process, assessment plays a critical role. There are two major classifications for questions in the exam assessment are named closed-ended (e.g. multiple-choice, true/false, and matching questions) and open-ended questions (e.g. short answer and essay questions). In any natural language, the assessment of open-ended questions is a much more difficult process since it requires a special text analysis. Most automatic assessment management systems support close-ended questions [1]. On the other hand, open-ended questions have been studied for many years but it is still an under-research area. Thus in this research, short answer questions of open-ended questions class will be addressed. There is a need to compare the student's short answer with the model answer and assign a score for the student's answer based on the similarity score between them. Similarity is defined as determining whether two concepts (e.g. word, sentence, or paragraph) are similar to each other or not. There are two ways that concepts can be similar: lexically or semantically. If text words have similar character sequences, then concepts are lexically similar. Semantic similarity is described as deciding whether or not two concepts have similar meanings. There are two categories for measuring semantic similarity corpus-based similarity and knowledge-based similarity [2], [3]. In knowledge-based similarity, it uses semantic data from the knowledge graph to determine how similar the concepts are to one another. The path connecting

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

two concepts in a knowledge graph represents the semantic distance between concepts. There are two measures of knowledge-based similarity that are semantic similarity and semantic relatedness. While in corpus-based similarity, it is a semantic similarity metric that determines how similar two terms are which is based on the large corpora's information such as HAL, LSA, ESA, and NGD [2], [4], [5]. In this research, the corpus-based similarity is the focus research similarity metric, especially the use of Latent Semantic analysis LSA. This research explores the Latent Semantic Analysis (LSA) algorithm as a metric the corpus-based text similarity. In this approach, local weighting schema, and hybrid local and global weighting schema are employed as two weighting schemas for the data representation.

Many automatic grading systems have been conducted on English language. However, few works have been done on Arabic language. Arabic is considered one of the most popular and complex natural languages. Moreover, The Arabic language has a high degree of ambiguity, extensive morphology, complicated morpho-syntactic agreement rules, and a significant number of irregular forms [6], [7]. The aim of this research is to build an Automatic Arabic Short Answer Grading (AASAG) system using semantic similarity approaches. It is used to measure the semantic similarity between the Student Answer (SA) and Model Answer (MA) [8], [9]. Comparing to the manual short answer grading, AASAG system also aims to reduce time, effort, and achieve fairness in the scoring process of students' answers. The proposed model is applied on one of the Arabic scarce publicly available dataset which is called (AR-ASAG).

This study introduces two contributions. Firstly, proposing an effective automatic Arabic short answer scoring model using semantic similarity approaches since it is especially well suited for languages with limited resources. The model is based on Latent Semantic Similarity LSA which is used to measure the semantic similarity between the Student Answer (SA) and Model Answer (MA). Secondly, employing two weighting schemas for the data representation. Where it is used for determining the data important to effectively measure the semantic similarity. The effects of hybrid weighting schema are more effective and it achieves a promising positive results for the Arabic language.

This paper is organized as follows: Section II presents related work on automatic short-answer grading systems. Section III explains the proposed short answer grading system. Section IV shows the experimental results, and finally, the research conclusions is given in Section V.

II. RELATED WORK

The proposed model aimed to build an Automatic Arabic Short Answer Grading (AASAG) model using semantic similarity approaches. The model is based on LSA which is used to measure the semantic similarity between the Student (SA) and Model Answer (MA). It is based on two weighting methods used for filling the cell values are local weighting schema, and hybrid local and global weighting schema.

Various approaches have been proposed for automatic short answers grading systems. Here, we go over some works that are closely related and deals with automatically grading for short answers questions.

In [10], they introduced AR-ASAG an Arabic Dataset for Automatic Short Answer Grading Evaluation. Moreover, they proposed Automatic Short Answer Grading approach which is based on COALS (Correlated Occurrence Analogue to Lexical Semantic) algorithm. The proposed approach achieved promising results for Arabic language. The approach was tested on the introduced AR-ASAG dataset. The Dataset contains 2133 pairs of (Model Answer, Student Answer) in several versions (txt, xml, Moodle xml and .db). In [11], the authors presented an Automatic Arabic Essay Scoring (AAES) using Vector Space Model (VSM) and Latent Semantic Indexing (LSI). AAES was applied on one question with four model answers and 30 student responses. The proposed model is mainly based on two processes. First, the information retrieval techniques were used to retrieve the most important information from the electronic essay. Then, VSM and LSI were used to be used to measure the similarity degree between the student and the instructor essay. The authors in [12], proposed an automatic Arabic essay grading (AAEG) system using Support Vector Machine. It is main idea is to extract features from the student and model answers. Moreover, finding the related words from the student answers using the AWN. The proposed model was applied on multiple language such as English and French. The model has been applied on kaggle dataset with 40 questions with 120 model answers. AAEG with AWAN achieved a better accuracy than AAEG without AWAN. In [13], the authors proposed an automatic scoring systems for short Arabic texts using the sentence embedding approach. The proposed model applied on three different datasets are: AraScore, AR-ASAG, and two translated answer sets. The model achieved the best results with AraScore dataset. Moreover, authors in [14] proposed automatic scoring system for Arabic Short answers using Longest Common subsequence (LCS) and Arabic WordNet (AWAN). First, AWAN is used for providing the synonyms of each student answer. Then, LCS is used to modify the proximity degree of the student and model answer. The model was applied on dataset with 330 student's answers. It achieved 0.81 value as a Root Mean Square Error (RMSE) and 0.94 value as a Pearson correlation r. In [15], the authors proposed an automatic grading for Arabic short answer questions using optimized deep learning model. They used a hybrid LSTM and GWO model to predict the short answer grade questions for the students in science. They used a dataset gathered about science subject from various schools in the Qalyubia-Governorate of the Egypt Arab Republic. The proposed LSTM-GWO model achieved the best accuracy compared to LSTM, SVM, SVM-GWO, Ngram, Word2vec, Arabic WordNet, and MaLSTM. It achieved the lowest RMSE value the best R square value, and the highest Pearson correlation coefficient value. The authors in [16] proposed a scoring model for short answers grading called Ans2vec. The model is based on measuring the similarity between the student and model answer. Ans2vec model applied on three different datasets: Texas, Cairo University and SCIENTSBANK. Ans2vec achieved 0.63 as a Pearson correlation value.

