Arabic Language Documents' Similarity and its Challenges (A Review)

Aledinat Lowai Saleh, Syed Abdullah Fadzli, Yousef El-Ebiary

Faculty of Informatics and Computing, UniSZA, Malaysia

ABSTRACT

Measuring the similarity of documents is of great importance and has influences on many areas and subjects such as Information Retrieval (IR), redrafting discovered, classified documents, and conversational agent programs. Recently, several scientific studies have been interested in implementing the similarity of documents in Arabic language applications. In this study, we review most of the techniques used in measuring the similarity between documents in Arabic; these techniques are classified into three types, mainly lexical, semantic, and finally hybrid. Hence, most of the techniques and methods adopted to measure the similarity between documents were reviewed to determine the most appropriate ones to measure the similarity in Arabic. However, the measurement of the percentage of similarity in documents in the Arabic language represents a major challenge faced by most of the researchers due to the nature of the Arabic language, such as complex morphological processes, ambiguity, and a general lack of resources.

Keywords: *Documents' Similarity; Standard Arabic Language, Arabic Corpus; datasets; Lexical-Based; Hybrid Similarity; Semantic-Based.*

I. INTRODUCTION

The similarity measurement in the documents among the word containts needed because it is the smallest component of a document or a sentence. However, the measurement of the similarity between words extracted from their context without any modification negatively affects the proportion of similarity, thereby resulting in not presenting accurate and good results. Therefore, the similarity between words must be measured based on their grammatical and semantic features. It is worth mentioning that the similarity measurement is used in many fields in natural language processing, such as information retrieval (IR) [1]-[2]-[3]-[4], document classification [5], essay scoring [6], machine translation [7], and Plagiarism Detection [8], As well as the similarity between Arabic documents in terms of the approximation of the distance. A new technique to reduce the distance between the Arabic documents by finding a unified form for a group of words was suggested [9]. The focus on the Arabic language is due to its significance as one of the most important languages used around the world. In other words, more than 300 million people speak this language spread over twenty countries in the world [10]. However, the implementation of the NLP tasks particularly in Conversation Agent (CA) presents a major challenge in the Arabic language. The Arabic as a language spoken widely has many features, thus posing challenges faced by the Conversation Agent (CA). The first challenge in Arabic is the presence of three categories namely classical, modern and colloquial categories. For instance, the classical Arabic (CA) is used widely in the Qur'an and it becomes more complicated compared to the modern Arabic (MA) language in terms of using grammar and vocabulary. In addition, the CA contains diacritical marks that facilitate and help pronounce the words and reveal them in its states of grammer. The second category is the Modern Arabic() In this classification, all the diacritics are deleted to make writing and reading tasks more faster and easy as well. That specific kind of language classification most widely spoken as an official language especially in Arab countries as it's their mother-tongue communication method that use in daily basis between people, education, and media. In addition, most computer and electronic research based on the Arabic language uses this category [11]. In the colloquial category, the third classification in Arabic, all components of language such as grammar and vocabulary are less developed and are rather simple compared to the modern Arabic. This category is the most common type of language used by people in their regular daily conversations and informal written messages [12]. However, when using the modern Arabic language, the Arab people make mistakes and confuse the modern language with the Colloquial category. In addition, each Arab country has its own Colloquial category, which is distinct from other countries, and therefore, the challenges faced by the Arabic conversation agents (CA) increase in terms of

recognizing the pronunciation of people and understanding the meaning. Arabic morphology is the other obstacle that make the the language is more complicated based on differences of morphological procedures also existing stacking phenomenon [13]. Whereby the Arabic words can contain affixes called suffixes and prefixes within different groups, thus resulting in a very complex formation and the production of a large number of words that fall within the same meaning [14]. Consequently, the presence of these features in Arabic increases the challenge for the Conversational Agent (CA) to determine the grammatical status of the spoken words of the user. The third challenge is capitalization. In other words, in Arabic language no uppercase letters of names such as country names, city names and people names are used. In Latin language generally, nouns begin with a capital letter [15]. Hence, it is difficult for the Conversation Agents (CA) in Arabic to determine these words types whether they are nouns or verbs, thereby making it difficult or impossible to detect these nouns. Short vowels and the existing Arabic documents are considered the fourth challenge. There is an urgent need for the pronunciation of words to eliminate ambiguity. However, the Arabic documents that use the modern language do not contain diacritical marks, thus creating a kind of obscurity.The fifth challenge is lack of resources. In their study, Khan et al. [16] describe plagiarism and theft by combining more than one method to measure the similarity of plagiarism detection in the Arabic documents. The proposed system contains two main elements; one element is concerned with the retrieval of documents, whereas the other element is concerned with the analysis of similarity. The component creates queries to retrieve documents from a certain suspicious document and uses the Google Search API applications to retrieve the source of the candidate of Web documents. The second component, a similarity analysis for each document, intends to identify the stolen and spoofed parts of the target document. The system was fully evaluated using a local set of data. Regarding the level of data retrieval of documents, the system gave results of more than 75% accuracy in terms of f-score. On the other hand, concerning the level of calculation of the detailed similarity, the results f-score is above 70%. Looking quickly at any free encyclopedia gives us the basics of Arabic language. Depending on the number of speakers, there are more than 27 sub-languages, and the modern Arabic language is classified as their main language. However, the Classical Arabic (CA) language is taught extensively and used effectively and officially throughout the Muslim world. The Modern Arabic (MA) language is derived from the classical Arabic language, which has witnessed the art of writing since the 6th century. The Arabic language is unique compared to other languages, and at the same time, the Islamic religion is sacred for its rituals since the seventh century until now. This is largely due to the presence of revelation, memorization, synthesis, and transmission of the Qur'an's Bible to Muslims.

