AUTOMATED ARABIC ESSAY SCORING BASED ON HYBRID STEMMING WITH WORDNET

Mohammad Alobed[1*], Abdallah M M Altrad[2], Zainab Binti Abu Bakar [3], Norshuhani Zamin[4]

[1,2,3] Faculty of Computer Science and Information Technology, Al-Madinah International University, Malaysia

[4] College of Computing and Informatics, Saudi Electronic University, Saudi Arabia

Email: shamona.2018@gmail.com[1*](corresponding author), abdallah.mahmoud@mediu.edu.my[2], zainab.abubakar@mediu.edu.my[3], n.zamin@seu.edu.sa[4]

DOI: https://doi.org/10.22452/mjcs.sp2021no2.4

ABSTRACT

Schools, universities, and other educational institutions have been forced to close their doors because of the coronavirus outbreak. E-learning has become an option and has long been discussed about the need to integrate it into the educational process-learning uses a variety of evaluation methods, one of which is the essay. This research introduces a new model for Arabic Automated Essay Grading (AAEG) that has been developed to reduce human bias mistakes and costs while saving time. However, (AAEG) is still in its infancy. The model relies on new hybrid stemming with Arabic WordNet (AWN). The primary goal of stemming is reducing inflectional forms of words to root words. The hybrid method is based on different techniques: Extended Light Stemmer, ISRI, and looking at tables (AWN). Data used in this study consists of 3050 words with their roots were retrieved from (AWN) and then stemmed using algorithms (Light10, ISRI, Hybrid...). For evaluation, the metrics used were accuracy, precision, recall, and F1-score. While comparing the performance of the different stemming algorithms, the hybrid stemming method had the greatest results, therefore the (AAEG) will improve with Hybrid Stemming.

Keywords: Information Retrieval, Similarity measure, Automated Essay Grading, Arabic WordNet, Stemming

1.0 INTRODUCTION

Teachers spend around 30% of their time marking students' answers [1, 2]. As a result, the advantages of adopting automated essay scoring include reduced human bias, errors, and expenses, as well as time saving [3]. The research on the Automated Essay Scoring System was done in a variety of languages, including English, French, and a little bit of Arabic [4]. However, (AAEG) is still in its infancy [5]. When using an automated essay scoring system, it's feasible to offer students with personalized feedback remarks on their answers online, whereas manual evaluation takes a lot of time to do [6]. Public schools, universities, and institutes have expressed interest in automated essay scoring.

In the wake of the Coronavirus outbreak, schools, colleges, and other educational institutions were forced to lock their doors in order to prevent the virus from spreading. So, educational institutions began using E-learning, a long-discussed alternative to classroom instruction. From this perspective, the relevance of electronic examinations as an alternative to traditional paper exams becomes apparent.

Using an automated scoring system, the authors in [7] created an essay question for Arabic short answers. Cosine similarity is used to calculate the degree of similarity between the answer reference and a student's answer in this study. After a series of natural processing steps (tokenization, stop word removal, normalization, etc.), the next stage is root extraction and synonym search, followed by TF-IDF word weight calculation. After that, cosine similarity is used to measure proximity. Throughout the evaluation process, the Recall method is employed. The achieved correlation is 95.4 percent, the validity ratio is 84.5 percent, and the percentage recall value for the correlation is 62%.

As the name suggests, Automated Essay Scoring (AES) is a system that uses computers to grade essays. It is intended to reduce mistakes and injustice due to human bias by reducing time and costs. E-rater is an automatic scoring method established by the Educational Testing Service in 1998 to rate GMAT essays. Essays are scored using mathematical and natural language processing (NLP) techniques. How well E-rater models perform in a real-world setting is determined by the quality of training and evaluation data. The correlation between human assessors and the system varied between 0.87 and 0.94 [8].

Students' answers to essay questions in Arabic will be evaluated using a new model that will offer a score that is similar to that supplied by the teacher manually. The model relies on the semantics of the Arabic WordNet (AWN) to achieve accuracy and to avoid weaknesses in stemming, after using a hybrid method of stemming with (AWN) tables.

2.0 LITERATURE REVIEW

An ontology is a collection of concepts and terminology inside a domain, as well as the connections between these concepts. A common understanding of knowledge structure is required for people or software agents, which is why ontologies were created. Each term-concept pair is called a sense. A set of senses is called synset. Concepts and senses are described by attributes [9]. Ontologies can be represented as graphs. A simple example of an ontology is shown in (Fig 1).