Finally, the authors in [17] proposed an Arabic short answer scoring system using deep learning. The authors assured that the system provide a kind of help for the Arab teachers to save their times in doing other activities toimprove the education quality. They proposed AraScore used the baseline model, RNN, LSTM, and Bi-LSTM. In addition, they used two transformer-based language models called BERT and ELECTRA. The best results have been achieved when using ELECTRA with 0.78 as QWK score value. This paper was considered one of the first deep learning researches in the Arabic short answer scoring system.

According to the previous related works, almost all of the use of LSA in the automatic Arabic scoring of short answer questions has been ignored. The main goal of the proposed system is to evaluate the student's answers in a fair and accurate manner using LSA which is one of the most popular a corpus-based similarity techniques. And is the most suitable method for capturing the most descriptive features of the text semantics. In addition, most related works applied on small size of the dataset. Finally, they achieved a low value of accuracy for Arabic language.

Table 1 briefly describes a comparison of different related and significant works in automatic Arabic scoring systems.

III. THE RESEARCH CONTRIBUTION

The proposed model aimed to provide an Automatic Arabic Short Answer Grading (AASAG) model. The model focused on the semantic similarity metric that uses the term-weighting of the model answer (MA) and the student answers (SA). The model is applied on Arabic language. Arabic is one of the most popular languages and there are hundreds of millions are speaking in Arabic. Although, few research works have been done on Arabic language in contrast of English language. As we will mention after that the accuracy of the Arabic grading model has been affected and improved by applying the semantic space method. In addition, the model is based on two term-weighting methods are local weighting schema, and hybrid local and global weighting schema. The accuracy of the Arabic grading system is improved by using the hybrid local and global term-weighting schema. We used Latent Semantic Analysis (LSA) technique for semantic space method to measure the semantic similarity between the SA and MA for three main reasons.

First, LSA is one of the most popular corpus-based similarity techniques. Second, LSA overcomes the large number of document vector space dimensions by implementing mathematical calculations, and therefore implementing the document vector space with reduced dimensionality. Third, LSA is the most suitable method for capturing the most descriptive features of the text semantics [5].

The proposed model is applied to one of the Arabic scarce publicly available datasets which is called (AR-ASAG). AR-ASAG [10] is the first Arabic publicly and freely available dataset. It contains questions taken from the cybercrimes teaching course and the responses of three classes of master students. There are a total of 2133 student responses in the dataset as shown in Table 2. There is a suggested model response for each question. Two human experts assessed the responses independently on a scale from 0 (totally inaccurate) to 5 (perfect answer). Both of the experts were instructors in computer science. AR-ASAG considered the gold standard is the average grade of the two experts.

There are several versions of the AR-ASAG Dataset, including TXT, XML, XML-MOODLE, and Database (.DB). Table 3 displays a question-answer pair with three representative student responses and the two manual grades given by the experts.

The proposed model consists of three main modules namely data pre-processing, LSA implementation, and finally, the semantic score module as shown in Fig 1. In the following, each component of this model will be described in details starting from the input datasets to the output artifacts.

A. DATA PREPROCESSING MODULE

Data preprocessing is one of the most important steps in enhancing the performance of the proposed model. The main purpose of this process is to get the data into the best possible form for data analysis and modeling. It consists of the following actions:

1) TOKENIZATION

It is the process of breaking up texts into words. Arabic sentences are broken up using punctuation to indicate the end of each sentence. The text is divided into sentences using a series of punctuation symbols, such as commas (,), semicolons (;), question marks (?), exclamation points (!), colons (:), and periods. The space is used to separate words. For example: “عرف مصطلح أمن المعلومات” (Define the information security?); after applying tokenization: ‘ “عرف” “المعلومات” ‘ “أمن“” ‘ “مصطلح” (“Define,” “the,” “information,” “security”)

2) STOP WORDS REMOVAL

Some words don't convey any information about the content and less meaningful. Such words must be eliminated since they are high-frequency such as “الى” (to), “فى” (in),“على” (on). In the proposed model, Farasa Arabic stop words list will be used [18].

