English Semantic Similarity based on Map Reduce Classification for Agricultural Complaints

Esraa Rslan[1], Rasha M.Badry[4]
Information Systems Department
Faculty of Computers and Information, Fayoum University
Fayoum, Egypt

Mohamed H.Khafagy[2]
Computer Science Department
Faculty of Computers and Information, Fayoum University
Fayoum, Egypt

Kamran Munir[3]
Computer Science and Creative Technologies Department
University of West of England
Bristol, United Kingdom

***Abstract*—Due to environmental changes, including global warming, climatic changes, ecological impact, and dangerous diseases like the Coronavirus epidemic. Since coronavirus is a hazardous disease that causes many deaths, government of Egypt undertook many strict regulations, including lockdowns and social distancing measures. These circumstances have affected agricultural experts' presence to help farmers or advise on solving agricultural problems. For helping this issue, this work focused on improving support for farmers on the major field crops in Egypt Retrieving solutions corresponding to farmer query. For our work, we have mainly focused on detecting the semantic similarity between large agriculture dataset and user queries using Latent Semantic Analysis (LSA) based on Term Frequency Weighting and Inverse Document Frequency (TF-IDF) method. In this research paper, we apply SVM MapReduce classifier as a framework for paralleling and distributing the work on the dataset to classify the dataset. Then we apply different approaches for computing the similarity of sentences. We presented a system based on semantic similarity methods and support vector machine algorithm to detect the similar complaints of the user query. Finally, we run different experiments to evaluate the performance and efficiency of the proposed system as the system performs approximately 77.8%~94.8% in F-score measure. The experimental results show that the accuracy of SVM classifier is approximately 88.68%~89.63% and noted the leverage of SVM classification to the semantic similarity measure between sentences.**

***Keywords*—Agricultural system; semantic textual similarity; text classification; latent semantic analysis; part of speech**

I. Introduction

The semantic similarity of sentences has many real applications like Intelligent Question Answering (IQA) system. When a question is asked, the existing answer can be returned if a similar question is found in the database. In this paper, we provided a solution for calculating semantic similarity between sentences that based on vectoring sentences using their syntactic and semantic features. Semantic Textual Similarity (STS) is focusing on finding the similarity between two sentences. Similarity between the sentences is based on the explicit or implicit semantic relationships between them[1]. These relationships can be identified or measured by finding semantic relations among them. Many algorithms are presented for textual similarity. We can group them based on the algorithm or method that we used to perform the semantic similarity process.

Agriculture has a huge impact in the economy of countries. Since over a huge number of the population in Egypt is dependent on agriculture. Moreover, it considers to be one the source national economy, foreign currency, Livelihood, and food supply [2]. Further, it creates job opportunities to a large scale of the population.

This paper uses the an English approach based on latent semantic analysis [3],[4] for measuring the semantic similarity between English sentences of agricultural data and user query to find the appropriate solution for the complaints of farmers. The proposed system used SVM classification in MapReduce Hadoop environment to classify the agricultural dataset complaints based on crop name to improve the efficiency of the semantic similarity process.

Therefore, the aim of the approach is providing the support for experts and farmers in the system in Egypt. The complaints' associations are distributed over around 4242 villages and 198' centers' across Egypt [5]. In Arabic script format, these complaints' are submitted to support for farmers in their agriculture problems. Storage all farmer agriculture problems stored on a public cloud which hosting analytics toolkits [6], [7].

In our approach, first; the farmer submits his agriculture problem in the Arabic language; then, Google machine translation is used to translate the problem from Arabic into English. Second, Analyses of the complaints through data analytics techniques to extract (most) term frequency and classify the query to which crop class using support vector machine in map/reduce model. The classification process might take some time to correctly classify the crops. Third, Building an automated support response by searching for similar complaints within the agriculture complaint datasets. We saved our dataset on the public cloud to store massive data or the number of complaints as big data. Our key focus has been on calculating the semantic similarity between Arabic and English cross-language sentences using LSA. We consider different methods like term frequency weighting and inverse document frequency to identify words in each complaint. The rest of the paper is presented as follows: Related Work in Section II describes a few Semantic Textual Similarity approaches. In Section III, Proposed System, we present our proposed LSA with SVM classification. Section IV, Discussion and Results describe the experimental results of

these methods. Finally, the Section V, Conclusion will be presented.