II. LITERATURE REVIEW (APPROACHES)

Methods of documents' similarity as in the figure below (Fig. 1) were divided into three main approaches, namely the similarity indicator techniques which differ based on details: some technologies at the character-level, word-level, and group-level information.

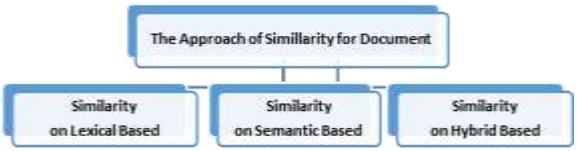

Figure 1. similarity approaches

A. Lexical-Based Similarity

Lexical similarities are combined to measure the degree of similarity between couple of words. Also the mechanism of similarity among two terms measured as lexical similarity between same word. The similarity between two terms is measured as the lexical similarity between similar word vectors extracted from a corpus as a second-word vector [17]. The lexical similarity is the same type of similarity that measures the degree of convergence between two dependent sequences based on character and term matching [18]. It is determined by the percentage of similarity between the two linguistic varieties by comparing a set of standardized lists of words and the calculation of those forms that show similarity in shape and meaning. Hence, the percentage of more than 85% that is usually variable in the presence of speech is likely to be the tone of the language. Unlike clarity, the lexical similarity is bi-directional or mutual [19]. Hence, the vector generally represents the text and all the words in it. A set of concepts and techniques such as Inverse Document Frequency (IDF) and the use of Term Frequencies (TF) and other metrics such as cosines are used to measure the distance between two parts of the text represented. The texts are often classified based on the technique used for measurement. The size of the vector depends on the number of words and terms in the documents, which are often huge. For text in medium-size vectors, it produces a scatter as well as short document vectors. As a result, the user can reduce the size of the vectors using trunk techniques or a stem technique to accelerate calculations to measure the similarity between languages. Sometimes, the user

needs to convert and translate the text or the document from one language to another through the use of a dictionary or the use of more sophisticated technology for translation and conversion and the use of vector and comparison (in the same language). In general, the use of a dictionary for translation is unsatisfactory because there is a single meaning for a set of multiple words. The question is which of these meanings will be chosen. The most sophisticated translation method that is somewhat satisfactory is expensive. Hence, the best option is to use traditional translation methods with all their advantages and disadvantages. In [20], radical algorithms were used to retrieve a more significant number of documents related to the user's query, and are intended to assess the impact of three different Arabic stem vectors on the performance of retrieving information in Arabic for the Arabic language. Light ten stemmer in term means the average precision that achieved the evaluation of three different stemmers techniques, thus ranking them, and choosing the best method in terms of performance. [21] N-gram similarity techniques that are used give a score of more than 80% accuracy. In one of the studies [22], 71.42% was detected as plagiarism. In addition, the percentage of partially stolen articles or documents was 28.85%. There are two types of lexical based similarity, as explained below:

a) Character-Based Similarity

Five algorithms were reviewed as shown in (Fig. 2) below, which demonstrate the methods used in the Arabic language and their efficiency in measuring similarities.

An in-depth learning approach was suggested to tackle

Figure 2. Character-Based Similarity

the task of Arab NER [23], whereby a neural network structure based on the LSTM bi-directional and CRF was provided with many excessive parameters commonly used to assess the impact on the overall performance of their system. The representation of an introductory is embedded to include words and descriptions based on the characters, and it also eliminates the need for any knowledge of the task or engineering features. The latest results in the standard group were obtained, whereby the ANERcorp corpus got a degree F1 of 90.6% [24]. Three works were carried out, including (NN) which is Neural Networks, that could enhance the (SVM) Support Vector Machine. These works classified the group as text areas in the pictures or the text. To improve the system performance, it applied the way of the group on the basis of a majority vote, whereby the three works resulted in outputs. Experiments were conducted using images of the basic truth, and the experimental data show that the results of the proposed method are an effective way.

- **Levenshtein distance similarity**

This indicates the minimum number of characters to insert, delete, and replace operations, which are necessary to convert s1 to s2. According to the Information Theory, the computer science and linguistics, Levinstein's distance technology, is a measure to calculate the difference between two sequences [25]. Unofficially, the distance between the Levenshtein between two words is the minimum number of edits with individual specifications (insert or delete or modify) needed to change the word to another. This is named after the world of the Soviet mathematician Vladimir [26]. A system based on the basis of stemming methods and processes editing Levenshtein to evaluate the students' exams online was provided [27]. Therefore, a hybrid method for an automatic essay scoring system (AES) for Arabic articles was suggested [28]. The system relies mainly on light and heavy stemming of words, whereby it relies only on a string-based algorithm (Levenshtein).