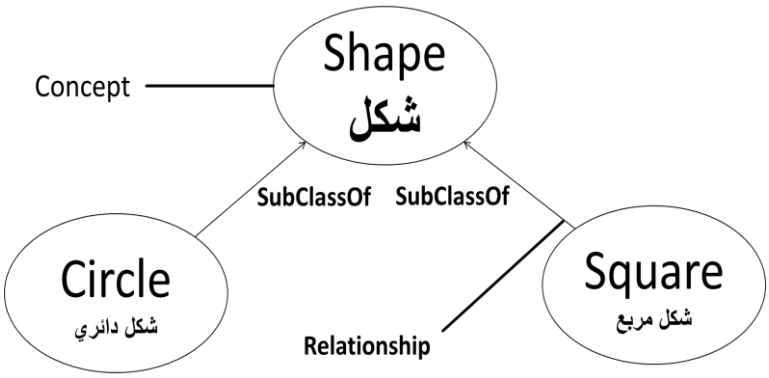

Fig 1. Ontology overview

There are, of course, additional approaches that employ formal requirements for knowledge representation, such as vocabularies, taxonomies, thesauri, topic maps, and logical models, as shown in (Table 1). However, unlike taxonomies or relational database schemas, ontologies provide the important advantage of offering a means to describe connections, thereby enabling users to link numerous ideas together in a number of flexible ways.

Table 1. Relationships and senses

A lexicon	A collection of words, concepts, and phrases connected with a certain knowledge domain. There are no definitions, no explanations of how words are linked. A lexicon is similar to a vocabulary.
A glossary	Like lexicon but with glosses added
A taxonomy	• A set of terms that are organized in a hierarchical structure. • Most of the time, this relationship is by hypernym, also called the "is-a" relationship [10].
Folksonomy	Lexicon and a taxonomy but not deeply structured as a taxonomy [11]
Polysemy	Is the ability of a word to possess several meanings. E.g., bright: 'shining'; 'intelligent' [12]
Synonym	Words that have the same meaning or that are closely related in meaning E.g., automobile/car; answer/reply [12]
Synset	Is a set of synonyms that define a concept or word meaning [13, 14]
Hyperonymy	Relation between a concept and its superordinate E.g., Food is a hypernym of cake [12]
Hyponymy	Relation between a concept and its subordinate [12]
Holonymy	Relation between a whole and its part E.g., Car is a holonymy of wheel [12]
Meronymy	Relation between a part and its whole E.g., Wheel is meronymy of car [12]
Antonomy	Words that have opposite meanings E.g., Good ⇔ Bad; Day ⇔ Night [12]
Troponym	More specific way of doing an action E.g., Run is troponym of move [15]
Entailment	If by doing *X* you must be doing *Y* E.g., Buying entails paying [15]
Gradation	Antonyms and meronymy relationships E.g., Temperature Hot, Warm, Cold

Like a taxonomy, an ontology entails relationships like hypernym. An ontology may have hypernym relationships between multiple overlapping taxonomies. Ontologies include additional types of relationships that are usually binary. Hyponym/hypernym (Is-A relationships), also called subordination/superordination, subset/superset, or is-a relation. Other binary relations commonly used in ontology are: part-whole, property, and value. Meronym/holonym (Part-of relationships), also called "part-whole" or "has a relationship".

Concepts, relations, properties, and axioms are the main components of an ontology. A concept represents a set of entities within a domain. Relations show how concepts are related to one another. This concept is described by the properties attribute. Finally, axioms are descriptions of concepts, properties, and relations via logical expressions. (Table 2) represents a logic axiom syntax for ontology.

Table 2. A logic syntax for ontology

Superclass of all	T
A is a subclass of B	$A \sqsubseteq B$
Class intersection	$A \sqcap B$
Class union	$A \sqcup B$
Class equivalence	$A \equiv B$

Each binary relation and the two entities x and y connected by it is called a "semantic triple". The type of triple stored in specialised databases is called a "triple store". These triples can be connected whenever they share an entity in common. This is how graph databases are constructed from semantic triples.

2.1 Arabic Ontologies

Many Arabic ontologies were developed to facilitate processing of natural Arabic languages. Arabic WordNet is the most widely used in Arabic ontology. WordNet is an electronic lexical database, which contains the most commonly encountered terms (nouns, adjectives, adverbs, verbs), terms grouped by context in this dictionary. WordNet provides short, general definitions, and records conceptual-semantic and lexical relations between these synonym sets. WordNet has a similar structure as ontologies. There are many different ways that senses can be related to one another in wordnet [16].

Arabic WordNet is considered one of Modern Standard Arabic's best semantic and lexical thesauruses. AWN consists of terms (nouns, verbs, adjectives, and adverbs), identified with their origins, definitions (synsets), and associations between these definitions. Some core relations are shown here in (Table 3) [17].

Table 3. Relations of Arabic WordNet

Relationships available in WordNet	
All parts of speech	
Synonymy	Words that have the same meaning
Antonymy	Words that have opposite meanings
Glossary	Store a gloss for every synset
Similar	Synsets that have similar meaning
Verb only	
Troponymy	More specific way of doing an action
Entailment	If by doing verb (X) you must be doing verb(Y)
Principle	Relation between verbs and adjectives
Nouns only	
Hypernymy	Relation between a concept and its superordinate
Hyponnymy	Relation between a concept and its subordinate
Meronymy	Relation between a part and its whole
Attribute	Relation between noun and adjective synsets
Adjectives only	
Participle	Relation between verbs and adjectives
Pertain	A lexical relation between two words
Attribute	Relation between noun and adjective synsets
Adverbs only	
Pertain	A lexical relation between two words
Synsets semantic relations	

