Arabic Semantic Similarity Approach for Farmers' Complaints

Rehab Ahmed Farouk[1], Mohammed H. Khafagy[2], Mostafa Ali[3], Kamran Munir[4], Rasha M.Badry[5]
Department of Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum 63511, Egypt[1,2,3,5]
Department of Computer Science and Creative Technologies, University of the West of England[4]
BS16 1QY, Bristol, United Kingdom[4]

***Abstract*—Semantic similarity is applied for many areas in Natural Language Processing, such as information retrieval, text classification, plagiarism detection, and others. Many researchers used semantic similarity for English texts, but few used for Arabic due to the ambiguity of Arabic concepts in both sense and morphology. Therefore, the first contribution in this paper is developing a semantic similarity approach between Arabic sentences. Nowadays, the world faces a global problem of coronavirus disease. In light of these circumstances and distancing's imposition, it is difficult for farmers to physically communicate with agricultural experts to provide advice and find suitable solutions for their agricultural complaints. In addition, traditional practices still are used by most farmers. Thus, our second contribution is helping the farmers solve their Arabic agricultural complaints using our proposed approach. The Latent Semantic Analysis approach is applied to retrieve the most problem-related semantic to a farmer's complaint and find the related solution for the farmer. Two methods are used in this approach as a weighting schema for data representation are Term Frequency and Term Frequency-Inverse Document Frequency. The proposed model has also classified the big agricultural dataset and the submitted farmer complaint according to the crop type using MapReduce Support Vector Machine to improve the performance of semantic similarity results. The proposed approach's performance with Term Frequency-Inverse Document Frequency-based Latent Semantic Analysis achieved better than its counterparts with an F-measure of 86.7%.**

Keywords*—Semantic similarity; latent semantic analysis; big data; MapReduce SVM; COVID-19; agriculture farmer's complaint

I. Introduction

The semantic analysis field has an essential role in the research related to text analytics. Measuring the semantic similarity between sentences is a long-standing problem in the Natural Language Processing (NLP) field [1], [2]. With the growth of text data over time, NLP became essential to be worthy of attention for Artificial Intelligence (AI) experts [3], [4]. Semantic similarity is used for several fields in NLP like information retrieval, text summarization, plagiarism detection, question answering, document clustering, text classification, machine translation, and others [5], [6]. It is defined as determining whether two concepts are similar in meaning or not [7]. The concepts are words, sentences, or paragraphs. Each concept takes a score. When the concept has a high score refers to high similarity or semantic equivalence to another concept [8]. Concepts can have two ways to be similar that are either lexically or semantically. Concepts are lexical similarly if words have similar character sequences and are performed using a String-based algorithm. Concepts are semantic similarly if words depend on information acquired from massive corpora, even if they have a different lexical structure. Semantic similarity can be done by a corpus-based algorithm or knowledge-based algorithm [9], [10]. Several research works of semantic similarity have been developed for English sentences. On the other side, few research works have been used for the Arabic language because Arabic is considered a complex morphological language [11]. However, the Arabic language considers the fifth most spoken language in the world. Also, it participates in the most critical foreign languages with over 300 million speakers and a wide range of functionalities that no other language can have [12]. Therefore, this paper will apply a semantic similarity approach to the Arabic dataset.

Currently, the world faces a huge disaster that threatens the world is the global Coronavirus disease (COVID-19) pandemic. COVID-19 causes destructive economic, political, and social crises in each country. All fields have been affected by the global Coronavirus, especially the agriculture field. In our life, Agriculture plays a critical role in the entire life of the economy. It can be one source of Livelihood, contributes to national revenue, the supply of food, and marketable surplus. Moreover, it provides job opportunities to a huge percentage of the population and supplies the country with an important portion from its foreign exchange through agriculture exports. Therefore, due to COVID-19 that compounds pre-existing vulnerabilities in the field of agriculture in Egypt. Initial analyzes of this epidemic have shown disrupting access to agricultural inputs, including employment, extension, advising services, and producing markets for farmers. Most significant countries became deserted that people stay indoors, either by choice or by the government, to reduce the spread of this pandemic. Because of this, the curfew and distancing imposed by COVID-19 cause many problems for farmers. So, it is difficult for farmers to communicate and interact with agricultural experts to present their complaints and find suitable solutions. Therefore, it is essential to find an appropriate way to help in solving the farmers' complaints. Agriculture Research Center (ARC) and Virtual Extension and Research Communication Network (VERCON) [13] in Egypt provide a large group of farmers' complaints and their solutions in Arabic deployed on a public cloud. The agricultural experts have resolved these complaints. Thus, under these difficult circumstances of the spread of the COVID-19, this paper aims to develop an approach for farmers to help, support, and find the

most suitable solution for their agricultural complaints. The proposed approach is based on latent semantic analysis (LSA) to measure the semantic similarity score between Arabic farmers' complaints and the Arabic agricultural dataset, further retrieving the related solution to the farmer. As an example, the farmer's complaint is " مهاجمة دوده ورق القطن لحقول البرسيم فما هي المقاومه؟" and its English equivalent is "Cotton leaf worm attack alfalfa fields, what is the resistance?". After applying the proposed model, the recommended solution is " تقاوم دودة ورق القطن فى البرسيم باستخدام احد المبيدات الموصى بها مثل لانيت 90% بمعدل 300 جم/ف ثم الرى بالسولار بمعدل 200 لتر للفدان" and its English equivalent is "The cotton leafworm is resisted in alfalfa by using one of the recommended pesticides such as Lannet 90% at a rate of 300 g/f, then irrigation with diesel at a rate of 200 liters per feddan."

To improve the performance of the semantic similarity approach, we used the classification. Text Classification (TC) is an active research field and an essential in information retrieval technology [14]. It aims to classify text documents into one or more predefined categories. TC is applied in many applications like sentiment analysis, sentence classification, and document classification [15], [16]. TC can use many methods such as Decision Trees, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes (NB), K-Nearest Neighbor (KNN), etc. [17], [18].

SVM is an extremely powerful classifier in the machine learning field and is widely used in text classification [19]. However, it is fast and easy to implement. Therefore, we applied SVM on the agricultural dataset to classify Arabic complaints into crops. But, SVM didn't achieve better accuracy results. Thus, to improve performance in classifying our Arabic agricultural dataset, we resort to a parallel programming model like MapReduce. So, this paper applied the classification by MapReduce SVM using Hadoop to classify the Arabic agricultural dataset according to its crop type.

Most of the previous works applied Arabic semantic similarity to small datasets and achieved low accuracy results. Moreover, fewer of them tested on agriculture datasets and didn't use the classification.

Thus, the main objectives in this paper are as follow:

1) Applying the proposed model on a big agricultural dataset with real complaints facing Egypt's farmers.

2) The proposed model can help farmers, especially in the circumstances of COVID-19, by providing advice and finding appropriate solutions for their complaints to enhance agriculture productivity.

3) Developing a semantic similarity model between Arabic complaints and obtaining better results.

4) Using a parallel programming model like MapReduce based on SVM to classify the agricultural dataset and improve the performance of a semantic similarity model.