- **N-gram similarity**

N-gram is a Sub-Arrangement of N terms that gives text arrangement. N-gram Similarity algorithm was used to compare N-gram in two strings but for each character of them, the space also computed to divide a similar n-gram number by a maximum number of the n-gram [29]. The distance is calculated to divide a similar n-gram number by a maximum number of the n-gram. For natural language processing statistical, models N-grams were widely used. Like phonemes and sequencing using the distribution of n g in speech recognition, the analysis designed words so that each n-gram of n-words is made and used to determine the language [30] The TF-IDF model was designed for the documents examined using a research-based consensus matching algorithm, taking into account the lexical and grammatical changes. Subtle correlations between unique n-gram statements and their documents are then examined using the underlying Semantic Analysis. Next, a subset of the document and similarity measures are derived from the individual value analysis calculations. The performance of the suggested method was confirmed by experiments on different data sets, thus showing promising abilities in literal estimation

and some types of smart similarities. It also presents a comparison between indexing 3 grams and 4 grams in retrieving Arabic documents. The similarity between the query and the documents is calculated using a single term and a two-chapter query, based on the collection of Arabic language documents. The development and indexing of the large Arabic version using N-gram for inclusion in the Google N-gram viewer was discussed [31] to create a data set to start the process of digitization of the Arabic content and preparing for the inclusion in the Google N-gram viewer.

- **Longest Common Substring (LCS)**

LCS used in matching the patient's records, clinical setting, and summarizing texts, but it is not used to compare the titles. The calculation of the longest text string of all the strings of the existing text is taken and implemented and divided to the longest string character [32].

- **Hamming distance (HD)**

According to one of the studies [33], different string distances were examined, including the distance of (CS) which is Cosine Siimilarity that with L.Dist., H. Dist. And Hashes (ASCII-Based Segmentation over symbol Sub-Sequences to disclose the string matching of strings that were used. Similarity indicator differ based on details, such as some techniques at the character-level, word-level, and group-based. The usefulness of the methods that are assessed is to deal with the multiplication patterns for a simultaneous similarity.

- **Smith-Waterman**

The sizing was improved by adding the cost of the named affine gap [34] which offers extra costs for inclusion: an open gap, a penalty corresponding to the beginning of a series of mismatched characters and the extension of the gap, and penalty for continuation. In addition, the conversion between the symmetrical sounding characters gives a different weight more than the match / mismatch weights, for example, five-character matching units, three similar sounding units, and -3 units of matching.

b) Statement-Based Similarity

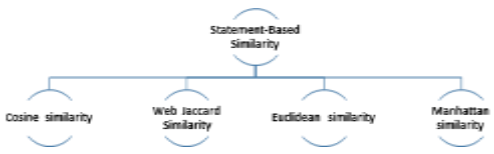

Figure 3. Statement-Based Similarity

- **Cosine similarity**

The usage of cosine method to measure the similarity ratio between two deffernt texts by represent each text as vector [35]. Each term in the text identifies the dimension in the Euclidean space. The value in this dimension matchs the recurrence of each word for t1 and t2 texts.

$$SIM\,(t_1,t_2)=\frac{\sum_{i=1}^{n} t_{1i}.t_{2i}}{\sqrt{\sum t_{1i}^{\,2}}\times\sqrt{\sum t_{2i}^{\,2}}} \quad (1)$$

The cosine similarity is used (1) for the Arabic language [21] to develop an application that measures the similarity between the descriptions course for the same subject. The results indicate that this method has shown better fulfillment compared with other methods like Dice's Similarity. Several methods and techniques for measuring the lexical similarity in the retrieval of Arab information (IR) were studied, concluded the cosine similarity as better measurement compared to other techniques and methods [36]. In another study [37], the Cosine similarity was used to study the classification of Arabic texts, whereby VSM , k-Nearest Neighbors, Naïve Bayes, classification tree, and Neural Network are the methods employed in this study [38]. The cosine similarity was also compared with the algoritm called K-NN from the LSA method and it runs automatic to record Arabic articles [38].

- **Web Jaccard Similarity**

This part determines the similarity ration between words using calculation for the frequency of the words in such group then convert and divide that in numbers of the words in a group [39].

- **Euclidean similarity**

Euclidean distance using the following equation:

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2}$$

The library contains both procedures and functions to calculate the similarity between sets of data. The best procedure is used when calculating the similarity between the small numbers of datasets. The procedures parallelize the computation and are therefore more convenient for the similarities measurement on large datasets.

- **Manhattan similarity**

MD is a method using to measure distance between two different points, a sum of absolute differences in Cartesian coordinates. On the other hand, It is aggregate of the difference among X and Y. also, it is assumed that thare are two different points which are Point A and Point B, in case that found MD amoung both of them then sould add the difference in the X axis

and the absoulute asix of Y. Mathematically, the measurement between two points according to (MD) found that the axes at the right angles. Moreover, P1 in (X1, Y1) with P2 in (X2, Y2). (MD) = | x1 - x2 | + | y1 - y2 |