II. Related Work

Words can be similar in two ways lexical or semantical. Similar lexical words, if the words have the same sequence of character. Similar semantical words, if they have almost the same meaning, used in the same way, used in the same context. The String-Based algorithm is based on lexical similarity. Corpus-Based and Knowledge-Based algorithms are based on Semantic Similarity. String similarity measures operate on word sequences and character composition. It can be categorized into two sets: Character-Based Similarity, Term-based Similarity Measures. Character-Based Similarity like longest Common SubString (LCS) algorithm. N-gram algorithm Smith-Waterman [8]. Term-based Similarity Measures like Cosine similarity measure, Euclidean distance, Jaccard similarity, and Block Distance that also called Manhattan Distance.

Corpus-Based Similarity: Latent Semantic Analysis (LSA) [3] is the most popular technique of Corpus-Based Similarity. Hyperspace Analogue to Language (HAL), Generalized Latent Semantic Analysis (GLSA), Explicit Semantic Analysis (ESA), Normalized Google Distance (NGD). Knowledge-Based Similarity can be categorized into three groups like: (1) node-based/ information content (IC): like Resnik (res), and Conrath (jcn), (2) edged-based like Lesk, and vector pairs, and (3) hybrid where it combines both node and edge-based. We used LSA corpus-based algorithm in our work that depending on the corpus and word embedding to compute the semantic similarity degree between the sentences.

Nagoudi et al. [9] presented a word embedding representations for calculating the semantic similarity between Arabic and English sentences. This paper used machine translation and word embedding approach to get the properties of words like semantic and syntactic. Machine translation is used to translate English complaint into the Arabic one for applying a classical monolingual comparison. Word embedding methods are applied to measure the semantic similarity. The proposed method is used Bag-of-word alignment, IDF, and part of speech weighting to determine the most descriptive words in each sentence. The performance of this approach is evaluated on the four datasets of the shared task of SemEval in 2017. The results achieved the best accuracy rate compared to the other systems in the semantic text similarity in Arabic-English cross-language of SemEval 2017.

Wafa Wali et al. [4] proposed several methods for calculating the semantic similarity among two English sentences, which consider semantic and syntactic knowledge. It presented a technique for measuring sentence similarity, which combined the three components: lexical similarity, semantic similarity, and syntactic-semantic similarity. Lexical similarity included the common words, the semantic similarity used for finding the synonymy words, and the syntactic-semantic similarity based on common semantic arguments, thematic role, and semantic class. The word-based semantic similarity is measured for estimating the semantic degree among words by exploring the WordNet "[10] is a" taxonomy. Furthermore, the semantic argument is determined by the VerbNet database. The experiments are applied on the Microsoft Paraphrase Corpus and shown the metric F-score compared to other metrics. The results are shown that the proposed technique could support using several sentence features like semantic arguments and properties in measuring the sentence similarity. Therefore, this technique can be applied in many applications, such as plagiarism detection.

The author in [11] presented both the design and implementation of an evaluation system for English short answers. Handwritten Short Answer Evaluation System (HSAES) is an automated short answer system for determining the answer in answer papers and testing each short answer's marks depending on the model's knowledge during training. The proposed system was used the Optical Character Recognition (OCR) tools for extracting handwritten texts. Natural Language Processing is applied to retrieve the main feature from person tested datasets for answer keys and the handwritten of answer papers. The proposed system was used the cosine similarity approach for measuring the semantic similarity among sentences. Marks were given to each sentence in the evaluated answer paper. The developed model was applied for assessing the un-scored short answer marks.

Chandrathlake et al. [12] focused on providing an accuracy level for English news posts written on social media. The proposed system performed many functions: extract the news item's content, search the Internet for finding the similar posts in online articles sources, match the returned content with the online article sites' content and finally generate the accuracy level. Many Natural Language Processing techniques are used for developing this model like web scrolling, text summarization, URL ranking, and semantic similarity methods like Word2vec, part of speech, and cosine similarity. This system achieved an accuracy of 70% for the news posts on social media comparing with the trustable online news in the social media.