5) Testing and validating the proposed model performance by implementation multiple experiments and applying previous models on our Arabic agricultural dataset.

The remaining parts of the paper will be structured so that Section II presents related work; Section III includes materials and methods; Section IV covers discussion; Section V shows the experimental results. Finally, in Section VI, the conclusion is produced.

II. Related Work

This section will introduce related work about Arabic text classification and Arabic semantic similarity. Mostafa et al. [20] Proposed two models to classify the Arabic farmers' complaints based on different diseases that may affect crops. The Arabic complaint is classified into its respective crop and a specific disease in the first model. The second model could classify the complaint directly into diseases. Each preprocessed complaint is represented into a binary vector form using the vector space model by helping the crop lexicon. Experiments are conducted on the dataset by changing the training percentage with many trials using SVM and KNN classifiers. The results are shown that the proposed model is performed on par with the human expert and can be applicable for real-time operations. Moreover, Raed Al-khurayji and Ahmed Sameh [21] presented an approach that depends on a Kernel Naive Bayes classifier to solve the non-linearity problem of Arabic text classification. First, they applied preprocessing techniques on Arabic datasets like tokenization, stop word removal, and light stemmer. Then, they used the TF-IDF technique on Arabic words for feature extraction to convert them into the vector space. Experimental results are shown that the proposed approach achieved good accuracy and time compared with other classifiers. While Abutiheen et al. [22] proposed the Master-Slaves (MST) technique to classify Arabic texts. The proposed approach consisted of two phases. In the first phase, Arabic corpus text files are collected. These text files are classified manually into five categories. In the second phase, four classifiers were implemented on the Arabic collected corpus. The four classifiers were NB, KNN, Multinomial logistic regression, and maximum weight. NB classifier was applied as Master and the others as Slaves. The slave classifiers' results were used to change the NB classifier probability (Master). Each document in a corpus was represented as a vector of weights. The results of the MST have achieved a good improvement in accuracy compared with the other techniques.

Schwab et al. [23] presented a technique that depends on word embedding for measuring semantic relations among Arabic sentences. This technique relies on the characteristic of semantic words in the model of word embedding. This technique has applied three methods: no weighting method, Inverse Document Frequency (IDF) weighting method, and part-of-speech (POS) weighting method. No weighting method is used by summing the word vectors of each sentence. To improve the results, use the IDF weighting method to calculate IDF weight for each word and add the word vectors with IDF weights for each sentence. Also, use the POS tagging method that supposes weight for each POS and calculate POS for each word, then for each sentence, sum the words vectors with POS weights. This technique is evaluated the results on a small dataset. This paper demonstrated how weighing IDF and POS tagging supports highly descriptive word determination in any sentence. The performance of both IDF and POS weighting techniques achieved better results. While Amine et al. [24] proposed an Arabic search engine method depending on the MapReduce method. This method is used for finding semantic similarity among an Arabic query and the large corpus of

existing documents in the Hadoop Distributed File System (HDFS). It is also used to obtain the most relevant documents. It uses two measures in MapReduce: Wu and Palmer (WP) measure and Learock and Chodorow (LC) measure. The results appeared that WP and LC obtained better results than the existing approaches of semantic similarity. Mahmoud et al. [25] suggested a semantic similarity technique in paraphrase identification for Arabic. This technique depends on the combination of various NLP like the TF-IDF technique and the word2vec algorithm.TF-IDF technique is used to ease the identifying of highly descriptive terms in each sentence. The word2vec algorithm is used for representations of distributed word vectors. Also, word2vec can minimize computation complexity and optimize the likelihood of word prediction in producing a model of sentence vector. This paper applied the similarity using various comparison metrics, like Cosine Similarity and Distance Euclidean. Finally, the proposed technique was tested on the Open-Source Arabic Corpus OSAC and achieved a reasonable rate. In [26] used a semantically reduced dimensional vector to represent high dimensional Arabic text. It has been accomplished by extending the standard vector space model (VSM) to enhance the representation of text that utilizes Linguistic and semantic properties from Arabic WordNet and Name Entities' gazetteers. If synonyms and similar terms obtain from the same root in clusters, the vector size reduces, and the shorter NE represents the chosen cluster members. The word similarity is also determined using distributional similarity to collect similar terms into clusters. Results demonstrated the size, form of the analysis windows, and the text's nature and category based on how much it reduced. In [27] suggested a method for finding the semantic similarity among two Arabic texts. This approach used hybrid similarity measures that are edge-counting semantic approach, cosine similarity, and N-gram similarity. The edge-counting semantic approach determined the value of a threshold. If the first approach's similarity result was lower than the threshold value, then cosine similarity is applied. Moreover, if the cosine similarity value compared with the predefined threshold was not appropriate, use N-gram similarity. This hybrid approach addresses problems of writing mistakes like repetitive, incomplete, and substituted characters. The hybrid similarity results outweigh the results of any of the three measures that have been used individually.

III. Material and Methods

The proposed model aims to measure the semantic similarity's score between the current farmer's complaint and the available historical agricultural problems to provide an adequate solution to the farmer's complaint. Our proposed model was applied and tested on the agriculture problems dataset and their solutions. ARC and VERCON. It contains complaints of various causes, such as harmful weeds, fungal diseases, and other diseases that affect plants and their solutions. It also includes complaints belongs to 31 governorates and their directorates. It contains more than 10,000 complaints. The complaints are written in the Arabic language in an unstructured form and not well-formatted. These complaints related to different crops such as "rice, okra, wheat, corn, cotton, beans, etc...". It is also associated with different categories, which are "Administrative, Productivity, Marketing, and Environmental". Due to the variety of crops in the agriculture dataset, the proposed model applied a classification method to classify the farmer complaints dataset according to the crop type.

Table I shows some examples of the dataset's complaints related to different crops and their English translation.

TABLE I. Examples for Arabic Farmers' Complaints and their English Equivalent

Complaints in the English Language	Complaints in the Arabic Language
Yellow spots on the leaves of onion plants.	وجود بقع صفراء على اوراق نباتات البصل.
The presence of whiteflies strongly in cotton.	وجود ذبابة بيضاء بشدة فى القطن.
The wheat was infested with aphids.	اصابة القمح بحشرة المن.
The presence of spots on the upper surface of okra leaves with the appearance of a spider thread on its lower surface	وجود بقعة على السطح العلوى من اوراق الباميا مع ظهور خيط عنكبوتي على سطحها السفلى.
The lack of water in the Arimon canal for more than a month exposes the existing winter crops to fallow, such as clover.	عدم وجود مياة بترعة اريمون منذ اكثر من شهر مما يعرض المحاصيل الشتوية القائمة للبوار مثل البرسيم.

The proposed model consists of four phases, as shown in Fig. 1 Preprocessing, MapReduce SVM classification, and Latent Semantic Analysis. The last phase is the ranking and selection to choose the most semantically relevant solution to the farmer's complaint. The next sections explain in detail the phases of the proposed model.