Nouns	
hypernyms	*X* is a (kind of) *Y*
hyponyms	*Y* is a (kind of) *X*
holonym	*X* is a part of *Y*
meronym	*Y* is a part of *X*
coordinate terms	nouns sharing a common hypernym
Verbs	
hypernym	*X* is a (kind of) *Y*
troponym	*Y* is doing *X* in some manner
entailment	if by doing *X* you must be doing *Y*
coordinate terms	verbs sharing a common hypernym
Adjectives	
related nouns	
participle of verb	
Adverbs	
Pertain	root adjectives

Arabic WordNet has the following four components (tags):

- Item: The term concepts
- Word: the terms (phrases)
- Form: The words source
- Link: The Concept Relationship

The (AWN) is used wildly, since it is a wide ontology free to use and covers multiple domains. (Table 4) presents Arabic WordNet statistics [18].

Table 4. Arabic WordNet database statistics

POS	**Arabic WordNet (AWN)**	
	Word Forms	**Synset**
Noun	13,330	7961
Verb	5595	2536
Named entities	1426	1155
Broken plurals	405	126
Total	**20,756**	**11,778**

2.2 Similarities measures

In computing similarity, two data items are compared numerically. Increasing similarity increases the similarity value. A similarity is usually not a negative number; it is more likely to be between 0 and 1. In computing, dissimilarity (distance) is a measure of how two data objects differ. With increasing resemblance, the dissimilarity value decreases. Dissimilarities are common in [0, 1] [19].

Cosine similarity is a computation based on angles. It calculates the cosine angle between two vectors and determines how closely two texts are related. In the case of near vectors, the angle is small and the significance is great. The cosine value of 1 is 100% similar, while the value of 0 is 100% different [20].

$$\cos(d_1, d_2) = \frac{d_1 * d_2}{|d_1||d_2|} \quad (1)$$

Mathematical models called vector space models are commonly used to represent text as an identifiably organized vector. Using these models, we can determine if different texts have comparable meanings, regardless of whether they have the same words in them [20]. The number of times each term appears in a text is known as Term Frequency (TF). The document frequency of a term is the number of documents in which the term occurs.

$$TF(t) = (\text{Number of times term t appears in a document}) / (\text{Total number of terms in the document}) \quad (2)$$

Inverse Document Frequency (IDF) used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score.

$$IDF_{(W,J)} = \log \frac{\text{Total number of documents}}{\text{Number of documents with term w in it}} \quad (3)$$

Term frequency–inverse document frequency is a term weighting scheme commonly used to represent textual documents as vectors.

$$TF\text{-}IDF = TF * IDF \quad (4)$$

2.3 Pre-processing

The process of cleaning and preparing data for analysis is called pre-processing. In the pre-processing stage of tokenization, document strings are divided into smaller pieces. A stop word is a phrase that has no meaning or relevance in the search process. Next, stop words are removed from the text once it has been converted to a token list.

Stemming is a method of generating verb stems or roots by either extracting affixes connected to its root using a dictionary or by developing a collection of linguistic rules to detect verb patterns and therefore extracting the root [21].

In Arabic Language, there are two mains stemming approaches: the root-based approach and the light stemming approach. Light stemming algorithms extract only prefixes and suffixes from terms, while root algorithms delete prefixes, suffixes and infixes.

Light stemming approaches such as light1, light2, light3, light8, light10, and extended-light10 are used to extract prefixes and suffixes from words (Table 5). The primary disadvantage of light-based stemming is that it leads to a number of errors. Furthermore, eliminating suffixes and prefixes may result in additional uncertainty. As a result, stemming rules with a list of affixes are critical for properly determining if a list of suffixes and prefixes may create a known root.

Table 5. Light stemming

	Prefixes	**Suffixes**
Light1	ال، وال، بال، كال، فال	-
Light2	ال، وال، بال، كال، فال، و	-
Light3	ال، وال، بال، كال، فال، و	ه، ة
Light8	ال، وال، بال، كال، فال، و	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي
Light10	ال، وال، بال، كال، فال، لل، و	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي
Ext Light10	فل، ولل، وبال، لل، فال، كال، بال، وال، ال ل، ب، و، تت، فب، وب، ول	ت، هم، نا، هما، تي، وا، ي، ة، هـ، ية، يه، ين، ون، ات، ان، ها

Tashaphyne is a light stemmer that eliminates diacritics and normalizes input (Hamza, for example). The stemming procedure removes a large number of common prefixes and suffixes [22].

Abainia et al. presented a novel method for light stemmers called ARLStem to deal with prefix, suffix, and infix. List of prefixes that will be deleted from ARLStem (ال،لل،فب،بال،كال،وال،فلل،ولل،فبال،وبال،فكال،ست،سي،سا،سن،لن،لت،لي،لأ) and suffixes (كي ، كم ، ها ، كما ، كنّ ، هما ، هنّ ، ان ، ين ، ون ، تان ، تين ، ون ، تما ، تنّ ، نا ، تم ، وا ، تا ، ت ، ا ، ن). ARLStem, on the other hand, begins by normalizing the word and deleting the connector letter Waw ("و"). It also transforms feminine to masculine, plural to singular, and removes verb prefixes and suffixes [23].