Taieb et al. [13] proposed a Features-based Measure of Sentences Semantic Similarity (FM3S) approach for computing the semantic similarity between English sentences. The proposed method combined three methods: the noun semantic similarity, the verb semantic similarity and the common word. This approach used the information content-based measure in computing similarity between keywords using the WordNet [14]. The experiments are performed and tested on the Microsoft Paraphrase Corpus (MPC) and scored the best results compared with other metrics for high similarity thresholds. The results showed that FM3S proved the importance of syntactic information, compound nouns, and verb tense in computing the semantic similarity.

Xiaolin Jin et al. [10] proposed a model based on Word2vec for measuring the semantic similarity between English sentences. This method was presented to solve the low universality problem and the contextual information's absence in calculating the word based on the dictionary. This method improved the approaches based on the Chinese dictionary, e.g., HowNet and Tongyici Cilin. It also used the word vector model as a weighing parameter for measuring the word

similarity after comparing the words' similarity by giving different weights to the three methods. The experiments were conducted on this algorithm and achieved a high Pearson coefficient. The proposed method could include most words that could effectively solve the word similarity calculation problem in the dictionary.

Many work related to SVM in the parallel environment (or distributed system) have introduced in Ngoc et al. [15], Wen et al [16], and Rao [17] . There are many researcher papers using Cloudera and Hadoop [18] Map/reduce. Studies using SVM [19] in parallel environment for semantic classification are proposed. However, there isn't work which combine them such as: Hadoop Map/ Reduce, SVM classification, parallel system, and semantic similarity. Our new system uses all of them.

III. Proposed System

This section presents the proposed system main steps as shown in Fig. 1. The proposed system has five steps: (1) translate the user query (farmer complaint) from Arabic into English language (2) preprocessing the farmer query; (3) classification method using SVM in map/reduce model (4) Finding the word vector, building the sentence vectors matrix using LSA; and computing the similarity between sentence vectors by using vector similarity methods like cosine similarity (5) The problems are ranked and then select the one with the highest semantic score.

A. Translation

In this step, the complaint text is translated into English language. We used Google Cloud translator API [20] to translate the Arabic sentence into English one.

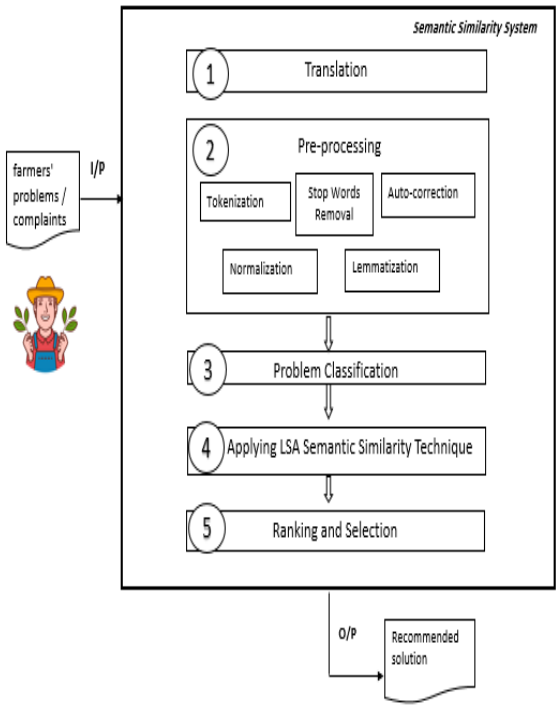

Fig. 1. The Proposed System.

B. Preprocessing

The farmer who describes the problem information like: the crop name, planting and watering method, and soil type. The farmer's query may contains useful words that effect the text processing phase. Pre-processing is important for removing the noise rows or data from historical complaint/response. It is focused on the historical agriculture dataset and farmer query to be used in this step [21]. Data pre-processing has many steps, like as: "tokenization", "stop words removal", "auto-correction", "normalization", and "lemmatization".

1) Tokenization : is the process of dividing written text into units (tokens) [21].White spaces, commas, semicolon and punctuations are used as a segment point in various languages especially in Arabic and English.