A. Preprocessing Phase

The preprocessing phase is an essential step for Natural Language Processing (NLP) tasks. It transforms input text into a more desired form for performing better for further steps [28]. Unfortunately, the complaints' meaning is difficult to understand and interpret since farmers typically write complaints without following the Arabic grammar rules.

Data preprocessing includes four operations: tokenization, stop word removal, complaints auto-correction, normalization, and lemmatization, as shown in Fig. 1.

- Tokenization: It is a method for breaking texts into tokens. Words are separated from their neighboring words by blanks such as white space, periods, commas, semicolons, and quotations [29]. For example, The Arabic complaint is "وجود بقع صفراء لها مظهر مسحوقى فى صفوف طوليه على ورقه القمح".

After applying the tokenization, the Arabic complaint is:

And its equivalent English is: "The presence of yellow spots that have a powdery appearance in longitudinal rows on the wheat leaf."

"وجود", "بقع", "صفراء", " لها", "مظهر", "مسحوقى", "فى", "صفوف", "طوليه" , "على" , "ورقه" , "القمح".

- Stop Word Removal: The most popular undesirable term is either a punctuation mark or a stop word. Therefore, they are eliminated from complaints since they do not have any meaning or indications about the content. We used an online Arabic stop words list for elimination [30]. Examples of these unimportant words in the Arabic language such as:

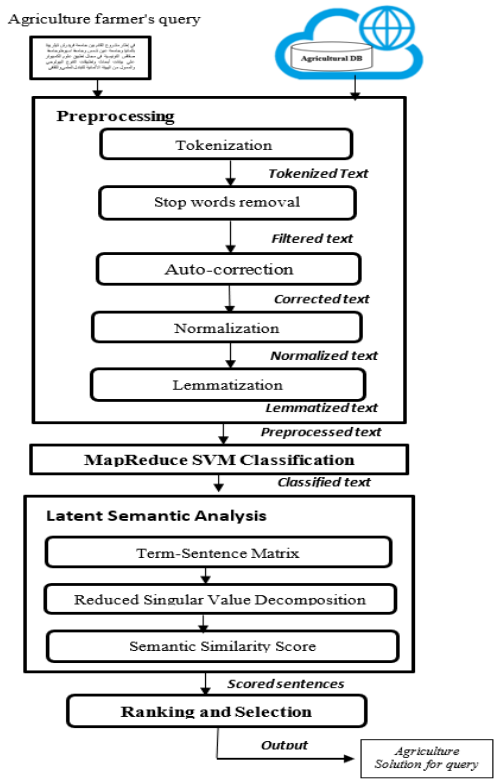

Fig. 1. The Proposed Architecture.

(كلما, اولا, حول, اين , طالما, التى, من, الى, على, فوق, تحت , الخ.

And in the English language are (Whenever, first, about, where, as long as, which, from, to, on, above, below, etc.)

In addition, eliminating all symbols such as: (@, #, &, %, and *).

- Auto-correction: The farmers may write their complaints with spelling errors. For example, the crop name of tomatoes "طماطم" may be incorrectly written in slag way as "اوطه" , the crop name of corn "ذره" may be incorrectly written as "زره" and the name of rice crop "أرز" may be incorrectly written as "رز" [20]. Therefore, we use auto-correction to solve these problems by substituting the incorrect word with the correct one.

- Normalization: Text normalization is transforming the input text into a canonical (standard) form. It is critical for noisy texts like comments on social media and text messages that are popular in abbreviations and misspellings. Moreover, it concentrates on removing the inconsistent language variations. For example, In English, the word "croooop" can be transformed into its canonical form "crop" and also in Arabic like "محصووووول" is normalized into "محصول" [31], [32]. So, we applied some methods for normalization such as the letters "إ ، أ ، آ" will convert into one form "ا." Also, the letter "ة" will replace by "ه," the letter "ي" will convert to "ى." Also, remove the diacritics from the words such as "أَشْجَارٌ" will convert to "أشجار". Furthermore, if there are more spaces between words, then we removed them. We also convert numbers into words such as "15" will convert to "خمسه عشر."

- Lemmatization: It is an essential step in the preprocessing phase and a significant component for many applications of natural language processing. It is an operation to find the base form for a word. For example, in Arabic words like (ثمار) has the root "ثمر." Also, in English, "fruits" has the root "fruit." We used an online Farasa lemmatization [33].

B. MapReduce SVM Classification Phase

Text Classification is the process of distributing each document to its labeled class [34]. MapReduce is a popular programming model developed by Google. It can process massive datasets in a parallel manner and achieves a high performance [35], [36]. The main idea of MapReduce comes from the divide and conquer algorithms which are used to divide a large problem into smaller subproblems. Therefore, we apply MapReduce SVM to classify big data preprocessed agricultural complaints according to their crop type, such as rice, wheat, okra, etc. MapReduce SVM uses the Hadoop framework to share the classification between many machines using HDFS to store the preprocessed agricultural complaints to classify and store the classification result. MapReduce model is divided into two tasks which are Map and Reduce [37]. It divides the dataset into smaller chunks and then assigns each chunk to a single map task. Map tasks' number is equal to the number of data chunks. Thus, each map task processes each data chunk in a parallel way. The model shuffles and sorts the Map outputs and transfers them to the Reduce tasks. The Reduce task is a summarization step that all associated records are processed together by a single entity. The Map and Reduce tasks are mathematically represented in (1) and (2), respectively [38].

$$Map: \quad (K_1,V_1) \longrightarrow [(K_2,V_2)] \tag{1}$$

$$Reduce: \quad (K_2,[V_2]) \longrightarrow [(K_3,V_3)] \tag{2}$$

The (K_1, V_1), (K_2, V_2), and (K_3, V_3) represent the key-value pairs for map and reduce tasks.

After this phase, each complaint is classified according to its crop type. Table II shows the number of Arabic agricultural complaints in each crop after applying the MapReduce SVM model.

C. Latent Semantic Analysis Phase

In this phase, LSA is applied to measure the semantic similarity among the agriculture dataset and the farmer complaint. It is a technique used for representing documents as a vector. It helps to find the similarity between agricultural complaints by calculating the distance between vectors.

There are three main steps for the LSA-based algorithm:

- Creating the input matrix (Term-Sentence matrix)
- Applying reduced singular value decomposition (RSVD) on the created matrix
- Calculating the semantic similarity score between the farmer's complaint and complaints document.

TABLE II. THE NUMBER OF ARABIC AGRICULTURAL COMPLAINTS IN EACH CROP

Crop Name (English)	Crop Name (Arabic)	Number of Arabic Complaints	Crop Name (English)	Crop Name (Arabic)	Number of Arabic Complaints
Wheat	قمح	890	Apples	التفاح	837
Rice	ارز	754	Orange	البرتقال	458
Cotton	قطن	735	Lettuce	الخس	579
Tomatoes	طماطم	692	Cabbage	الكرنب	699
Beans	الفاصوليا	572	Garlic	الثوم	276
Potato	البطاطس	379	Guava	الجوافة	497
Clover	البرسيم	437	Okra	الباميا	238
Peach	الخوخ	288	Banana	الموز	798
Onions	البصل	583	Peas	البسلة	436
Apricot	المشمش	400	Watermelon	البطيخ	972
Lentils	العدس	389	Mandarin	اليوسفى	697
Grapes	العنب	457	Cowpea	اللوبيا	793
Eggplant	الباذنجان	389			

These steps will be explained in detail in the following sections.