3) LEMMATIZATION

It plays a vital role in many applications of natural language processing and is a crucial step in the pre- processing stage. Lemma is the process of finding the basic form of the words. For example, in Arabic words such as “الوسائل”(Means) has

TABLE 1. Summary of related works in automatic arabic scoring systems.

Author	Technique	Dataset	Advantages	Limitations
L.Ouahrani et al. [10]	COALS (Correlated Occurrence Analogue to Lexical Semantic)	AR-ASAG	Is best suited for languages with limited resources, like Arabic	low accuracy
R.Abbas et al. [11]	VSM (Vector Space Model) and LSI(Latent Semantic Indexing)	One question with four model answers and 30 student responses	- Provide an electronic assessment closer to traditional assessment in about only 10 seconds. - It was applied on multiple language such as English and French	Small dataset
A. E. E. Elalfi et al. [12]	Word2vec, and SVM (Support Vector Machine)	Kaggle dataset	Is considered the word synonyms to generate model answer alternatives	Small dataset
A.ElNaka et al [13]	The sentence embedding approach, and feed forward deep neural network architecture	AraScore-Datase, AR-ASAG, and Hewlett Foundation SAS	Scaling an efficient and unbiased content scoring applications across different Arabic educational domains	Achieved a good results only with the new proposed AraScore dataset
H.A. ABDELJABER [14]	LCS Longest Common subsequence , and AWAN (Arabic WordNet)	Dataset collected from educational institution of 330 students' answers	The model can be applied in many Arabic applications such as scoring short answers of Arabic essay questions and detecting plagiarism of Arabic textual assignments	- Small size of dataset - Feedback and correction for the assessment process are not provided
M. Abdul Salam et al [15]	LSTM (Long Short Term Memory) and GWO (Grey Wolf Optimizer).	Datasets collected in the Science subject for students in seventh grade in Egypt.	Using a hybrid optimized deep learning model.	It required a higher training time than the traditional deep learning model.
WH.Gomaa et al. [16]	It is based on Skip-thought vectors	Three different datasets: Texas, Cairo University and SCIENTSBANK	Best RMSE value of 0.81 and Pearson correlation r value of 0.94	Available only for English - Not support the Arabic language
O. Nael et al. [17]	Baseline model, RNN, LSTM, and Bi-LSTM	ASAP Short Answer Scoring dataset hosted by kaggle	- Using a deep learning pre-trained language models - Best QWK score of 0.78	- It isn't a good solution to use a translated dataset using Google Translate API - less data quality

TABLE 2. About AR-ASAG dataset.

Dataset	AR-ASAG
Language	Arabic
Course	Cybercrime
Questions	48
Answers	2133
Availability	Yes

the root "وسيلة"(Mean). For such a MADAMIRA lemmatization system is used [19].

B. LSA IMPLEMENTATION MODULE

Applying LSA is the second module in the proposed model. This module is used to measure the semantic similarity between the Student Answer (SA) and Model Answer (MA). LSA is one of the powerful unsupervised analytical technique. It is one of the most well-known algorithm for information retrieval applications. Through singular value decomposition (SVD), it can expose the hidden structure of individual words, groups of words, sentences, or texts. Additionally, it generates measures of word-word, word-document, and document-document interactions that are highly linked with a number of association and semantic-related human cognitive processes [20]. There are two steps for LSA-based algorithm: term weight representation and applying SVD Calculations.

C. TERM WEIGHT REPRESENTATION

In term weight representations, the terms of the student and model answer are weighted to build the term incidence matrix which is represented as a m × n matrix (A). A matrix with $A = [a_{1j}, a_{2j}, \ldots a_{nj}]$ represents each row as a term and each column as a sentence. The word importance is represented by the cell value (a_{ij}). In the proposed model, the weighting schema for the cell value $W(a_{ij})$ is calculated using [21] (1)

TABLE 3. Question-answer pair with three representative student responses and the two manual grades.

Sample Question	عرف مصطلح أمن المعلومات Define the information security?	Mark 1	Mark 2	AVG
Model Answer	حماية وتأمين كافة الموارد المستخدمة في معالجة المعلومات من منشآت نفسها والأفراد العاملين فيها وأجهزة الحاسب المستخدمة فيها ووسائط المعلومات التي تحتوي على البيانات وذلك في جميع مراحل تواجد المعلومة (التخزين – النقل – المعالجة) Protecting and securing all resources used in processing information such as facilities, individuals, computers, and the information media that contain data, in all information stages (storage- transmission - processing).			
Student Answer 1	هي الوسائل والأدوات والسياسات والإجراءات المستخدمة لضمان خصوصية و تكامل وتوافر المعلومات ومنع الوصول اليها او حذفها او مسها بدون حق It is the means, tools, policies and procedures used to ensure the information privacy, integrity and availability and to prevent unauthorized access, or deletion.	4	4	4
Student Answer 2	هو الوسائل التقنية والفنية والإدارية اللازم توفيرها لحماية معالجة المعلومة (تخزينها، نقلها، معالجتها) It is the technical and administrative means that must be provided to protect the processing of information (storage, transmission, processing).	3	3.5	3.25
Student Answer 3	هو تأمين المنشأة نفسها وموظفيها وأنظمتها و أجهزتها و الوسائط التي تحتوي على المعلومة في كل مراحل تواجدها وتخزين (نقل ومعالجة) It is the security of the facility, employees, systems, devices, and media that contain information in all information stages (transmission - processing).	4	3.5	3.75

and (2).