2) Stop words removal: is the process of removing unnecessary words. There are some words that are less importance, less useful, and less informative. These words are called stop words such as words in English complaints like "the", "is", "and", "an","a", etc. To enhance and generate a better solution, it is necessary to eliminate and remove these words using a predefined list. We also used the WordNet database that has a list of all English words. WorldNet is a huge lexical database of English verbs, nouns, pronouns, adjectives, and adverbs that are used for knowledge-based semantic similarity. WordNet's Relations make it a useful tool for natural language processing and computational linguistics.

3) Auto-correction: is used to correct errors made by the farmer when entering the complaint text. The complaints can also contain words written in a slang language. Auto-correction is used to solve such problem by replacing the incorrect word with the correct one.

4) Normalization: is the process of transforming the input text into a standard form. It focuses on removing inconsistent variations or unwanted data such as: "riice" is transformed to "rice".

5) Lemmatization: is the process of finding the base form of words, such as: "fruits" is transformed to "fruit".

C. Classification

In this phase, the classification is made semantically using SVM map reduce approach which is applied on the agriculture dataset and the farmer query in a parallel manner. The classification is paralleled between several machines using a Hadoop cluster with MapReduce [19], programming model for our work. Our approach is based on the English data set. The dataset is classified into a number of crop names like Wheat, Rice, Cotton, Local Bean, Tomato, Corn, Onion, and Beet Each group contains the SVM using Hadoop Map (M)/Reduce (R) is applied to classify the farmer query based on which crop class belongs to find the suitable solution.

D. LSA

Once the farmer complaint text and historical agriculture dataset are pre-processed and classified, and word vectors, the next step is to build a semantic model to compute the semantic similarity between the farmer query and the historical

agriculture dataset. LSA Algorithm builds in three main steps, Input Matrix Creation, Singular Value Decomposition (SVD), and Sentence Selection. Almost all previous works perform the first two steps of latent semantic similarity algorithm are in the same way. There is some difference in the word weighs like term-frequency and part of speech tagging which used to fill in the input matrix. Another difference is that they select words in the two sentence to measure the similarity [3] , [22]. The developed semantic model is based on LSA. LSA [23] is one of the most and important corpus-based techniques used for measuring semantic similarity. It consist of three steps are input matrix creation, singular value decomposition (SVD), and sentence selection.

A word co-occurrence matrix is calculated where the rows filling with the main words and columns filling with the sentences and the cells values have word occurrence counts. This matrix has an important underlying corpus so SVD dimensionality reduction is applied using a mathematical techniques. Such dimensionality reduction is highly used to: (i) minimize the output dimensionality and (ii) increase overall performance. Finally, the semantic score is calculated for each farmer complaint; then the sentences are ranked according to the semantic score to select the closest solution to the farmer query.

In this phase, an input matrix is computed for the farmer query and historical agriculture dataset. Each row in the matrix represents the word or term in the farmer query. Each column represents the problem. The cell value is the result of the intersection between term and problem. There are two ways that are used for filling the cell values, which are Term Frequency-Inverse Document Frequency (TF-IDF) or Term Frequency (TF). In TF-based LSA, the cells are filled with the term frequency (TF_i) of terms in the complaint query (C_j) according to Eq. 1.

$$W(t_{ij}) = tf_{ij} \quad (1)$$

Where $W(t_{ij})$ the weight of a term (i) in each problem (j), and tf_{ij} is the frequency of a term (i) in each problem text (j). In LSA bas based on TF, the cells are filled with the weight of term (i) in

problem statement (Cj) according to Eq. 2.

$$TF - IDF_{ij} = TF_{ij} * IDF_{ij} \quad (2)$$

Where TF-IDFij is TF is the frequency of a term (i) in each complaint statement (j), and IDF explain the importance of N terms between all problems.

E. Ranking and Selection

The semantic similarity is measured as the cosine value output between these sentences vectors. LSA system is generalized by changing rows with texts and columns with samples and can be used to compute the similarity between sentences, paragraphs, and documents. After applying the SVD matrix, the cosine similarity method will be calculated between the user complaint and each historical data problem to find the most suitable answer to identify the similarity among them. The cosine is calculated as Eq. 3:

$$\text{cosine similarity}(X, Y) = \frac{X.Y}{\|X\|.\|Y\|} \quad (3)$$

Ordering the agriculture problems based on to the semantic similarity result as shown in Fig. 2 decision function, and then select the problem (complaint) with the highest score based on the semantic similarity score.