1) Term-sentence matrix: In this phase, an input matrix is created for the farmer's complaint query and classified complaints document. Each row in the matrix represents the word or term in the farmer's complaint or classified complaints document [39]. Each column represents the complaints. The cell value is the result of the intersection between term and complaint. There are two methods used as a weighting schema for data presentation for filling the cell values: Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF).

In TF-based LSA, the cells are filled with the term frequency (TF_i) of terms in the complaint statement (C_j) as in (3).

$$W(t_{ij}) = tf_{ij} \quad (3)$$

Where $W(t_{ij})$ is the weight of a term (i) in each complaint statement (j) and tf_{ij} is the frequency of a term (i) in each complaint statement (j).

TF_IDF-based LSA, the cells are filled with the weight of (TF_IDF) of the term (i) in complaint statement (C_j) as shown in (4) and (5).

$$TF\_IDF_{ij} = TF_{ij} * IDF_{ij} \quad (4)$$

Where TF_IDF_{ij}. TF is the frequency of a term (i) in each complaint statement (j), and IDF reflects the importance of term among all sentences

$$IDF_{ij} = log \frac{N}{ComplaintFreq(f)} \quad (5)$$

Where N represents the number of complaints in the collection, and $ComplaintFreq\ (f)$ is the number of complaints containing the term.

2) Reduced singular value decomposition: Singular value decomposition (SVD) is an algebraic method that plays an essential role in text mining and natural language processing. SVD is used to improve the term sentence matrix, remove noise, and determine the relationships between terms and complaints statements [40]. SVD decomposes the Term Sentence Matrix into three matrices that detect all the important properties and features of the matrices.

Equation (6) shows the SVD decomposition of the m × n matrix.

$$SVD = USV^T \quad (6)$$

Where U is the m-dimensional matrix, V is the n-dimensional matrix, and S is the diagonal matrix.

Moreover, RSVD is applied to improve and enhance the performance of SVD and reduce the matrix dimensionality.

3) Semantic similarity score: After applying RSVD, RSVD results are used to calculate semantic similarity between the farmer query and classified complaints document. The semantic score is calculated using the most common similarity method, which is the cosine similarity. Equation (7) represents the calculation of cosine similarity.

$$Cos\ similarity\ (A,B) = \frac{A.B}{||A||*||B||} \quad (7)$$

Where $Cos\ similarity\ (A,B)$ is the similarity score between the farmer query and complaints document, A is the weight of the term in the query, and B is the weight of the term in the complaint statement.

D. Ranking And Selection Phase

In this phase, rank the complaints according to the semantic score, then select the complaint of the highest score. Finally, retrieve the solution of the complaint with the highest score to the farmer query.

IV. DISCUSSION

F-measure is used to evaluate the performance for the proposed classification approach and semantic similarity approach.

A. Classification Evaluation

The performance of the MapReduce SVM classifier using Hadoop is evaluated. Also, we compared our classification results with the previous classification works as in Mostafa et al. [20] and Mohammad et al. [41]. Authors in [20] applied two classifiers that are SVM and KNN, on the same agricultural dataset to classify agricultural complaints into crops.

Moreover, authors in [41] used two classifiers that are the Naive Bayes algorithm (NB) and the Hybrid Naive Bayes with Multilayer Perceptron network (NB-MLP), to classify the dataset into positive or negative sentiment. Therefore, we applied NB and NB-MLP algorithms to our agricultural dataset to classify complaints into crops.

TABLE III. EVALUATION OF MAPREDUCE SVM CLASSIFIERS COMPARED WITH PREVIOUS MODELS

	NB			NB-MLP			SVM			KNN			MapReduce SVM		
Class	***Precision***	***Recall***	***F-Measure***	***Precision***	***Recall***	***F-Measure***	***Precision***	***Recall***	***F-Measure***	***Precision***	***Recall***	***F-Measure***	***Precision***	***Recall***	***F-Measure***
Wheat	91.71	94.54	93.10	92.54	95.91	94.19	93.68	97.49	95.55	93.96	97.58	95.74	94.2	98.4	96.3
Rice	90.83	92.32	91.57	94.72	95.73	95.22	89.13	97.27	93.02	96.85	89.71	93.14	97.4	98.3	97.8
Cotton	92.52	94.84	93.67	93.03	95.45	94.22	93.06	95.32	94.18	83.87	93	88.2	94.7	97.4	96
Beans	93.77	95.61	94.68	94.18	96.48	95.32	96.82	75.43	84.8	95.24	97.62	96.42	97.5	98.8	98.1

Table III shows a comparison between our MapReduce SVM evaluation results and previous works that used NB, NB-MLP, SVM and KNN classifiers. We evaluated the results on four crops, such as wheat, rice, cotton, and beans, familiar with authors in [20].

As a conclusion, the evaluation results of MapReduce SVM achieved better results than previous classifiers of NB, NB-MLP, SVM and KNN.

B. Semantic Similarity Evaluation

The proposed semantic similarity approach using TF-based LSA and TF_IDF-based LSA are tested and evaluated. And finally, the results of the proposed approach are compared with the previous models. The tests applied on twenty-five crops which are Okra, Mandarin, Watermelon, Wheat, Rice, Cotton, Beans, Tomatoes, Potato, Peach, Apricot, Lentils, Onions, Clover, Apples, Eggplant, Grapes, Orange, Banana, Guava, Peas, Cowpea, Cabbage, Garlic, and Lettuce.

TABLE IV. THE NUMBER OF ARABIC QUERIES IN EACH CROP

Crop Name (English)	Crop Name (Arabic)	Number of Arabic Complaints	Number of Arabic Complaint Queries
Wheat	قمح	890	222
Rice	ارز	754	188
Cotton	قطن	735	183
Tomatoes	طماطم	692	173
Beans	الفاصوليا	572	143
Potato	البطاطس	379	94
Clover	البرسيم	437	109
Peach	الخوخ	288	72
Onions	البصل	583	145
Apricot	المشمش	400	100
Lentils	العدس	389	97
Grapes	العنب	457	114
Eggplant	الباذنجان	389	97
Apples	التفاح	837	209
Orange	البرتقال	458	114
Lettuce	الخس	579	144
Cabbage	الكرنب	699	174
Garlic	الثوم	276	69
Guava	الجوافة	497	124
Okra	الباميا	238	59
Banana	الموز	798	199
Peas	البسلة	436	109
Watermelon	البطيخ	972	243
Mandarin	اليوسفى	697	174
Cowpea	اللوبيا	793	198

In addition, we tested 25 % of different queries on each crop. Table IV shows the number of Arabic complaint queries for each crop.