As an alternative to Light10, Motaz has developed an Arabic light stemmer that would eliminate prefixes and suffixes from the word. Prefixes (ال، وال، بال، كال، فال، لل، و) and suffixes (ي ، ة ، ه ، ية ، يه ، ين ، ون ، ان ، ها) are employed.This is followed by normalization of Hamza (ء), Yaa (ي), Taa (ة) and the removal of the prefix before the stemmer begins [24].

Darwish suggested Farasa, which segmented the word into a series of prefixes, stem, and a stem of suffixes. These prefixes and suffixes will be deleted. The prefixes are (ف ، و ، ل ، ب ، ك ، ال ، س) and suffixes are (ا ، ة ، ت ، ك ، ن ، و ، ي ، ات ، ان ، ون ، وا ، ين ، كما ، كم ، كن ، ه ، ها ، هما ، هن ، هم ، نا ، تما ، تم ، تن). It includes a collection of lexicons for stem, Gazetter, Buckwalter, and AraComLex [25].

In Arabic, the Khoja stemmer is one of the most well-known stemmers, which is responsible for finding the root of any given word. It is based on a root-stemmer. First, the Khoja algorithm strips out all the prefix and suffix combinations before using pattern matching to find the roots. However, nouns have proven to be particularly difficult to decipher [22, 26].

Following the same procedure as Khoja stemmer, ISRI is used to determine the root-base of Arabic words. After eliminating potential suffixes, the ISRI stemmer looks for probable matches within a group pattern. Input word stemming should be terminated when there are three or less characters left. There is no way for ISRI to check roots against dictionaries of any type. In the absence of a dictionary, the extracted roots may not be valid, or they may consist of a nonsensical collection of letters.

Summarizing, (Table 6) compares all prior stemmers based on certain criteria, including the kind of stemmer used, stop-words and lexicon used, as well as prefixes and suffixes.

Table 6. Stemmers' comparison

Stemmer	**Type**	**prefixes**	**suffixes**	**Dictionary**
Light10	light	7	10	NO
Ext-Light10	light	16	16	NO
Tashaphyne	light	241	293	NO
ARLStem	light	19	23	NO
Motaz	light	7	10	NO
Farasa	light	7	21	NO
Khoja	root	11	28	YES
ISRI	root	15	28	NO

3.0 RESEARCH METHODOLOGY

This research introduces a new model for (AAEG). The model can evaluate students' answers to Arabic essay questions with a score that is close to that provided by the teacher manually. The model relies on the new hybrid stemming with (AWN). Hybrid methods are based on different techniques: Using the Extended Light Stemmer, ISRI, and Look in Tables (AWN) to check the extracted stems for accuracy, finding a noun or adverb in Arabic WordNet to preserve without stemming, and finding all synonyms for all reference answer words in Arabic WordNet, then using the tf-idf method to weight words, and applying semantic similarity through the cosine similarity method to find the angle between the answers of students and the reference answer. Data used in this study consists of 3050 words with their roots were retrieved from (AWN) and then stemmed using stemming algorithms (Light10, ISRI, Hybrid...). After that, the stemming outcomes were compared. For evaluation, the metrics used were accuracy, precision, recall, and F1-score. While comparing the performance of the different stemming algorithms, the hybrid stemming method had the greatest results. Therefore, the (AAEG) will improve with hybrid stemming. The steps of Proposed technique as shown in (Fig 2)

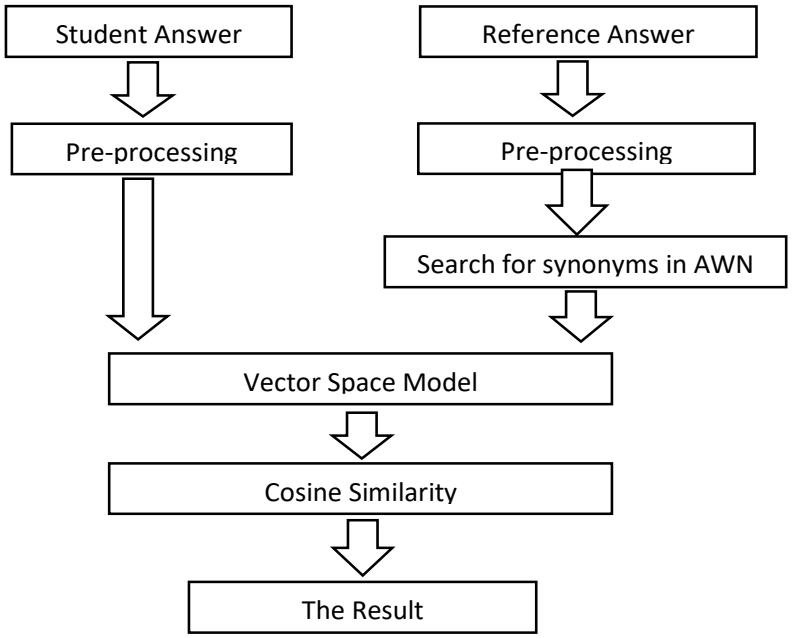

Fig 2. The components of the proposed technique

3.1 Pre-processing

The steps of Pre-processing as shown in (Fig 3).