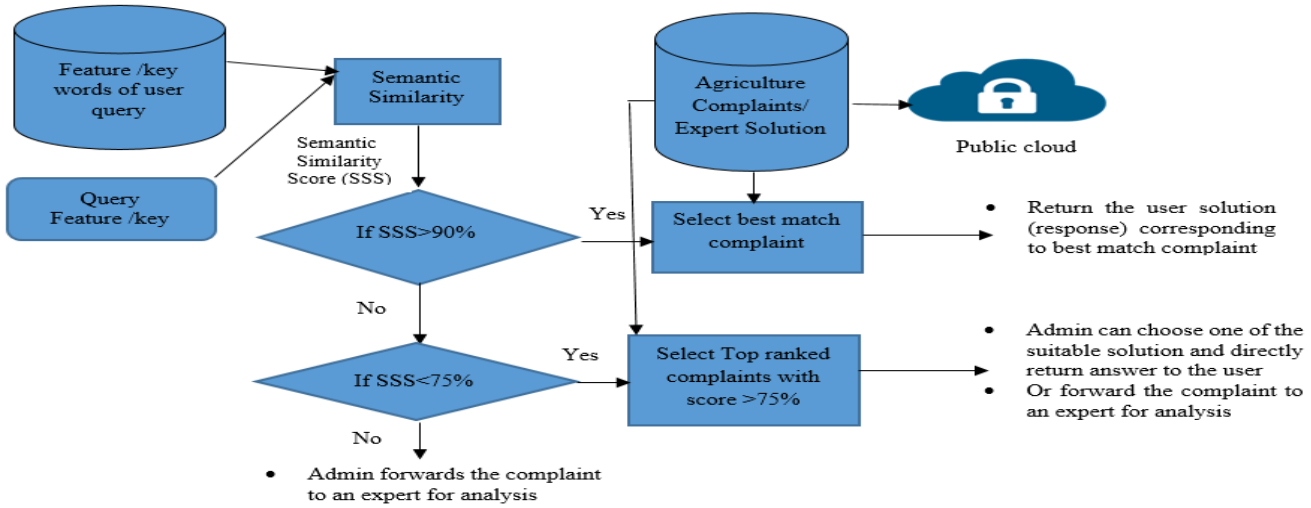

Fig. 2. Decision Function.

IV. Decision and Results

A. Environmental Setup

We used python programming language to implement our LSA model. The dataset is divided into 80% training and 20% testing with 10 experiments. We train the agriculture data set. The experiments are performed in devices with the following properties as shown in Table I.

B. Dataset

The dataset collected from Egypt's Virtual Extension Research Communication Network(VERCON) [5] and Agriculture Research Center (ARC), it contains historical complaints and solutions provided by the experts in text forms in English language complaints. The agricultural dataset was deployed on a public Cloud. The dataset is important because it has complaints/solutions from different agricultural problems that contain data for the main crop like: corn, cotton, wheat, and rice, also problem categories like environment, irrigation, pest, weed, diseases, and farming. Table II shows some examples of VERCON dataset. Table III presents the number of complaints in VERCON dataset in each crop.

C. Result Analysis

Consider the farmer query example: as presented in Table IV; firstly the farmer query is translated into English using google API. Secondly we apply preprocessing on the farmer query. Third, classify farmer query based on crop name by using SVM classifier in Hadoop Map/Reduce. Fourth create term frequency matrix. Fifth; compute semantic similarity score from the generating LSA matrix using TF-IDF or TF as shown in Tab. 3. The semantic similarity using TF-IDF achieved better result than the TF, because TF-IDF method shows the important features in each complaint however TF shows the number of term occurrence that appears in a complaint. Finally, ranking the complaints according to the semantic similarity score, and return solution of the farmer query with the highest similarity result.