TABLE V. EXAMPLES FOR ARABIC COMPLAINTS QUERIES AND THEIR ENGLISH EQUIVALENT

Complaints queries in the English language	**Complaints queries in the Arabic language**
The presence of aphids inside the okra fruits**.**	وجود حشرة المن داخل ثمار الباميا.
The presence of brown spotting on apricot trees**.**	وجود تبقع بنى على اشجار المشمش.
The appearance of black spots on the potato leaves.	ظهور بقع سوداء على اوراق البطاطس.
The appearance of oval spots of different sizes on the onion leaves	ظهور بقع بيضاوية مختلفة الحجم على اوراق البصل.
The appearance of a minute white layer on the peach trees.	ظهور طبقة بيضاء دقيقه على اغصان اشجار الخوخ.
High rates of apple fruit fall.	ارتفاع نسب تساقط ثمار التفاح.
Yellowing of mandarin trees, complete yellowing with their fall.	اصفرار شجرة اليوسفى اصفرار كامل مع تساقط معظم اوراق اليوسفى.

TABLE VI. EXAMPLES OF EVALUATION RESULTS FOR TF-BASED LSA APPROACH

	Evaluation Methods		
Crops Name	***Average Precision***	***Average Recall***	***Average F-Measure***
الباميا (Okra)	83.5%	80.3%	81.9%
اليوسفى (Mandarin)	86.5%	83.3%	84.9%
البطيخ (Watermelon)	84.7%	82.0%	83.3%
القمح (Wheat)	85.6%	82.3%	83.9%
الأرز (Rice)	86.2%	83.2%	84.7%
القطن (Cotton)	85.5%	83.3%	84.4%
الفاصوليا (Beans)	83.5%	81.7%	82.6%
الطماطم (Tomatoes)	86.0%	80.4%	83.1%
البطاطس (Potato)	80.7%	78.5%	79.6%
الخوخ (Peach)	86.4%	82.3%	84.3%
المشمش (Apricot)	81.5%	80.4%	80.9%
العدس (Lentils)	85.0%	83.5%	84.2%
البصل (Onions)	79.8%	76.7%	78.2%
البرسيم (Clover)	83.5%	80.3%	81.9%
التفاح (Apples)	80.0%	78.4%	79.2%
الباذنجان (Eggplant)	82.5%	80.3%	81.4%
العنب (Grapes)	79.5%	77.7%	78.6%
البرتقال (Orange)	81.5%	80.8%	81.1%
الموز (Banana)	85.4%	81.5%	83.4%
الجوافة (Guava)	81.7%	79.5%	80.6%
البسلة (Peas)	85.7%	80.6%	83.1%
اللوبيا (Cowpea)	80.0%	78.5%	79.2%
الكرنب (Cabbage)	80.5%	76.3%	78.3%
الثوم (Garlic)	81.0%	79.5%	80.2%
الخس (Lettuce)	83.7%	79.8%	81.7%

Table V shows some examples of the Arabic queries complaints related to different crops and their English translation.

In TF-based LSA, average Precision, Recall, and F-measure values for the twenty-five crops are shown in Table VI.

In TF_IDF-based LSA, we also apply the previous Arabic queries in Table IV on each crop of the previous twenty-five crops. Thus, average Precision, Recall, and F-measure values of the TF_IDF-based LSA for the twenty-five crops are shown in Table VII.

As a conclusion, by comparing the evaluation results of the TF-based LSA approach with the TF_IDF-based LSA approach, we conclude that the results of TF_IDF-based LSA approach achieved the best results since TF-IDF measures how important a term in complaints that give high weight for important terms while TF shows the only number of times that a term appears in a complaint.

TABLE VII. EXAMPLES OF EVALUATION RESULTS FOR TF_IDF-BASED LSA APPROACH

	Evaluation Methods		
Crops Name	*Average Precision*	*Average Recall*	*Average F-Measure*
الباميا (Okra)	85.4%	83.0%	84.2%
اليوسفي (Mandarin)	88.3%	85.2%	86.7%
البطيخ (Watermelon)	87.0%	84.9%	85.9%
القمح (Wheat)	86.5%	83.3%	84.9%
الأرز (Rice)	86.7%	84.0%	85.3%
القطن (Cotton)	86.8%	84.3%	85.5%
الفاصوليا (Beans)	84.3%	82.7%	83.5%
الطماطم (Tomatoes)	87.4%	83.7%	85.5%
البطاطس (Potato)	81.6%	80.5%	81.0%
الخوخ (Peach)	86.7%	84.4%	85.5%
المشمش (Apricot)	82.1%	81.1%	81.6%
العدس (Lentils)	85.4%	83.6%	84.5%
البصل (Onions)	81.3%	79.4%	80.3%
البرسيم (Clover)	84.8%	81.5%	83.1%
التفاح (Apples)	82.5%	80.4%	81.4%
الباذنجان (Eggplant)	84.0%	83.2%	83.6%
العنب (Grapes)	81.5%	79.3%	80.4%
البرتقال (Orange)	82.6%	81.7%	82.1%
الموز (Banana)	87.8%	84.6%	86.2%
الجوافة (Guava)	83.4%	81.3%	82.3%
البسلة (Peas)	87.0%	82.8%	84.8%
اللوبيا (Cowpea)	81.0%	79.8%	80.4%
الكرنب (Cabbage)	82.0%	78.5%	80.2%
الثوم (Garlic)	82.6%	81.5%	82.0%
الخس (Lettuce)	84.5%	82.3%	83.4%

TABLE VIII. EXAMPLES OF EVALUATION RESULTS FOR PREVIOUS MODELS

	Evaluation Methods		
Crops Name	*Average Precision*	*Average Recall*	*Average F-Measure*
الباميا (Okra)	79.3%	76.7%	78.0%
اليوسفي (Mandarin)	83.9%	80.4%	82.1%
البطيخ (Watermelon)	81.0%	77.9%	79.4%
القمح (Wheat)	81.8%	79.5%	80.6%
الأرز (Rice)	82.5%	79.4%	80.9%
القطن (Cotton)	83.5%	80.0%	81.7%
الفاصوليا (Beans)	80.3%	78.5%	79.4%
الطماطم (Tomatoes)	83.4%	77.7%	80.4%
البطاطس (Potato)	77.8%	76.0%	76.9%
الخوخ(Peach)	83.5%	79.4%	81.4%
المشمش (Apricot)	80.4%	78.0%	79.2%
العدس (Lentils)	82.9%	81.2%	82.0%
البصل (Onions)	78.0%	75.6%	76.8%
البرسيم (Clover)	79.4%	78.7%	79.0%
التفاح (Apples)	78.0%	76.8%	77.4%
الباذنجان (Eggplant)	81.0%	79.9%	80.4%
العنب (Grapes)	77.9%	76.8%	77.3%
البرتقال (Orange)	80.6%	79.9%	80.2%
الموز (Banana)	82.4%	78.3%	80.3%
الجوافة (Guava)	80.5%	78.2%	79.3%
البسلة (Peas)	84.1%	79.6%	81.8%
اللوبيا (Cowpea)	78.8%	75.9%	77.3%
الكرنب (Cabbage)	78.4%	75.8%	77.1%
الثوم (Garlic)	78.6%	76.4%	77.5%
الخس (Lettuce)	80.3%	78.6%	79.4%

Finally, we compared our results with the previous models as Schwab et al. [23]. They applied three methods that are no weighting method, IDF weighting method, and POS weighting method. Schwab concluded that applying both the IDF and POS weighting methods achieved better results in performance. Therefore, we apply the previous models (IDF and POS weighting methods) to our agricultural dataset.