B. Semantic-Based Similarity

The unique meaning of symbols can be exploited to evaluate the similarity of the parts of the text. It can take advantage of this form of web content like Wikipedia and collections of texts classified as an Explicit Semantic Association (ESA) [41] or the corpora wherein link among the text that required as is the case about Latent Semantic Association (LSA) [42] maybe overlapping search engine results. The individual could be depends on the knowledge infrastructure on the web. When using WordNet-based Lom Common Subsumer (LCS) and its multiple shapes, they are considered good examples. A study by [43] Makes use of wikipedia classes of articles as opposed to calculate the word with its identify. in any other observe, the semantic that means that represented and classifies the gadget in contrast to the semantic evaluation as a unique approach and makes use of a flat vector. illustration is in phrases of wikipedia articles [44]. In another study, The know-how-based totally representation of the textual content primarily based on the babelnet semantic community is used to create symbolize and calculates documentation units, connection among information units, and evaluate graphs to decide the similarity between documents. According to [46], in the semantic similarity to the aggregation of text advantages, amount of the textual content representation is decreased by way of 27% in comparison to the vector space model based totally at the stem. further, it's far decreased with the aid of 50% as compared to the conventional bag-of-phrases model. Therefore, a study by [47] proposed a new method based totally on an algorithm parallel to the semantic similarity measures the usage of mapreduce and wordnet after the translation level to discover relevant files within the face of the arabic query. numerical effects have been acquired and furnished the performance and overall performance of the approved approach.. Another study [48] studied the effects of each one to extract the appropriate attributes of the sentences without knowing the semantic structure and the standard Arabic grammar, for consideration there is a suggestion that creating new automated structure in the group to see the lack of Arab resources available to the public. The exams monitor that the LSTM model accomplished the very best price of semantic similarity and outperformed considerably the contemporary cutting-edge strategies. A fingerprint-based semantic similarity detection gadget, so-known as FPSS, was also provided to hit upon plagiarism within the arabic texts [49].

a) Corpus-Based Similarity

The similarity depends on the similarity of the document to the Corpus between the Phrases that use information from a massive corpus, and a massive group of various forms of textual content for comparison purposes, which adversely affects the research in natural language processing in a Corpus-based language [50]. Fourteen Arabic corpora classified according to the target language, whereby the classifications depend on text-domain, size, and document date [51]. Also based on [52], the texts of all elements are not categorized concerning the period to which they belong; therefore, limitations exist related to their usability, The issue of comparing languages utilized in extraordinary durations, and tracking how they broaden the arabic language. as an instance, maximum currently written texts are probably to be in the contemporary arabic language, while there may be texts amassed from the historic period inside the classical arabic language. Thus, the types difference in text languages will affect the efficiency of the similarity measurements based on the corpus system. However, The layout of many companies adopts uncertain standards. Regarding the other system for recording the Arabic articles on short questions, measuring the similarity depends on employing the similarities in string-primarily based and listing-based texts. to evaluate the machine, the authors used a hard and fast of 210 short solutions (10 college students answering 21 questions) in the technology consultation of the plenary. The students' answer was recorded against one typical answer for each question. The exceptional link became said in 0.82 while using n-gram with the word stop removed.

b) Knowledge-Based Similarity

It's a type of Semantic Based that measures the ratio of similarity among the words that extaracted by using semantic information systems [53]. Most of the semantic systems used Tagged WordNet [54], and the Natural Language (NL) Toolkit (NLTK) [55]. The WordNet is words gathered in one database that contains a synonym called synsets [56]. Within the following sections, we describe several measurements which have been determined to work properly at the hierarchy wordnet. these kinds of measurements are as input of a pair of principles and the return cost refers to the semantic association count on. the six techniques beneath have been selected based at the perceived overall performance in other language processing, whereby the computational efficiency is relatively high applications. Our assessment of the similarity is

conducted using the following measurements: Leacock & Chodorow and Lesk and Wu & Palmer and Resnik, and Lin and Jiang & Conrath. It seems each one the ones word by phrase with the aid of selecting any of the spouses worried that lead to any idea of the concepts which cause the pinnacle of the idea of similarity. We use an application called WordNet-based standards, as is available in WordNet: similarity package [57].

C. Hybrid-Based Similarity

The current scientific research has used hybrid methods to measure the similarity and has incorporated more techniques to obtain good results [58]. Eight semantic similarity methods were tested; six of these techniques are based on knowledge, whereas the other methods are based on corpus-based. The methods were used separately and they were also merged. The bottom line was that the combined and combined methods resulted in greater results and efficiency. In one of the studies, a hybrid method was used which combines lexical similarity methods to compute the similarity between documents using the user's expressions on the one hand and the text patterns on the other employing the Urdu language [59]. The Levenshtein algorithm was used to measure the similarity between two strings. The Kuhn-Munkres algorithm using to specify the variation between the order of words in the document [60]. In another study [61], the Hybrid style was used to gain access to the matrix weights that belong to every word, and this technique is used to overcome the challenge of the similarity of the series set words arranged in Urdu or other languages in conjunction with the arrangement in Arabic. In their study [62], the similarity technique based on lexical similarity was used to develop and evaluate the short-grading system to answer. The measurement test was conducted using three stages or steps. Hence, thirteen criteria were used on a lexical basis; seven of these criteria were based on the phrase, and the remaining six criteria were based on personal information. When using the character-based and particular necessities are among the ideas in preference to phrases, but they may be without problems modified to degree the similarity of the sentence–based, the methods have yielded better correlation results by appling (MP). The second stage measures the similarity using the corpus-based on the measurement of the similarity techniques [63]. DISCO1 (calculates the similarity of the first-order between two words depending on their collections) and DISCO2 (calculates the similarity of the second-order between two words depending on the distribution groups of similar words). In findings show that the performance of DISCO1 is performing well than DISCO2. Finaly, the dimension of similarity became assessed by all and sundry among the lexical measurement techniques and the corpus-primarily based. the great correlation cost was received from mixing n-gram similarity with DISCO1 techniques. By means of proposed data set, the authors in [62] focused on the part of speech (POS), which distinguishes the way to determine the grammatical feature of words within the similarity.