$$W(a_{ij}) = L(t_{ij}) \tag{1}$$

$$W(a_{ij}) = L(t_{ij}) \times G(t_{ij}) \tag{2}$$

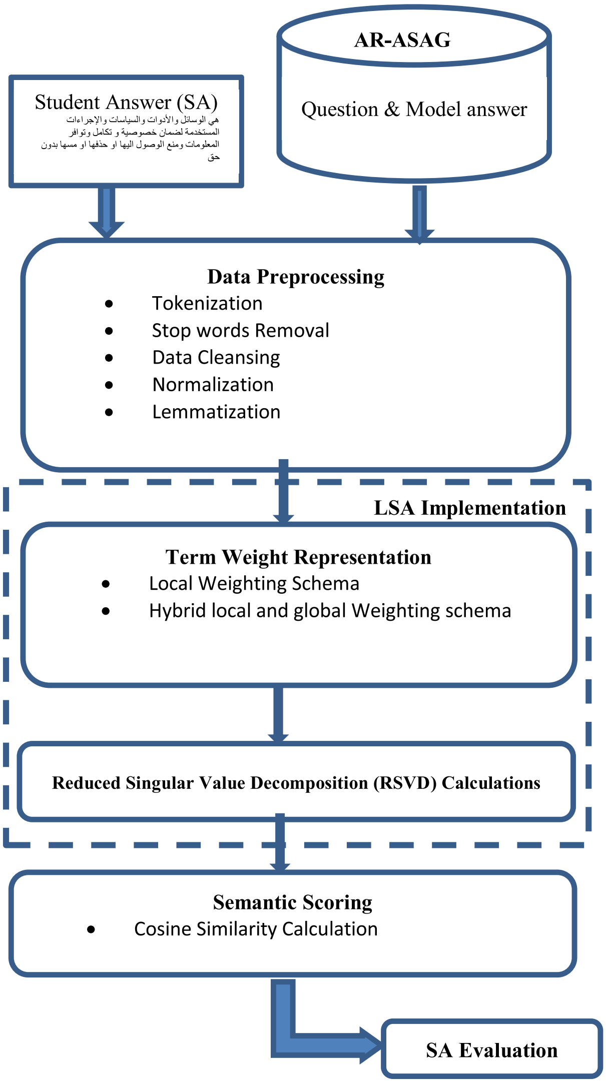
where $L(t_{ij})$ is the local weight for the term i in sentence j, and

$G(t_{ij})$ is the global weight for the term i in the whole document.

There are various weighing techniques that can be used to determine the local weight, and the global weight. These techniques include the following:

- Local weight
  There are various methods for calculating the local weight [22], as follows:
  a) Binary Representation (BR): $L(t_{ij}) = 1$, if a term i exists in a sentence j, otherwise $L(t_{ij}) = 0$.
  b) Term Frequency (TF): $L(t_{ij}) = tf(_{ij})$, where $tf(_{ij})$ is the number of times that the term i occurs in a sentence j.
  c) Augment weight (AW): AW is calculated by (3)

$$L(t_{ij}) = 0.5 + 0.5 \times tf\,(_{ij})/tf\,(_{max}) \tag{3}$$



Student Answer (SA)
هي الوسائل والأدوات والسياسات والإجراءات
المستخدمة لضمان خصوصية و تكامل وتوافر
المعلومات ومنع الوصول اليها او حذفها او مسها بدون
حق
AR-ASAG
Question & Model answer
Data Preprocessing
Tokenization
Stop words Removal
Data Cleansing
Normalization
Lemmatization
LSA Implementation
Term Weight Representation
Local Weighting Schema
Hybrid local and global Weighting schema
Reduced Singular Value Decomposition (RSVD) Calculations
Semantic Scoring
Cosine Similarity Calculation
SA Evaluation

FIGURE 1. The proposed architecture for AASAG model.

where tf $(_{ij})$ is the number of times that the term i occurs in a sentence j, and tf$(_{max})$ refers to the frequency of the most frequent term in sentence j.

- Logarithm Weight (LW): LW is calculated by (4)

$$L(t_{ij}) = 1 + log(tf\,(_{ij})) \tag{4}$$

- Global weight
  There are various methods for calculating the global weight [22], as follows:

a) No Global Weight (NG): $G(t_{ij}) = 1$.
b) Inverse Sentence Frequency (ISF): ISF is calculated by (5)

$$G(t_{ij}) = 1 + Log(n/n_i) \tag{5}$$

where n is the total number of sentences, and n $(_i)$ is the number of sentences that contain the term i.