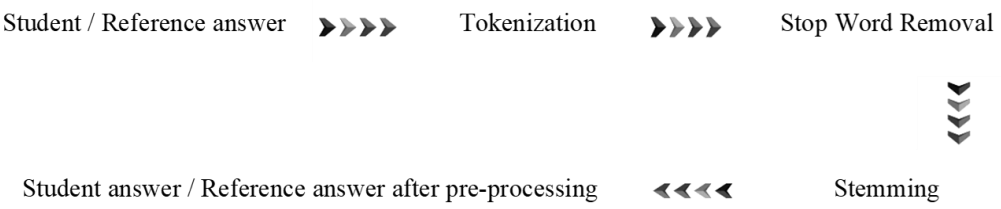

Fig 3. Pre-processing

Hybrid stemming method: Research suggests the following methods of evaluating stem correctness. The new technique makes it feasible to delete letters ("ب" ,"أ", and "و") from words' beginnings, add letters ("ال") to words' beginnings, or remove letters ("هـ") from words' ends. These (AWN) tables are then used to evaluate extracted stems, address problems with broken plural stems, and find terms that can be retained without stemming. The components of the proposed technique are shown in (Fig 4).

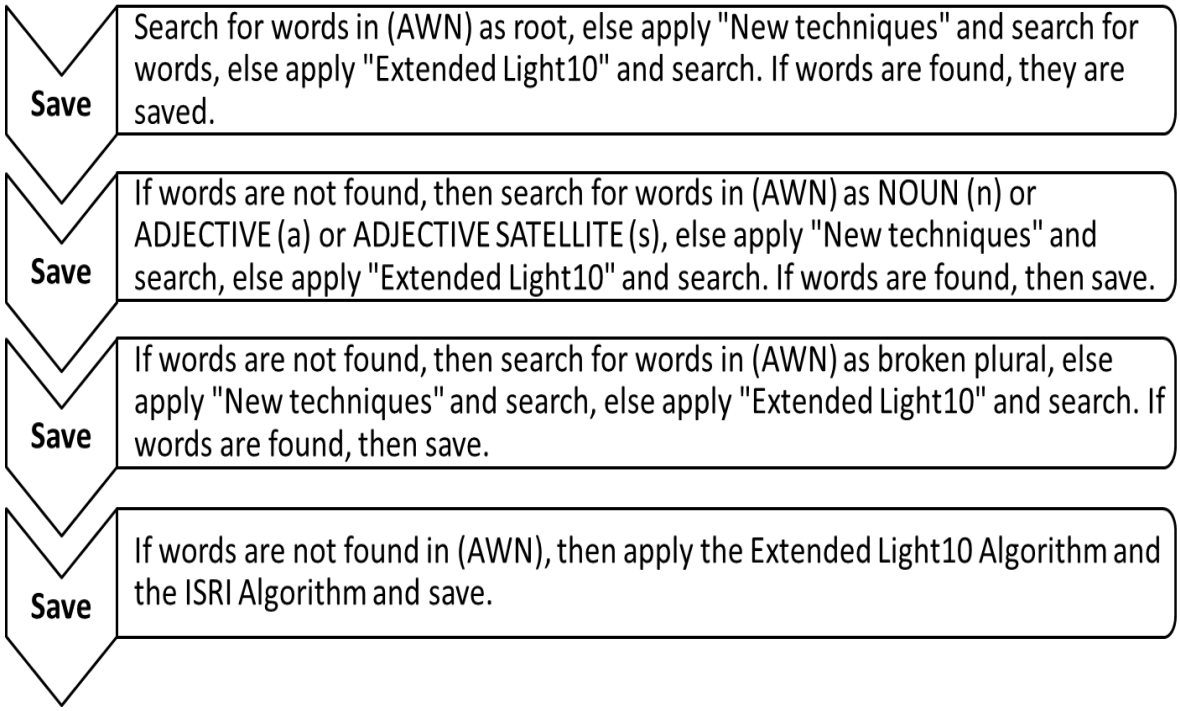

Fig 4. The components of the proposed technique

4.0 PERFORMANCE EVALUATION METRICS

In classification issues, the confusion matrix is a common tool. Both binary and multiclass classification issues may be solved using this technique. An example of a binary classification confusion matrix. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are statistical metrics that may be used to evaluate the accuracy of document categorization algorithms. (Fig 5) illustrates the Confusion Matrix [23].

		Actual Value	
		Positive	**Negative**
Predicted	**Positive**	(True Positive)	(False Positive)
	Negative	(False Negative)	(True Negative)

Fig 5. confusion matrix for binary classification

This matrix is based on the following (Table 7):

Table 7. Possible classification outcomes

True Positives (TP)	Yes, as expected, and yes, as labeled.
True Negatives (TN)	No, as expected, and no, as labeled.
False Positives (FP)	Yes, as expected, and no, as labeled.
False Negatives (FN)	No, as expected, and yes, as labeled.