III. RELATED WORKS IN ARABIC DOCUMENTS OR TEXT SIMILARITY

The proposed methods and techniques are shown in Table 1 about calculating the similarity consistent with the scope of the text file, in addition to approach type that used (lexical, semantic, and hybrid). Table 1 also demonstrates the data used or the experimental approach in the experiment group and the results Obtained. Based on Table 1, the approaches similarity of document using the underlying analysis method of the Covariance is better than others, whereas for the semantic similarity in the sentence, the hybrid method of combining word combinations with the list feature supplies better results. Consequently, measuring the similarity using the semantic similarity achieved the best results, whereby the proposed methods were used to measure the similarity according to the scope.

Table 1. Comparison For Similarity Approaches

AUTHOR	YEAR	SCOPE	DATASET OR CORPUS	TECHNIQUE	RESULTS
A.Al Ramahi And Mustafa [21]	2012	Document	Two datasets: the first contains descriptions of 104 courses, and the second contains descriptions of 30 courses selected from the Jordanian universities.	**LB**	Matching techniques provide documents. N-gram accuracy level exceeds 80 %.
Soori et al. [22]	2013	Document	150 diverse documents, a number of articles distributed between the newspaper and source document, a total of 330	**LB**	The results showed that plagiarism of documents ratio exceeded 70%
H. Froud et al.[68]	2010	Document	The Contemporary Arabic Language Corpus comprises 12 categories	**LB**	Using Stemming to reduce the size of texts, JACCARD technique showed a better performance more than others
Awajan A. [46]	2016	Document	Various data sets such as news, sports ... etc	**SB**	Reduced the size of the documents representation by 27%
A. Hussein [30][65]	2016	Document	Answers of 30 students on a historical course.	**LB**	Outperformed the results on Checker Plagiarism X with N-gram technique

Alzahrani [69]	2016	Sentence	A set of data (pairs of sentences)	**SB**	Semantic similarity algorithm based on machine translation, got the highest relationship (r = 0.8657), whereas the algorithm produces the maximum similarity of the translation of medium closely 0.7206
Alameer et al. [70]	2017	Sentence	Network semantically-based stores Arabic keywords in the field of computer science.	**HB**	The hybrid technique showed satisfactory results and better accuracy
Ferrero et al. [71]	2017	Sentence	A collection of data containing a pair of sentences totaling 2412	**LB**	Bag-Of-Words link rate : 77.39 % weighting on the link rate : 73.75%
Almarsoomi et al. [72]	2013	Word	A data set of 70 randomly selected word pairs using high, medium, and low similarities.	**SB**	The AWS scale got an excellent rate Pearson correlation that0.894 cmpared with human average 0.893
Froud [73]	2012	Word	Couple of sets of data contain 506 documents from the Saudi Press Agency	**SB**	Stemming Light technique is better than the Stemming technique
Selamat et al. [74]	2008	document	Middle East, Al Jazeera News, and others.	**SB**	The results showed the average accurancy as follow: SOM 87% , GHSOM 93%
Lowai, S.E. & Fadzli [9]	2018	documents	BBC News, divided into multiple topics	**SB**	The results showed a preference to reduce the dimension between documents, which led to a reduction in the size of text representation by 42%
Zarzour et al.[75]	2018	document	Use two sets of varied data, consisting of ten topics	**SB**	Results showed a preference by reducing the distance between documents while retaining accuracy.

Where LB: Lexical based , SB : Semantic Based , HB: Hybrid Bsed.