TABLE 4. Question-answer pair with three representative student responses.

Sample Question	عرف مصطلح أمن المعلومات؟ Define the information security?
Model Answer	حماية وتأمين كافة الموارد المستخدمة في معالجة المعلومات من منشآت نفسها والأفراد العاملين فيها وأجهزة الحاسب المستخدمة فيها ووسائط المعلومات التي تحتوي على البيانات وذلك في جميع مراحل تواجد المعلومة (التخزين – النقل – المعالجة) Protecting and securing all resources used in processing information such as facilities, individuals, computers, and the information media that contain data, in all information stages (storage- transmission - processing).
Student Answer 1	هي الوسائل والأدوات والسياسات والإجراءات المستخدمة لضمان خصوصية و تكامل وتوافر المعلومات ومنع الوصول اليها او حذفها او مسها بدون حق It is the means, tools, policies and procedures used to ensure the information privacy, integrity and availability and to prevent unauthorized access, or deletion.
Student Answer 2	هو الوسائل التقنية والفنية والإدارية اللازم توفيرها لحماية معالجة المعلومة (تخزينها، نقلها، معالجتها) It is the technical and administrative means that must be provided to protect the processing of information (storage, transmission, processing).
Student Answer 3	هو تأمين المنشأة نفسها وموظفيها وأنظمتها و أجهزتها و الوسائط التي تحتوي على المعلومة في كل مراحل تواجدها وتخزين (نقل ومعالجة) It is the security of the facility, employees, systems, devices, and media that contain information in all information stages (transmission - processing).

c) Entropy Frequency(EF): EF is calculated by (6)

$$G(t_{ij}) = 1 + \sum\nolimits_{j=1}^{n} \left(\frac{pi \log pij}{\log n} \right) \tag{6}$$

where

$$p_{ij} = tf_{ij}/gf_i \tag{7}$$

where tf ($_{ij}$) is the number of times that the term i occurs in a sentence j, n is the total number of sentences j, and gf_i refers to the number of times that the term (t_i) occurs in the whole document.

In the proposed model, the Term Frequency (TF) weighting schema is used to calculate the local weight to fill the cell values for data representation. While the inverse sentence Frequency (ISF) will be used to calculate the global weight.

D. RSVD CALCULATIONS

Singular Value Decomposition (SVD) is an algebraic technique that can be used to determine the relationships between words and sentences [23]. For the matrix A (m × n), there is a SVD for matrix A of the following (8) form [23]:

$$\boldsymbol{M} = \boldsymbol{U \Sigma V^T} \tag{8}$$

where U is an m × n unitary matrix,

Σ is an m × n is the diagonal matrix with non-negative real numbers on the diagonal representing the scaling values, and

$\mathrm{V^T}$ (the conjugate transpose of V) is an n × n real or complex unitary matrix.

Further in order to improve the performance of the SVD, Reduced SVD (RSVD) is used to reduce the noise and reduce the number of dimensions that are irrelevant.

TABLE 5. An example of the proposed system steps.

Process	Result	Translation
Tokenization	"هو" ,"الوسائل","التقينية", "و","الفنيية " ,"و", "الإدارية"," اللازم", "توفيرها ","لحماية"," معالجة", "المعلومة", "تخزينها","نقلها", "معالجتها"	“Is”, “Mean”, “The Technical”, “And” ,“The Technical”, “And”, “Administrative”, “Necessary For”, “Providing”, “Protecting”, “Processing”, “Information”, ”Stored”, “Transferred”, “Processed”
Stop words removal	"الوسائل","التقينية", "الفنيية", "الإدارية", "اللازم","توفيرها", "لحماية" , "معالجة", "المعلومة", تخزينها", "نقلها", "معالجتها"	“Means”, “Technical”, “Technical”, “Administrative”, “Necessary”, “Provide”, “Protecting”, “Processing”, “Information”, ” Stored”, “Transferred”, “Processed”
Data cleansing	"الوسائل", "التقنية", "الفنية", "الإدارية", "اللازم","توفيرها", "لحماية" ,"معالجة", "المعلومة","تخزينها", "نقلها", "معالجتها"	"Means", "Technical", "Technical", "Administrative", "Necessary", "Provide", "Protecting", "Processing", "Information", ” Stored”, “Transferred”, “Processed”
Normalization	"الوسائل", "التقنيه", "الفنيه" ,"الادارية", "اللازم","توفيرها", "لحمايه","معالجة", "المعلومه","تخزينها", "نقلها", "معالجتها"	"Means", "Technical", "Technical", "Administrative", "Necessary", "Provide", "Protecting", "Processing", "Information", “Stored”, “Transferred”, “Processed”
Lemmatization	"وسيلة", "تقني", "فني","إداري", " لازم", "توفير", "حمايه" ,"معالجه", "معلومه" ,"تخزين", " نقل" ,"معالجة"	"Mean", "Technical", "Technical", "Administrative", "Necessary", "Provide", "Protection", "Process", "Information", “Stored”, “Transferred”, “Processed”

E. SEMANTIC SCORING

After applying the SVD calculations, the semantic similarity score is calculated between the student and the model answer. Cosine similarity is one of the most common semantic similarity methods. It is calculated by using (9):

$$\boldsymbol{Cos\ similarity(A,B) = A.B/\|A\| * \|B\|} \tag{9}$$

where *Cos similarity* (*A*, *B*) is the similarity score between the student answer and model answer, A is the weight of the term in the student answer, and B is the weight of the term in the model answer statement.