Table VIII shows the average Precision, Recall, and F-measure values for the previous models. We also apply IDF and POS weighting methods on the same twenty-five crops used in our proposed TF-based LSA and TF_IDF-based LSA approach.

Finally, Fig. 2 show the F-measure evaluation results of our two proposed model TF_IDF-based LSA and TF-based LSA compared with the previous models of IDF and POS weighting methods.

The comparison figure shows that the proposed TF_IDF-based LSA achieves better results than the proposed TF-based LSA and previous models of IDF and POS weighting methods.

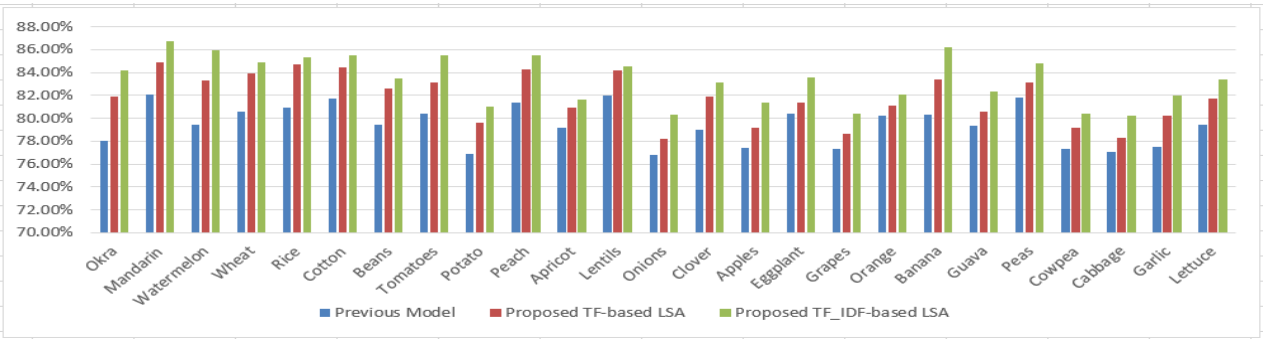

Fig. 2. Average F-Measure Values for Our Two Proposed Models and Previous Models.

V. RESULT

The LSA-based proposed model is applied to retrieve the most relevant complaint and its solution to the farmer query.

Consider the example in Table IX for an Arabic farmer query.

TABLE IX. EXAMPLE FOR ARABIC FARMERS' QUERY AND ITS ENGLISH EQUIVALENT

Arabic Farmer Query	English Farmer Query
وجود ديداان فى الباميـــــا	Presence of worms in okra

As shown in Table IX, the Arabic farmer query is " وجود ديداان فى الباميـــــا" and in English equivalent is "Presence of worms in okra".

The model will be followed step by step as follows:

Firstly, apply the preprocessing steps on the Arabic farmer query and all complaints in the dataset. Table X represents preprocessing steps on the Arabic farmer query.

Secondly, applying classification by MapReduce SVM approach using Hadoop to classify Arabic farmers' query into a crop that belongs to. As in Arabic farmer query "وجود ديدان باميا", this query is classified into باميا "okra" crop.

Thirdly, applying LSA steps that the input matrix is created for Arabic farmer query and all complaints that belong to okra crop. Then, apply RSVD to the input matrix.

TABLE X. PREPROCESSING STEPS FOR ARABIC FARMERS' QUERY AND ITS ENGLISH EQUIVALENT

Preprocessing steps	Arabic farmer query after preprocessing	English farmer query after preprocessing
Tokenization	"وجود" , " ديداان" , " فى" , " الباميـــــا"	'Presence' , 'worrrms' , ' in' , 'okra'
Stop Words Removal	"وجود" , " ديداان" , " الباميـــــا"	'Presence' , 'worrrms' , 'okra'
Auto-correction	"وجود" , " ديداان" , " الباميـــــا"	'Presence' , 'worrrms' , 'okra'
Normalization	"وجود," " ديدان," " الباميا"	'Presence' , 'worms' , 'okra'
Lemmatization	"وجود," " ديدان," " باميا"	'Presence' , 'worms' , 'okra'

Finally, the output of RSVD is used to measure the semantic similarity score between the farmer query and the complaints document.

According to the proposed two LSA methods, which are TF-based LSA and TF_IDF-based LSA, Table XI shows the results semantic similarity score between the farmer query and the complaints document.

As shown in Table XI, the results of the semantic similarity score using TF_IDF-based LSA are better than the semantic similarity score using TF-based LSA since TF-IDF shows how important a term is in complaints while TF shows the number of times that a term appears in a complaint.

Fourthly, rank the complaints according to the semantic similarity score of TF-based and TF_IDF-based LSA, as shown in Table XII and Table XIII, respectively, then select the complaint with the highest score.

TABLE XI. SEMANTIC SIMILARITY SCORE USING TF AND TF_IDF-BASED LSA BETWEEN THE FARMER QUERY AND THE COMPLAINTS DOCUMENT

Arabic farmer query	Complaints	Semantic similarity score using TF_IDF-based LSA	Semantic similarity score using TF-based LSA
وجود ديداان فى الباميـــــا **Presence worms in okra**	تلاحظ وجود حشرة المن على نباتات الباميا. Note the presence of aphids on okra plants.	0.977	0.941
	سأل المزارع عن كيفيه حفظ محصول الباميا و طريقة حفظها لاستعمالها فى الوقت الغير متوفره فيها. The farmer asked how to preserve the okra crop and how to preserve it for use at a time not available in it.	0.045	0.994
	وجود ديدان صغيره داخل ثمار الباميا. The presence of small worms inside the okra fruits.	0.9998	0.865
	وجود بقع دقيقة على سطحى الورقة لنبات الباميا تتحول هذة البقع الى اللون البنى و تجف الاوراق المصابة و تموت مما يؤدى الى صغر حجم النبات. The presence of minute spots on the two leaf surfaces of the okra plant, these spots turn brown, and the affected leaves dry and die, which leads to the small size of the plant.	0.315	0.713
	وجود ثقوب فى ازهار نباتات الباميا و جفاف وسقوط الازهار مع وجود ديدان فى بعض الازهار. The presence of holes in the flowers of the okra plants, drought, and fall of the flowers, with the presence of worms in some flowers.	0.994	0.687
	وجود بقع صفراء على بعض الأوراق فى نباتات الباميا مع وجود أوراق صفراء (تلاحظ وجود الحشرة الكاملة للذبابة البيضاء على الأوراق) The presence of yellow spots on some leaves in okra plants, with yellow leaves (the presence of the adult whitefly insect is noticed on the leaves)	0.654	0.933