IV. CONCLUSION

This study presents the techniques used to measure the similarity that is classified into three types and reviews the most important challenges to measure the similarity of documents in Arabic. The three classifications to measure the similarity in the Arabic language reviewed in this study are mostly based on lexical, semantic, and hybrid methods. The first type is the lexical similarity to measure the ratio of similarity between texts based on the letter or dependence on the sentence which has been reviewed by many algorithms, including what depends on the letter and some of which depends on the sentence. The second classification is the semantic similarity, which depends on the meaning of words or sentences to measure the proportion of similarity between parts of the texts; many companies use this approach to measure the similarity. The third classification proposed in this study is a hybrid similarity, which combines more than one technique or a technique to measure the similarity ratio. The Arabic language is a difficult language compared to the existing languages, and at the same time, it is an exciting language. In other words, it is considered one of the semitic languages, and in many countries, it's miles the professional language spoken by more than 380 M person [64]. Gulf area with North Africa countries are consider Arabic speakers officially. According to the classifications and other Arabic symbols, such as the (MSA), which is widely used in all official transactions and speeches as well as media [65]. With other Arabic type is language of (CA) or classical Arabic, in which the Qur'an was revealed, Mostly, all literary texts and poems use this format, whereby the Arab people use this language for over 14 centuries [66]. The third and final classification of the Arabic language is the general dialects, and this dialect varies according to the nature of the society and the nature of the geographical area [67]. The Arabic language contains 28 characters and the direction of writing from right to left. In addition, the Arabic language is a complex language in form and content, and the inherent characteristic of Arabic is ambiguity. At the moment, AWN lacks available resources compared to WordNet in English, which is found in abundance, and also lacks semantic relationships to information sources. Moreover, the challenge in Arabic is that most of the Arab companies do not cover all fields and all topics as each group of these companies specializes in a particular area rather than fully specialized in all areas due to the lack of sources of information. In conclusion, most of the Arab systems used the similarity method (cosine similarity), which showed the results efficiently compared with other lexical measurement methods. However, it is not possible to rely on the lexical similarity approach to measure the ratio of similarity between texts because of the presence of different and varied privileges that exist in this language most important thing is the morphology in Arabic. In addition, Arabic language lacks the resources that the Arab systems can use. Recently, the existence of hybrid methods in measuring the similarity of the Arabic language has become more efficient because it combines more techniques to improve the results of the measurement taking into account the work to increase the information sources of the Arabic language through research. Future studies can use the methods described in this study and how to combine them to get a better result.

ACKNOWLEDGEMENT

This research was supported by foundation from Universiti Sultan Zainal Abidin (UniSZA), therefore we thank our Universiti Sultan Zainal Abidin (UniSZA) that provided insight and expertise that greatly assisted the research.

REFERENCES

[1] Singh, J., Singh, P., & Chaba, Y. (2014). "*A Study of Similarity Functions Used in Textual Information Retrieval in Wide Area Networks*". International journal of computer science and information technologies, 5(6), 7880-7884.

[2] Noori, Z., Bandarl, Z., & Crockett, K. (2014). "*Arabic goal-oriented conversational agent based on pattern matching and knowledge trees*".

[3] El Mahdaouy, A., Gaussier, E., & El Alaoui, S. O. (2019). "*Should one use term proximity or multi-word terms for Arabic information retrieval?. Computer Speech & Language*", 58, 76-97.

[4] Mohamed, E., Elmougy, S., & Aref, M. (2019). "*Toward multi-lingual information retrieval system based on internet linguistic diversity measurement*". Ain Shams Engineering Journal.

[5] Ramesh, B., & Watzke, M. W. (2010). "*U.S. Patent No. 7,644,076*". Washington, DC: U.S. Patent and Trademark Office.

[6] Mezher, R., & Omar, N. (2016). "*A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring*". Journal of Applied Sciences, 16(5), 209.

[7] Meuschke, N., Stange, V., Schubotz, M., Karmer, M., & Gipp, B. (2019). "*Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations*". arXiv preprint arXiv:1906.11761.

[8] Hanumanthappa, M., Rashmi, S., & Reddy, M. V. (2015, January). "*Metrics for evaluating phonetics machine translation in Natural Language Processing through modified Edit Distance algorithm-A naïve approach*". In 2015 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-7). IEEE.

[9] Saleh, A. L., & Fadzli, S. A. (2018). "*A Proposed Method for Reducing the Dimension of Arabic Documents*". International Journal of Engineering & Technology, 7(3.28), 205-208.

[10] Versteegh, K. (2014). "*Arabic Language*". Edinburgh University Press. Versteegh, K. (2014). Arabic Language. Edinburgh University Press.

[11] Winder, R. B., & Ziyadeh, F. J. (2019). "*Introduction to Modern Arabic*" (Vol. 5500). Princeton University Press.

[12] Al-Saidat, E., & Al-Momani, I. (2010). "*Future markers in Modern Standard Arabic and Jordanian Arabic: a contrastive study*". European journal of social sciences, 12(3), 397-408.

[13] Kadri, Y., & Benyamina, A. (1992). "*A syntax semantic analyzer for the Arabic language*". Engineer thesis, University of Oran.

[14] Abuleil, S., Alsamara, K., & Evens, M. (2002, July). "*Acquisition system for Arabic noun morphology*". In Proceedings of the ACL-02 workshop on Computational Approaches to Semitic languages (pp. 1-8). Association for Computational Linguistics.

[15] Alanazi, S. (2017). "*A named entity recognition system applied to Arabic text in the medical domain (Doctoral dissertation*", Staffordshire University).

[16] Khan, I. H., Siddiqui, M. A., & Jambi, K. M. (2019). "*Towards Building an Arabic Plagiarism Detection System: Plagiarism Detection in Arabic*". International Journal of Information Retrieval Research (IJIRR), 9(3), 12-22.

[17] Elghannam, F. (2016). "*Automatic Measurement of Semantic Similarity among Arabic Short Texts*". Commun. Appl. Electron., 6(2), 16-21.

[18] Gomaa, W. H., & Fahmy, A. A. (2012). "*Short answer grading using string similarity and corpus-based similarity*". International Journal of Advanced Computer Science and Applications (IJACSA), 3(11).

[19] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. "*Ethnologue: Languages of the World*". Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

[20] Atwan, J., Mohd, M., Kanaan, G., & Bsoul, Q. (2014, December). "*Impact of stemmer on Arabic text retrieval. In Asia Information Retrieval Symposium*" (pp. 314-326). Springer, Cham.