F. STUDENT ANSWER (SA) EVALUATION

The Student Answer was evaluated compared to the model answer using the generated semantic score. Finally, the grades were assigned to the student answers compared to the model answers.

IV. RESULTS AND DISCUSSION

This section presents the results and evaluation of the proposed Automatic Arabic Short Answer Grading (AASAG) system which is based on LSA. Two experiments have been conducted using two different weighting for data representation are local and global weighting schema.

TABLE 6. Semantic similarity score using local and hybrid local and global weighting schema based LSA between the students answers and model answer.

Sample Question	Model Answers	Students' Answers	Semantic similarity score using local weight	Semantic similarity score using hybrid local &global weight
عرف مصطلح أمن المعلومات؟ Define the information security?	حماية وتأمين كافة الموارد المستخدمة في معالجة المعلومات من منشآت نفسها والأفراد العاملين فيها وأجهزة الحاسب المستخدمة فيها ووسائط المعلومات التي تحتوي على البيانات وذلك في جميع مراحل تواجد المعلومة (التخزين – النقل – المعالجة) Protecting and securing all resources used in processing information such as facilities, individuals, computers, and the information media that contain data, in all information stages (storage-transmission - processing).	هي الوسائل والأدوات والسياسات والإجراءات المستخدمة لضمان خصوصية و تكامل وتوافر المعلومات ومنع الوصول اليها او حذفها او مسها بدون حق It is the means, tools, policies and procedures used to ensure the information privacy, integrity and availability and to prevent unauthorized access, or deletion.	**0.821**	**0.843**
		هو الوسائل التقنية والفنية والإدارية اللازم توفيرها لحماية معالجة المعلومة (تخزينها، نقلها، معالجتها) It is the technical and administrative means that must be provided to protect the processing of information (storage, transmission, processing).	**0.736**	**0.755**
		هو تأمين المنشأة نفسها وموظفيها وأنظمتها و أجهزتها و الوسائط التي تحتوي على المعلومة في كل مراحل تواجدها وتخزين (نقل ومعالجة) It is the security of the facility, employees, systems, devices, and media that contain information in all information stages (transmission-processing).	**0.711**	**0.757**

TABLE 7. Students' answers grades based on the semantic similarity score using local and hybrid local and global weighting schema based LSA.

Student No	Student's Answers	Student Grade using local weight	Student Grade using hybrid local &global weight	AR-ASAG Mark 1	AR-ASAG Mark 2	AR-ASAG Gold standard (AVG)
Student answer 1	هي الوسائل والأدوات والسياسات والإجراءات المستخدمة لضمان خصوصية و تكامل وتوافر المعلومات ومنع الوصول اليها او حذفها او مسها بدون حق It is the means, tools, policies and procedures used to ensure the information privacy, integrity and availability and to prevent unauthorized access, or deletion.	**4.1**	**4.2**	4	4	4
Student answer 2	هو الوسائل التقنية والفنية والإدارية اللازم توفيرها لحماية معالجة المعلومة (تخزينها، نقلها، معالجتها) It is the technical and administrative means that must be provided to protect the processing of information (storage, transmission, processing).	**3.7**	**3.8**	3	3.5	3.25
Student answer 3	هو تأمين المنشأة نفسها وموظفيها وأنظمتها و أجهزتها و الوسائط التي تحتوي على المعلومة في كل مراحل تواجدها وتخزين (نقل ومعالجة) It is the security of the facility, employees, systems, devices, and media that contain information in all information stages (transmission - processing).	**3.6**	**3.8**	4	3.5	3.75

Consider the following example to investigate and evaluate the proposed model. Table 4 shows an example for question-answer pair with three representative student responses.

In order to evaluate and assign grades for the student answers, the steps of the proposed model were executed as follows: Firstly, the pre-processing steps on the students' answers and model answer were applied such as tokenization, stop word removal, data cleansing, normalization, and lemmatization as shown in Table 5.

Secondly, the term weight matrix is created using the local weight based LSA which is the first experiment. In addition, term weight matrix using the hybrid local and global weighting based LSA is created as the second experiment. Thirdly, SVD calculations for the created matrices are created and applied. Fourthly, the SVD values are used to measure the semantic similarity score between the student answer and the model answer using the cosine similarity method as shown in Table 6.

TABLE 8. The semantic evaluation results of using local weight and hybrid local &global weight.