TABLE XII. THE RANKED COMPLAINTS ACCORDING TO THE TF_IDF-BASED SEMANTIC SIMILARITY SCORE

Arabic farmer query	Complaints	Ranked semantic similarity score using TF_IDF-based LSA
وجود ديداان فى الباميـــا Presence worms in okra	وجود ديدان صغيره داخل ثمار الباميا. The presence of small worms inside the okra fruits.	0.9998
	وجود ثقوب فى ازهار نباتات الباميا و جفاف وسقوط الازهار مع وجود ديدان فى بعض الازهار. The presence of holes in the flowers of the okra plants, drought, and fall of the flowers, with the presence of worms in some flowers.	0.994
	تلاحظ وجود حشرة المن على نباتات الباميا. Note the presence of aphids on okra plants.	0.977
	وجود بقع صفراء على بعض الأوراق فى نباتات الباميا مع وجود أوراق صفراء (تلاحظ وجود الحشرة الكاملة للذبابة البيضاء على الأوراق). The presence of yellow spots on some leaves in okra plants, with yellow leaves (the presence of the adult whitefly insect is noticed on the leaves)	0.654
	وجود بقع دقيقة على سطحى الورقة لنبات الباميا تتحول هذة البقع الى اللون البنى و تجف الاوراق المصابة و تموت مما يؤدى الى صغر حجم النبات. The presence of minute spots on the two leaf surfaces of the okra plant, these spots turn brown, and the affected leaves dry and die, which leads to the small size of the plant.	0.315
	سأل المزارع عن كيفيه حفظ محصول الباميا و طريقة حفظها لاستعمالها فى الوقت الغير متوفره فيها. The farmer asked how to preserve the okra crop and how to preserve it for use at a time not available in it.	0.045

As shown in Table XII, after ranking semantic similarity score and selecting the highest score that is 0.9998 and its complaint is "وجود ديدان صغيره داخل ثمار الباميا" which is the nearest complaint to farmer query.

As shown in Table XIII, after ranking semantic similarity score and selecting the highest score that is 0.994 and its complaint is " سأل المزارع عن كيفيه حفظ محصول الباميا و طريقة حفظها لاستعمالها فى الوقت الغير متوفره فيها." which is not the nearest complaint to farmer query.

By comparing the results of both the TF_IDF-based LSA approach and the TF-based LSA approach, we conclude that the TF_IDF-based LSA approach is the best method for measuring the semantic similarity score.

Finally, based on the results of TF_IDF-based LSA approach in Table XII, the solution for the farmer query "وجود ديداان فى الباميـــا" according to the agricultural problems/solutions dataset is " جمع القرون المصابه واعدامها لعدم امكانية رش القرون قبل الاستهلاك" as shown in Table XIV.

TABLE XIII. THE RANKED COMPLAINTS ACCORDING TO THE TF-BASED SEMANTIC SIMILARITY SCORE

Arabic farmer query	Complaints	Ranked semantic similarity score using TF-based LSA
وجود ديداان فى الباميـــا Presence worms in okra	سأل المزارع عن كيفيه حفظ محصول الباميا و طريقة حفظها لاستعمالها فى الوقت الغير متوفره فيها. The farmer asked how to preserve the okra crop and how to preserve it for use at a time not available in it.	0.994
	تلاحظ وجود حشرة المن على نباتات الباميا. Note the presence of aphids on okra plants.	0.941
	وجود بقع صفراء على بعض الأوراق فى نباتات الباميا مع وجود أوراق صفراء (تلاحظ وجود الحشرة الكاملة للذبابة البيضاء على الأوراق). The presence of yellow spots on some leaves in okra plants, with yellow leaves (the presence of the adult whitefly insect is noticed on the leaves)	0.933
	وجود ديدان صغيره داخل ثمار الباميا. The presence of small worms inside the okra fruits.	0.865
	وجود بقع دقيقة على سطحى الورقة لنبات الباميا تتحول هذة البقع الى اللون البنى و تجف الاوراق المصابة و تموت مما يؤدى الى صغر حجم النبات. The presence of minute spots on the two leaf surfaces of the okra plant, these spots turn brown, and the affected leaves dry and die, which leads to the small size of the plant.	0.713
	وجود ثقوب فى ازهار نباتات الباميا و جفاف وسقوط الازهار مع وجود ديدان فى بعض الازهار. The presence of holes in the flowers of the okra plants, drought, and fall of the flowers, with the presence of worms in some flowers.	0.687

TABLE XIV. SOLUTION FOR THE FARMER QUERY

Farmer query	Complaint	Solution
وجود ديداان فى الباميـــا Presence worms in okra	وجود ديدان صغيره داخل ثمار الباميا. Presence of small worms inside the okra fruits.	جمع القرون المصابه واعدامها لعدم امكانية رش القرون قبل الاستهلاك (The infected pods were collected and destroyed because the pods could not be sprayed before consumption.)

VI. CONCLUSION

Agriculture has an important role in the economy of every country. Not only supplying foods for the whole population of a country but also it helps to connect and interact with all the relative industries of the country. Due to the world's current conditions from the spread of COVID-19, the imposition of a curfew, and adequate spacing between citizens, all fields are affected, especially the agriculture field. Farmers may have problems and complaints related to the agriculture process and the productivity of the percentage of the crops. It is difficult for farmers to communicate with agricultural experts to find appropriate solutions for their complaints. A semantic similarity approach for agriculture farmers' complaints is developed to solve these issues. This approach is based on LSA to measure semantic similarity between farmer query and the complaints document. The proposed model is applied to the MapReduce SVM using Hadoop for classifying the big agricultural dataset and the farmer complaint according to the crop type to improve the performance of the proposed approach. The results are evaluated on twenty-five crops and tested 25% of different complaint queries on each crop of them. These evaluations applied to our two proposed models of TF-based LSA, TF_IDF-based LSA, and previous work methods. The developed approach with TF_IDF-based LSA achieved better results than the TF-based LSA and previous work methods with an F-measure of 86.7%.

ACKNOWLEDGMENT

This work was supported by a Newton Institutional Links grant, ID 347762518, under the Egypt Newton-Mosharafa Fund partnership. The grant is funded by the 'UK Department for Business, Energy and Industrial Strategy' and 'Science and Technology Development Fund (STDF)' and delivered by the British Council. For further information, please visit www.newtonfund.ac.UK.

REFERENCES

[1] A. Pawar and V. Mago, "Challenging the Boundaries of Unsupervised Learning for Semantic Similarity," IEEE Access, vol. 7, no. January, pp. 16291–16308, 2019.

[2] M. K. and D. Chidambaram, "A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories," Int J Adv Comput Sci Appl, vol. 7, no. 8, pp. 231–237, 2016.

[3] D. Hussen Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. Mohammed Sadeeq, and K. Hussein Sharif, "State of Art for Semantic Analysis of Natural Language Processing," Qubahan Acad J, vol. 1, no. 2, 2021.