[21] Al-Ramahi, M. A., & Mustafa, S. H. (2012). "*N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation*". Abhath AL-Yarmouk Basic Sci. & Eng, 21(1), 85-105.

[22] Soori, H., Prilepok, M., Platos, J., Berhan, E., & Snasel, V. (2014). "*Text similarity based on data compression in Arabic*". In AETA 2013: Recent Advances in Electrical Engineering and Related Sciences (pp. 211-220). Springer, Berlin, Heidelberg.

[23] El Bazi, I., & Laachfoubi, N. (2019). "*Arabic named entity recognition using deep learning approach*". International Journal of Electrical & Computer Engineering (2088-8708), 9(3).

[24] Oulladji, L., Feraoun, K., Batouche, M., & Abraham, A. (2018). "*Arabic text detection using ensemble machine learning*". International Journal of Hybrid Intelligent Systems, 14(4), 233-238.

[25] Beernaerts, J., Debever, E., Lenoir, M., De Baets, B., & Van de Weghe, N. (2019). "*A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length. Expert Systems with Applications*", 115, 373-385.

[26] Levenshtein, V. I. (1966, February). "*Binary codes capable of correcting deletions, insertions, and reversals*". In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).

[27] Al-Shalabi, E. F. (2016). "*An automated system for essay scoring of online exams in Arabic based on stemming techniques and Levenshtein edit operations*". arXiv preprint arXiv:1611.02815.

[28] Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). "*A hybrid automatic scoring system for Arabic essays*". Ai Communications, 27(2), 103-111.

[29] Barrón-Cedeno, A., Rosso, P., Agirre, E., & Labaka, G. (2010, August). "*Plagiarism detection across distant language pairs*". In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 37-45). Association for Computational Linguistics.

[30] Hussein, A. S. (2015, May). "*Arabic document similarity analysis using n-grams and singular value decomposition*". In 2015 IEEE 9th international conference on research challenges in information science (RCIS) (pp. 445-455). IEEE.

[31] Alsmadi, I., & Zarour, M. (2018). "*Google n-gram viewer does not include Arabic corpus! towards n-gram viewer for Arabic corpus*". Int. Arab J. Inf. Technol., 15(5), 785-794.

[32] Friedman, C., & Sideli, R. (1992). "*Tolerating spelling errors during patient validation*". Computers and Biomedical Research, 25(5), 486-509.

[33] Kaur, H., & Maini, R. (2019). "*Granularity-Based Assessment of Similarity Between Short Text Strings*". In Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems (pp. 91-107). Springer, Singapore.

[34] Gotoh, O. (1982). "*An improved algorithm for matching biological sequences*". Journal of molecular biology, 162(3), 705-708.

[35] Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004, March). "*The similarity between Euclidean and cosine angle distance for nearest neighbor queries*". In Proceedings of the 2004 ACM symposium on Applied computing (pp. 1232-1237). ACM.

[36] I Hajeer, I. (2012). "*Comparison on the effectiveness of different statistical similarity measures*". International Journal of Computer Applications, 53(8).

[37] Al-Anzi, F. S., & AbuZeina, D. (2017). "*Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing*". Journal of King Saud University-Computer and Information Sciences, 29(2), 189-195.

[38] Ewees, A. A., Eisa, M., & Refaat, M. M. (2014). "*Comparison of cosine similarity and k-NN for automated essays scoring*". Int. J. Adv. Res. Comput. Commun. Eng, 3.

[39] Jaccard, P. (1901). "*Étude comparative de la distribution florale dans une portion des Alpes et des Jura*". Bull Soc Vaudoise Sci Nat, 37, 547-579.

[40] Khafajeh, H., Yousef, N., & Kanaan, G. (2010, April). "*Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus*". In Proceedings the European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS), Abu Dhabi, UAE.

[41] Gabrilovich, E., & Markovitch, S. (2007, January). "*Computing semantic relatedness using Wikipedia-based explicit semantic analysis*". In IJcAI (Vol. 7, pp. 1606-1611).

[42] Rensch, C. R. (1992). "*Calculating lexical similarity. Windows on bilingualism*", 13-15.

[43] Strube, M., & Ponzetto, S. P. (2006, July). WikiRelate! "*Computing semantic relatedness using Wikipedia*". In AAAI(Vol. 6, pp. 1419-1424).

[44] Liberman, S., & Markovitch, S. (2009). "*Compact hierarchical explicit semantic representation*". In Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09) (pp. 36-38).

[45] FRANCO SALVADOR, M. A. R. C. (2017). "*A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning (Doctoral dissertation)*."

[46] Awajan, A. (2016). "*Semantic similarity-based approach for reducing Arabic texts dimensionality*". International Journal of Speech Technology, 19(2), 191-201.

[47] Amine, E. L., MADANI, Y., EL AYACHI, R., & ERRITALI, M. (2019). "*A new semantic similarity approach for improving the results of an Arabic search engine"*. Procedia Computer Science, 151, 1170-1175.

[48] Mahmoud, A., & Zrigui, M. (2019, June). "*Deep Neural Network Models for Paraphrased Text Classification in the Arabic Language*". In International Conference on Applications of Natural Language to Information Systems (pp. 3-16). Springer, Cham.