Measures	Hybrid local & global weight	Local weight
Precision	83.23%	72.24%
Recall	81.41%	70.65%
F1-score	82.82%	70.36%

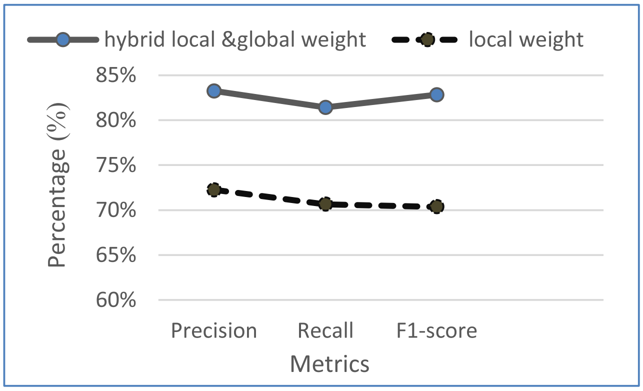


FIGURE 2. Precision, Recall, F1-score values for the proposed AASAG system.

Finally, the student's answers were scored and the grades were generated according to the calculated semantic score as shown in Table 7. It is noticed that the calculated students' grades using both methods of the proposed model are quite similar to both of the experts' marks. Therefore, the proposed model achieved the AR_ASAG gold standard.

The conducted experiments and results are presented to evaluate the proposed AASAG performance with different measures. Different measures are used to evaluate the performance of the proposed AASAG system measures such as Precision, Recall, and F1-score. In addition, Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE), Normal MAE (NMAE), and *Pearson correlation* (r: the higher the better) are used. They are also the most frequently used metric and they are well-known metrics used as a baseline to evaluate the model. In the following, we present the results of experiments for 40 answers for each question. As shown in Table 8, the proposed system achieved 83.23%, 81.41%, and 82.82% as Precision, Recall, and F1 score value respectively using hybrid local &global weight. While using local weight, the proposed system achieved 72.24%, 70.65%, and 70.36% as Precision, Recall, and F1 score value respectively. In conclusion, the results of the hybrid local and global weight-based LSA approach better results than those using the local weight-based LSA approach. In addition, Fig 2 shows the precision, recall, and F1-score for the proposed AASAG system (hybrid local &global weight, local weight).

It is noticed that the proposed AASAG model using hybrid local and global weight achieved better results than using local weight. Finally, Table 9 shows the comparative results acquired from using RMSE, MAE, and NMAE metrics with semantic analysis. They were compared with the previous system in [6] which is Grading System Assessment based on the stemming. The proposed AASAG system achieved better results than the grading system based on stemming similarity because the hybrid local &global weight in our model selects the most important and frequent features of the text.

TABLE 9. RMSE, MAE, and NMAE evaluation results of the proposed models compared with the previous stemming model.

Models	RMSE	MAE	NMAE	Pearson
Stemming model	0.862	0.813	0.2427	0.723
Proposed AASAG model using local weight	0.825	0.774	0.2305	0.731
Proposed AASAG model using hybrid local & global weight	0.798	0.738	0.2113	0.745

V. CONCLUSION

An automatic scoring system is a useful tool for grading open-ended questions such as short answers and essay questions. The automatic scoring system has numerous benefits; reducing the manual process, time, effort, and wasted resources. In addition, achieving fairness in the scoring process of students' answers. In this paper, Automatic Arabic Short Answer Grading is introduced. The proposed model is based on LSA which is one of the most popular corpus-based similarity techniques. It is applied to AR-ASAG which is an. AR-ASAG is an Arabic scarce publicly available dataset. Two experiments have been conducted using two different weighting for data representation local and global weighting schema. Two experiments have been conducted using two weighting schemas local weight, and hybrid local and global weight schema. The developed approach with hybrid local and global weight-based LSA achieved better results than using the proposed local weight-based LSA with (82.82%) as a F1-score value. And, it achieved 0.798 as a RMSE value. Further, the proposed two weighting methods achieved better results compared to the related work.

In future work, we focus on enhancing the accuracy of the grading system. Moreover, the proposed system will be tested on other additional languages. Finally, we will employ the Arabic WordNet for developing an effective system for scoring short answer questions.

REFERENCES

[1] S. W. N. Cheung, S. C. Ng, and A. K. F. Lui, "A framework for effectively utilising human grading input in automated short answer grading," *Int. J. Mobile Learn. Organisation*, vol. 16, no. 3, p. 266, 2022.

[2] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 13, no. 1, pp. 11–23, 2013.

[3] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. AAAI*, vol. 6, 2006, pp. 775–780.

[4] R. A. Farouk, M. H. Khafagy, M. Ali, K. Munir, and R. M. Badry, "Arabic semantic similarity approach for Farmers' complaints," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, 2021.

[5] E. Rslan, M. H. Khafagy, K. Munir, and R. M. Badry, "English semantic similarity based on map reduce classification for agricultural complaints," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 12, pp. 1–8, 2021.

[6] W. H. Gomaa and A. A. Fahmy, "Automatic scoring for answers to Arabic test questions," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 833–857, Jul. 2014.

[7] N. Y. Habash, "Introduction to Arabic natural language processing," *Synth. Lectures Human Lang. Technol.*, vol. 3, no. 1, pp. 1–187, Jan. 2010.

[8] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 567–575.

[9] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 752–762.