A total of 3050 words grouped in (noun, adverb, adjective, broken Plural, verb) with their roots were retrieved from Arabic WordNet and then stemmed using Light10, ExtendLight10, ISRI, Light10, and ISRI, Extend and ISRI, and Hybrid Arabic Stemmer. After that, the stemming outcomes were compared.

In fact, accuracy is probably the most commonly used metric of all. To find out how often the classifier is correct, the formula **(TP+TN)/Total** may be used [27].

Precision and recall are used to measure how successfully a system finds relevant materials when a user requests them as follows:

Precision = **Total number of documents retrieved that are relevant/Total** number of documents that are retrieved.

Recall = **Total number of documents retrieved that are relevant/Total number** of relevant documents in the database.

F1-score **Precision and recall** are taken into account in the calculation of an algorithm's performance using F1-score (also known as f-measure or f-score) [25].

Precision, recall, and F-score are useful metrics for assessing the performance of Boolean retrieval systems. However, they cannot be used to evaluate ranks. (Table 8) shows the confusion matrix for word stem prediction.

Table 8. Confusion matrix for words stem prediction

	Predicted Stem (1)			Predicted Stem (0)		
Stem (1) is true	✓	TP	👍	✗	FN	👎
Stem (0) is error	✗	FP	👎	✓	TN	👍

Predicted that stemming algorithms would fail to separate the noun, adverb, adjective, and broken plural; According to the confusion matrix as shown in (Table 9):

Table 9. Confusion matrix details

TP	Predicted value reflects the fact that stemming algorithms will be able to stem verbs.
TN	In fact, stemming algorithms was not succeed in stemming words (noun, adverb, adjective, broken Plural) as expected.
FP	Stemming algorithms are able to stem verbs as expected, however they have not been successful in stemming.
FN	Noun, adjective, adverb, and broken plural were not able to be stemmed by the stemming algorithms. Despite this fact, the predicted value matched the reality.

5.0 EXPERIMENTS AND RESULTS

The data used in this study consists of 3050 words grouped into (noun, adverb, adjective, broken plural, verb) with their roots retrieved from Arabic WordNet and then stemmed using stemming algorithms. Among the 3050 words obtained, there were 2391 words grouped into nouns, adverbs, adjectives, and broken plurals, and 659 of those were verbs.

It was predicted that hybrid stemming algorithms could correctly stem 659 verbs, but the actual correctly stem was 625 verbs, whereas light10 stemming algorithms could correctly stem 659 verbs, but the real correctly stem was 45 verbs as True Positives (TP).

Even though it was anticipated that hybrid stemming algorithms could not stem 2391 words (Nouns, Adverbs, Adjectives, and Broken Plural) correctly, the actual correct stem was 2358 words, whereas it was anticipated that light10 stemming algorithms could not correctly stem 2391 words (Nouns, Adverbs, Adjectives, and Broken Plural), the actual correct stem was 397 words, as True Negatives (TN).

Though hybrid stemming algorithms were expected to successfully stem 659 verbs, the actual number of verbs not correctly stemmed was 34. While light10 algorithms were projected to correctly stem 659 verbs, the actual number of verbs not correctly stemmed was 614, as a result of False Positive (FP).

Despite the fact that it was predicted that hybrid stemming algorithms would be unable to correctly stem 2391 words (Nouns, Adverbs, Adjectives, and Broken Plural), the actual not correctly stem was 33 words, whereas it was predicted that light10 stemming algorithms would not be able to correctly stem 2391 words (Nouns, Adverbs, Adjectives, and Broken Plural), the actual not correctly stem was 1994 words, as False Negatives (FN). The confusion matrix for stemming algorithms is shown in (Table 10).

Table 10. confusion matrix for stemming algorithms

	(TP)	**(TN)**	**(FP)**	**(FN)**	**Total**
Light10	45	397	614	1994	3050
Ext Light10	45	301	614	2090	3050
ISRI	326	1444	333	947	3050
L10 & ISRI	322	1490	337	901	3050
Ext & ISRI	317	1409	342	982	3050
Hybrid	**625**	**2358**	**34**	**33**	**3050**

Accuracy is the ratio of correctly classified data items to the total number of data items. For example, the accuracy of hybrid stemming:

Accuracy = TP+TN/TP+FP+FN+TN= (625+2358)/ (625+2358+34+33) =**0.978**

accuracy provides a measure of how well the model performs on the full set of data.

Precision is the ratio of relevant items to the total number of irrelevant ones. For example, the precision of the hybrid stemming:

Precision = TP/TP+FP=625/ (625+34) = **0.9484**

Precision shows us how much we can rely on the model when it predicts a positive outcome for a given item.