[49] Elhoseny, M., Zaher, M., Shehab, A., & Hassanien, A. E. (2017, December). "*FPSS: Fingerprint-based semantic similarity detection in big data environment*". In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 379-384). IEEE.

[50] Islam, A., & Inkpen, D. (2008). "*Semantic text similarity using corpus-based word similarity and string similarity*". ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 10.

[51] Alqahtani, M. M. A., & Atwell, E. (2015). "*A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus"*.

[52] Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: "*KACST Arabic corpus design and construction*". Language Resources and Evaluation, 49(3), 721-751.

[53] Budanitsky, A., & Hirst, G. (2001, June). "*Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*". In Workshop on WordNet and other lexical resources (Vol. 2, pp. 2-2).

[54] Miller, G. A. (1995). "*WordNet: a lexical database for English*". Communications of the ACM, 38(11), 39-41.

[55] Loper, E., & Bird, S. (2002). "*NLTK: the natural language toolkit"*. arXiv preprint cs/0205028.

[56] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). "*Introduction to WordNet: An on-line lexical database*". International journal of lexicography, 3(4), 235-244.

[57] Patwardhan, S., Banerjee, S., & Pedersen, T. (2003, February). "*Using measures of semantic relatedness for word sense disambiguation*". In International conference on intelligent text processing and computational linguistics (pp. 241-257). Springer, Berlin, Heidelberg.

[58] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). "*Corpus-based and knowledge-based measures of text semantic similarity*". In Aaai (Vol. 6, No. 2006, pp. 775-780).

[59] Kaleem, M., O'Shea, J. D., & Crockett, K. A. (2014, September). "*Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents*". In 2014 14th UK Workshop on Computational Intelligence (UKCI) (pp. 1-8). IEEE.

[60] Dasgupta, D., Hernandez, G., Garrett, D., Vejandla, P. K., Kaushal, A., Yerneni, R., & Simien, J. (2008, July). "*A comparison of multiobjective evolutionary algorithms with informed initialization and kuhn-munkres algorithm for the sailor assignment problem*". In Proceedings of the 10th annual conference companion on Genetic and evolutionary computation (pp. 2129-2134). ACM.

[61] Burkard, R. E., & Cela, E. (1999). "*Linear assignment problems and extensions*". In Handbook of combinatorial optimization(pp. 75-149). Springer, Boston, MA.

[62] Gomaa, W. H., & Fahmy, A. A. (2014). "*Automatic scoring for answers to Arabic test questions*". Computer Speech & Language, 28(4), 833-857.

[63] Kolb, P. (2008). "*Disco: A multilingual database of distributionally similar words*". Proceedings of KONVENS-2008, Berlin, 156.

[64] Farghaly, A., & Shaalan, K. (2009). "*Arabic natural language processing: Challenges and solutions*". ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 14.

[65] Zaidan, O. F., & Callison-Burch, C. (2014). "*Arabic dialect identification*". Computational Linguistics, 40(1), 171-202.

[66] Shoufan, A., & Alameri, S. (2015, July). "*Natural language processing for dialectical Arabic: A Survey*". In Proceedings of the Second Workshop on Arabic Natural Language Processing (pp. 36-48).

[67] Kanan, T., Ayoub, S., Saif, E., Kanaan, G., Chandrasekarar, P., & Fox, E. A. (2015). "*Extracting named entities using named entity recognizer and generating topics using latent Dirichlet allocation algorithm for Arabic news articles*". Department of Computer Science, Virginia Polytechnic Institute & State University.

[68] Froud, H., Benslimane, R., Lachkar, A., & Ouatik, S. A. (2010, September). "*Stemming and similarity measures for Arabic Documents Clustering*". In 2010 5th International Symposium On I/V Communications and Mobile Network (pp. 1-4). IEEE.

[69] Alzahrani, S. (2016). "*Cross-Language Semantic Similarity of Arabic-English Short Phrases and Sentences*". JCS, 12(1), 1-18.

[70] Alameer, A. Q. A. (2017). "*Finding the Similarity between Two Arabic Texts. Iraqi Journal of Science*", 58(1A), 152-162.

[71] Ferrero, J., Schwab, D., & Cherroun, H. (2017, October). "*Word embedding-based approaches for measuring semantic similarity of Arabic-English sentences*". In International Conference on Arabic Language Processing (pp. 19-33). Springer, Cham.

[72] Almarsoomi, F. A., OShea, J. D., Bandar, Z., & Crockett, K. (2013, October). "*AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity"*. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 504-509). IEEE.

[73] Froud, H., Lachkar, A., & Ouatik, S. A. (2012, October). "*Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model*". In 2012 Colloquium in Information Science and Technology(pp. 69-73). IEEE.

[74] Selamat, A., & Ismail, H. H. (2008, May). "*Finding English and translated Arabic documents similarities using ghsom*". In 2008 International Conference on Computer and Communication Engineering (pp. 460-465). IEEE.

[75] Zarzour, H., Al-Sharif, Z., Al-Ayyoub, M., & Jararweh, Y. (2018, April). "*A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques*". In 2018 9th International Conference on Information and Communication Systems (ICICS) (pp. 102-106). IEEE