[4] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," Procedia Comput Sci, vol. 117, no. September, pp. 256–265, 2017.

[5] M. Atabuzzaman, M. Shajalal, M. E. Ahmed, M. I. Afjal, and M. Aono, "Leveraging Grammatical Roles for Measuring Semantic Similarity between Texts," IEEE Access, vol. 9, pp. 62972–62983, 2021.

[6] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf, pp. 615–620, 2014.

[7] A. Y. Ichida, F. Meneguzzi, and D. D. Ruiz, "Measuring Semantic Similarity between Sentences Using A Siamese Neural Network," Proc Int Jt Conf Neural Networks, vol. 2018-July, pp. 1–7, 2018.

[8] B. Hassan, S. E. Abdelrahman, R. Bahgat, and I. Farag, "UESTS: An Unsupervised Ensemble Semantic Textual Similarity Method," IEEE Access, vol. 7, pp. 85462–85482, 2019.

[9] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity-A Survey," ACM Comput Surv, vol. 54, no. 2, pp. 1–35, 2021, doi: 10.1145/3440755.

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

[11] M. A. R. Abdeen, S. AlBouq, A. Elmahalawy, and S. Shehata, "A closer look at arabic text classification," Int J Adv Comput Sci Appl, vol. 10, no. 11, pp. 677–688, 2019.

[12] and O. A.-M. Mohammad, Adel Hamdan, Tariq Alwada'n, "Arabic text categorization using support vector machine, Naïve Bayes and neural network.," GSTF J Comput 51, vol. Volume 5, no. 1, pp. 40–44, 2016.

[13] "VERCON." http://www.vercon.sci.eg/ (accessed Sep. 02, 2021).

[14] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," Int J Grid Distrib Comput, vol. 11, no. 9, pp. 103–114, 2018.

[15] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," J Inf Sci, vol. 40, no. 4, pp. 501–513, 2014.

[16] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," Knowl Inf Syst, vol. 53, no. 3, pp. 805–831, 2017.

[17] C. A. Flores, R. L. Figueroa, and J. E. Pezoa, "Active Learning for Biomedical Text Classification Based on Automatically Generated Regular Expressions," IEEE Access, vol. 9, pp. 38767–38777, 2021.

[18] H. Al Saif and T. Alotaibi, "Arabic text classification using feature-reduction techniques for detecting violence on social media," Int J Adv Comput Sci Appl, vol. 10, no. 4, pp. 77–87, 2019.

[19] F. Ö. Çatak and M. E. Balaban, "A Map Reduce-based distributed SVM algorithm for binary classification," Turkish J Electr Eng Comput Sci, vol. 24, no. 3, pp. 863–873, 2016.

[20] M. Ali, D. S. Guru, and M. Suhil, Classifying Arabic Farmers' Complaints Based on Crops and Diseases Using Machine Learning Approaches, vol. 1037. Springer Singapore, 2019.

[21] R. Al-khurayji and A. Sameh, "An Effective Arabic Text Classification Approach Based on Kernel Naive Bayes Classifier," Int J Artif Intell Appl, vol. 8, no. 6, pp. 01–10, 2017.

[22] and K. B. A. Abutiheen, Zinah Abdulridha, Ahmed H. Aliwy, "Arabic text classification using master-slaves technique," J Phys Conf Ser, vol. 1032, no. May, 2018.

[23] E. M. B. NAGOUDI, J. Ferrero, and D. Schwab, "LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting," Proc 11th Int Work Semant Eval, no. June, pp. 134–138, 2018.

[24] A. El Hadi, Y. Madani, R. El Ayachi, and M. Erritali, "A new semantic similarity approach for improving the results of an Arabic search engine," Procedia Comput Sci, vol. 151, pp. 1170–1175, 2019.

[25] A. Mahmoud and M. Zrigui, "Semantic similarity analysis for paraphrase identification in Arabic texts," PACLIC 2017 - Proc 31st Pacific Asia Conf Lang Inf Comput, pp. 274–281, 2019.

[26] A. Awajan, "Semantic similarity based approach for reducing Arabic texts dimensionality," Int J Speech Technol, vol. 19, no. 2, pp. 191–201, 2016.

[27] S. Malallah, A. Qassim, and A. Alameer, "Finding the Similarity between Two Arabic Text," Iraqi J Sci, vol. 58, no. 1, pp. 152–162, 2017.

[28] A. El Kah and I. Zeroual, "The effects of Pre-Processing Techniques on Arabic Text Classification," Int J Adv Trends Comput Sci Eng, vol. 10, no. 1, pp. 41–48, 2021.

[29] A. Ayedh, G. TAN, K. Alwesabi, and H. Rajeh, "The Effect of Preprocessing on Arabic Document Categorization," Algorithms, vol. 9, no. 2, 2016.

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

[31] B. Li, Z. Li, T. Li, and J. Liu, "A portable embedded automobile exhaust detection device based," 2013 IEEE 3rd Int Conf Inf Sci Technol ICIST 2013, no. December, pp. 126–128, 2013.

[32] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," J King Saud Univ - Comput Inf Sci, vol. 29, no. 2, pp. 189–195, 2017.

[33] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," Proc 2016 Conf North Am chapter Assoc Comput Linguist Demonstr, vol. 2016, pp. 11–16, 2016.

[34] D. AlSaleh and S. Larabi-Marie-Sainte, "Arabic Text Classification using Convolutional Neural Network and Genetic Algorithms," IEEE Access, vol. 9, pp. 91670–91685, 2021.

[35] M. S. Shanoda, S. A. Senbel, and M. H. Khafagy, "JOMR: Multi-join optimizer technique to enhance map-reduce job," 2014 9th Int Conf Informatics Syst INFOS 2014, no. May, pp. PDC80–PDC87, 2015.

[36] M. H. Mohamed and M. H. Khafagy, "Hash semi cascade join for joining multi-way map reduce," IntelliSys 2015 - Proc 2015 SAI Intell Syst Conf, no. November, pp. 355–361, 2015.

[37] M. Aksa, J. Rashid, M. W. Nisar, T. Mahmood, H. Y. Kwon, and A. Hussain, "Bitmapaligner: Bit-parallelism string matching withmapreduce and hadoop," Comput Mater Contin, vol. 68, no. 3, pp. 3931–3946, 2021.

[38] N. K. Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework," J Big Data, vol. 2, no. 1, pp. 1–18, 2015.

[39] R. M. Badry and I. F. Moawad, A Semantic Text Summarization Model for Arabic Topic-Oriented, vol. 921, no. January. Springer International Publishing, 2020.

[40] A. V. Nimkar and D. R. Kubal, "A survey on word embedding techniques and semantic similarity for paraphrase identification," Int J Comput Syst Eng, vol. 5, no. 1, p. 36, 2019.

[41] M. S. Al-Batah, S. Mrayyen, and M. Alzaqebah, "Arabic Sentiment Classification using MLP Network Hybrid with Naive Bayes Algorithm," J Comput Sci, vol. 14, no. 8, pp. 1104–1114, 2018.