Recall evaluates how well the model can predict the positive units in a dataset. For example, the recall for the hybrid stemming:

Recall = TP/TP+FN=625/ (625+33) = **0.9498**

Precise and Recall are weighted together to give you the F1 Score. It is for this reason that this score takes into consideration both the good as well as the negative aspects of the test results. Only when both accuracy and recall are good can the F1 score increase. It is a superior metric to accuracy since it is a harmonic mean of precision and recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision) = **0.9491**

It was shown that while comparing the performance of the different stemming algorithms, the hybrid stemming method had the greatest results, followed by Light10 & ISRI stemming (Table 11).

Table 11. Performance comparison of the algorithms

	Accuracy	Precision	Recall	F1-score
Light10	0.1449	0.0682	0.02206	0.03335
Ext Light10	0.1134	0.0682	0.02107	0.03221
ISRI	0.5803	0.4946	0.256	0.3374
Light10 & ISRI	0.594	0.4886	0.2632	0.3421
Ext & ISRI	0.5659	0.481	0.244	0.3237
Hybrid	**0.978**	**0.9484**	**0.9498**	**0.9491**

The suggested model was able to properly stem 625 out of 659 verbs and 2358 out of 2391 words. Verbs were incorrectly stemmed in 34 of 659 cases, while words were incorrectly stemmed in 33 cases out of 3 291 total cases. The hybrid stemming outperformed other stemming algorithms according to performance assessment criteria (Accuracy, Precision, Recall, and F1-score), indicating that the suggested model might be beneficial to the Arabic Automated Essay Method as seen in (Fig 6).

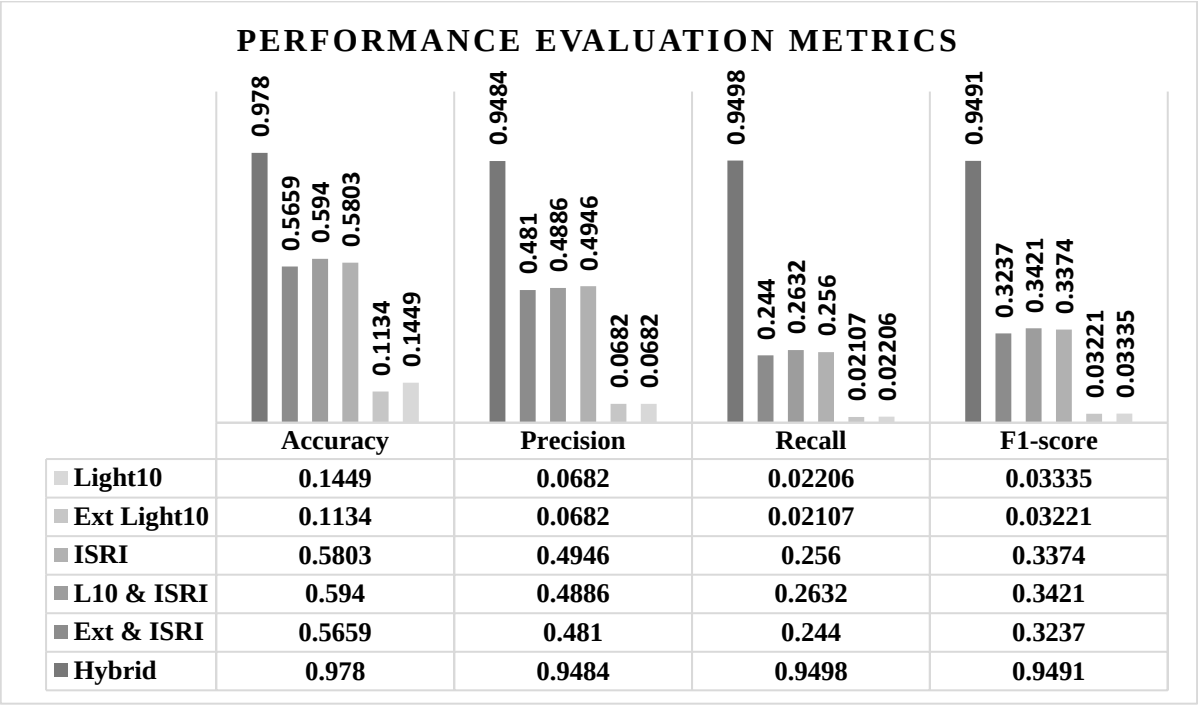

Fig 6. Performance Evaluation Metric

6.0 CONCLUSION

The goal of this study is to develop a new model that can evaluate students' answers to Arabic essay questions with a score that is near to that supplied by the teacher manually. The model depends on Arabic WordNet semantics to accomplish accuracy, prevent stemming weaknesses, evaluate the validity of extracted stems, fix a broken plural stem problem, search for a noun, adverb, or adjective in (AWN) to maintain without stems as well as the use of Arabic WordNet to look up synonyms for all answers in order to avoid penalizing students for not using the same words in

their answers compared to the reference answers. Finally, using semantic similarity through the cosine similarity method to determine the angle between students' answers and reference answers. There is potential for further work on the use of machine learning and neural network models to increase the precision of the Arabic essay score, as well as working on validating synonyms and vocabularies in WordNet.

REFERENCES

[1] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) System in Indian Context," *International Journal of Computer Applications*, pp. 72-77, 2010.