[10] L. Ouahrani and D. Bennouar, "AR-ASAG: An Arabic dataset for automatic short answer grading evaluation," in *Proc. Lang. Resour. Eval. Conf.*, Marseille, France, 2020, pp. 2634–2643.

[11] R. Abbas and A. S. Al-Qazaz, "Automated Arabic essay scoring (AAES) using vectors space model (VSM) and latent semantics indexing (LSI)," *Eng. Technol. J.*, vol. 33, no. 3, pp. 410–426, 2015.

[12] A. E. E. Elalfi, A. F. Elgamal, and N. A. Amasha, "Automated essay scoring using Word2vec and support vector machine," *Int. J. Comput. Appl.*, vol. 177, no. 25, pp. 20–29, Dec. 2019.

[13] A. ElNaka, O. Nael, H. Afifi, and N. Sharaf, "AraScore: Investigating response-based Arabic short answer scoring," *Proc. Comput. Sci.*, vol. 189, pp. 282–291, 2021.

[14] H. A. Abdeljaber, "Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet," *IEEE Access*, vol. 9, pp. 76433–76445, 2021.

[15] M. Abdul Salam, M. A. El-Fatah, and N. F. Hassan, "Automatic grading for Arabic short answer questions using optimized deep learning model," *PLoS ONE*, vol. 17, no. 8, Aug. 2022, Art. no. e0272269.

[16] W. Hassan and A. A. Fahmy, "Ans2vec: A scoring system for short answers," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, 2020, pp. 586–595.

[17] O. Nael, Y. Elmanyalawy, and N. Sharaf, "AraScore: A deep learning-based system for Arabic short answer scoring," *Array*, vol. 13, Mar. 2022, Art. no. 100109.

[18] K. Darwish and H. Mubarak, "Farasa: A new fast and accurate Arabic word segmenter," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 1070–1074.

[19] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2016, pp. 11–16, doi: 10.18653/v1/n16-3003.

[20] B. Li, Z. Li, T. Li, and J. Liu, "A portable embedded automobile exhaust detection device based," in *Proc. IEEE 3rd Int. Conf. Inf. Sci. Technol. (ICIST)*, Mar. 2013, pp. 126–128, doi: 10.1109/ICIST.2013.6747520.

[21] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 234–265.

[22] J.-H. Lee, S. Park, C.-M. Ahn, and D. Kim, "Automatic generic document summarization based on non-negative matrix factorization," *Inf. Process. Manage.*, vol. 45, no. 1, pp. 20–34, Jan. 2009.

[23] D. Kalman, "A singularly valuable decomposition: The SVD of a matrix," *College Math. J.*, vol. 27, no. 1, pp. 2–23, Jan. 1996.

RASHA M. BADRY received the B.Sc., M.Sc., and Ph.D. degrees from the Faculty of Computers and Artificial Intelligence, Helwan University, Egypt, in 2003, 2007, and 2015, respectively. She is currently a Lecturer with the Department of Information Systems, Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt, where she is also the Director of Crisis Management. She is the Director of the National Bank for Scientific Laboratories and Equipment, Supreme Council of Universities, Egypt; a member of the Big Data Research Group, Fayoum University; and a Researcher in international project with the University of the West of England (Newton-Mosharfa). Her research interests include NLP, machine learning, and data science.

MOSTAFA ALI received the B.Sc. and M.Sc. degrees from Assiut University, Egypt, in 2006 and 2013, respectively, and the Ph.D. degree from the Department of Computer Science, Mysore University, India, in 2020. He is currently a Lecturer with the Department of Information Systems, Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt. He is also the Vice Manager of the National Electronic Exam Center, Supreme Council of Universities (SCU), Egypt; a member of the Big Data Research Group, Fayoum University; and a Researcher in international project with the University of the West of England (Newton-Mosharfa). His research interests include text mining, NLP, machine learning, big data, and data science.

ESRAA RSLAN received the B.S. degree from the Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt, in 2012, the M.Sc. degree from the Faculty of Computers and Artificial Intelligence, Cairo University, Egypt, in 2018, and the Ph.D. degree from Fayoum University, in 2022. From 2019 to 2021, she was a Researcher in international project with the University of the West of England (Newton-Mosharfa). She is currently a Lecturer with the Faculty of Computers and Artificial Intelligence, Fayoum University, and a member of the Big Data Research Group, Fayoum University. Her research interests include big data, NLP, and database.

MOSTAFA R. KASEB received the B.Sc. degree in electronics and communications engineering from the Faculty of Engineering, Fayoum University, Egypt, in 2006, and the M.Sc. and Ph.D. degrees in computer engineering from the Department of Electronics, Communications and Computer Engineering, Faculty of Engineering, Helwan University, in 2011 and 2019, respectively. He is currently the Assistant Dean of Education and Student with the Faculty of Computers and Artificial Intelligence, Fayoum University; a member of the Big Data Research Group, Fayoum University; and a Researcher in international project with the University of the West of England (Newton-Mosharfa). His research interests include parallel and distributed systems, parallel processing, grid computing, cloud computing, database, parallel programming, business process management, and big data analysis.