[2] K. Yaman, A. Swati, M. Debanjan, R. S. Rajiv, K. Ponnurangam and Z. Roger, "Get It Scored Using AutoSAS -- An Automated System for Scoring Short Answers," *arXiv*, pp. 01-08, 2020.

[3] A. Shehab, M. Faroun and M. Rashad, "An Automatic Arabic Essay Grading System based on Text Similarity Algorithms," *International Journal of Advanced Computer Science and Applications*, pp. 263-268, 2018.

[4] A. Alzahrani, A. Alzahrani, F. Alarfaj, K. Almohammadi and M. Alrashidi, "AN AUTOMATED SCORING APPROACH FOR ESSAY QUESTIONS," *The Eurasia Proceedings of Educational & Social Sciences (EPESS)*, pp. 232-236, 2014.

[5] S. Al Awaida, B. Al-Shargabi and T. Al-Rousan, "Automated Arabic Essays Grading System based on F-Score and Arabic WordNet," *Jordanian Journal of Computers and Information Technology*, pp. 170-180, 2019.

[6] N. Hockly, "Automated writing evaluation," *ELT Journal*, pp. 82-88, 2019.

[7] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," *Institute of Electrical and Electronics Engineers (IEEE)*, pp. 697-702, 2017.

[8] C. Ramineni and D. Williamson, "Understanding Mean Score Differences Between the e-rater® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE® General Test," *ETS Research Report Series*, pp. 1-31, 2018.

[9] M. Jarrar, "The Arabic Ontology -An Arabic Wordnet with Ontologically Clean Content," *Applied ontology*, vol. 16, pp. 01-22, 2021.

[10] A. E. Bolock, S. Abdennadher and C. Herbert, "An Ontology-Based Framework for Psychological Monitoring in Education During the COVID-19 Pandemic," *Frontiers in Psychology*, vol. 12, pp. 28-79, 2021.

[11] B. R. and W. S., "Building an Ontology Based on Folksonomy: An attempt to represent knowledge embedded in filmed materials," *Journal of Internet Technology and Secured Transactions (JITST)*, vol. 1, pp. 93-98, 2012.

[12] A. Chaleplioglou, S. Papavlasopoulos and M. Poulos, "Polysemy and Synonymy Detection in Ontology Engineering," *WSEAS Transactions on Information Sciences and Application*, vol. 17, pp. 117-123, 2020.

[13] F. Christiane, "WordNet: An online lexical database and some of its applications," *MIT Press*, pp. 131-134, 1998.

[14] H. Bradley and K. Grzegorz, "Synonymy = Translational Equivalence," *arXiv*, pp. 01-08, 2020.

[15] G. A. a. B. R. Miller, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography*, vol. 3, pp. 235-244, 1991.

[16] Y. Suhad, S. Venus, E. Islam and Z. Rached, "Enhancement of Arabic Text Classification Using Semantic Relations of Arabic WordNet," *Journal of Computer Science*, pp. 498-509, 2015.

[17] T. Dimitrova and V. Stefanova, "On Hidden Semantic Relations between Nouns in WordNet," *Global Wordnet Association*, pp. 54-63, 2019.

[18] A. Manal, S. Majdi and A. Mohammad, "Towards an automatic extraction of synonyms for Quranic Arabic WordNet," *International Journal of Speech Technology*, pp. 04-05, 2015.

[19] M. Pannu, A. James and R. Bird, "A Comparison of Information Retrieval Models," *Proceedings of WCCCE 2014: The 19th Western Canadian Conference on Computing Education - In-Cooperation with ACM SIGCSE*, pp. 01-07, 2014.

[20] B. Naol, "Information Retrieval System By Using Vector Space Model," *International Journal of Scientific & Technology Research*, pp. 1562-1563, 2019.

[21] M. Mohammad, S. A. Afag, E. Z. Mohammed, E. A. Rihab and E. Yasir, "Developing Two Different Novel Techniques for Arabic Text Stemming," *Intelligent Information Management*, pp. 01-23, 2019.

[22] A. Almazrua, M. Almazrua and H. Alkhalifa, "Comparative Analysis of Nine Arabic Stemmers on Microblog Information Retrieval," *2020 International Conference on Asian Language Processing (IALP)*, pp. 60-65, 2020.

[23] A. Kheireddine, O. Siham and H. Sayoud, "A novel robust Arabic light stemmer," *Journal of Experimental \& Theoretical Artificial Intelligence*, vol. 29, pp. 557-573, 2017.

[24] S. Motaz and A. W., "Arabic Morphological Tools for Text Mining," *International Conference on Electrical and Computer Systems (EECS'10),* 2010.

[25] M. Eldesouki, W. Arafa and K. Darwish, "Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective," *The Egyptian Computer Journal*, vol. 36, pp. 30-49, 2009.

[26] S. Khoja and S. Garside, "Stemming Arabic Text," *Computing Department,Lancaster University, Lancaster, U.K.,* 1999.

[27] S. Motaz and A. W., "Arabic Morphological Tools for Text Mining," *International Conference on Electrical and Computer Systems (EECS'10),* pp. 01-07, 2010.