We apply accuracy measure is to calculate the accuracy of SVM classification before semantic process. SVM is also used to predict the farmer query belongs to which crop class before being combined with semantic similarity process. The results show that the performance of classification with accuracy is approximately 88.68%~89.63%, as shown in Table V.

We proposed semantic similarity approach when using TF-based LSA and TF-IDF-based LSA. The work was evaluated using different measures like F-measure, precision and recall. F-measure expresses a trade-off between the two measures, precision, and recall as shown in Tables VI, VII and VIII. We compare our proposed results with the other models like POS (Part of Tagging) [24], [25]. The work was tested and evaluated on our agriculture data set. We test our dataset on different crop such as Wheat Rice, Cotton, Local Bean, Tomato, Corn, and other crops as shown in the following tables.

The F-measure in Table VI using TF-IDF weights scores the highest one about 0.939 in cotton crop, then the TF (term frequency) about 0.899 in the TF in Table VII while part of speech (POS) in Table VIII is 0.889. We run different kind of queries and get the average of F-measure, precision and recall. The F-measure in in Table VI using TF-IDF weights of F-score approximately 77.8%~94.8%, then the TF (term frequency) approximately 75.7%~92.3 while part of speech (POS) is approximately 73.3%~91.4%.

TABLE I. Environmental Setting

Item	Description
programming language	python
Processor	dual-core processor
CPU	Pentium CPU speed of 6.00 GHz
GPU(Graphics processing unit)	Tesla V100-SXM2-8GB
RAM (Random Access Memory)	8GB

TABLE II. Example of Dataset Complaints

Complaint	Solution
How to treat piercings with rice	Fyuridan is used at a rate of 6 kg per acre.
Yellowing of the lower leaves and the drying of the edges of the leaves in wheat.	Yellowing of the leaves at this time is normal, especially the lower one, to reach the ripening stage of the crop.
White spots on the leaf, spikes and stems feel cotton.	These symptoms of microflora disease, and if the infection requires severe chemotherapy, is spraying with a sumateite at a rate of 35 cm / 100 liters of water.

TABLE III. numer of complaints in historical dataset

Crop Name (English)	#of Complaints	Crop Name (English)	#of Complaints
Wheat	1073	Mango	435
Rice	1021	Citrus	254
Cotton	937	Grapes	247
Local Bean	783	Eggplant	227
Tomato	757	Green Pepper	199
Corn	648	Cucumber	178
Onion	546	Zucchini	168
Beet	461	Orange	153
Potato	440	Garlic	151
Clover	321	Guava	146

TABLE IV. An Example of System Process

Process	An example
Farmer query	يقع صفراء على أوراق نباتات البصل.
Translation	Yellow spots on the leaves of onion plants.
Tokenization	Yellow, spots, on, the, leaves, of, onion, plants
Stop word removal	Yellow, spots, leaves, onion, plants
Lemmatization	Yellow, spot, leaf, onion, plant
Classification	Onion class
Solution	These are the symptoms of onion thrips infection and it is treated with a 50% Acylac pesticide at a rate of 500 cm/100 liters of water, or a 72% silicron pesticide at a rate of 750 cm/f.

TABLE V. THE EVALUATION OF THE SVM CLASSIFICATION IN DATASET

Crop Name	# of records	Correct classification	Incorrect classification	Accuracy
Wheat	1073	955	118	89.19%
Rice	1021	909	112	89.03%
Cotton	937	834	103	88.98%
Local Bean	783	691	92	88.23%
Tomato	757	668	89	88.50%
Corn	648	572	76	88.43%
Onion	546	482	64	87.13%
Beet	461	407	54	89.23%
Potato	440	388	52	88.23%
Clover	321	282	39	88.58%
Mango	435	383	52	89.36%
Citrus	254	224	30	88.41%
Grapes	247	217	30	88.35%
Eggplant	227	203	24	89.64%
Green Pepper	199	178	21	89.11%
Cucumber	178	160	18	89.35%
Zucchini	168	151	17	89.06%
Orange	153	137	16	88.68%
Garlic	151	135	16	89.54%
Guava	146	131	15	89.14%
Summary	9145	8106	1039	88.63%

TABLE VI. EVALUATION MEASURES FOR USING TF-IDF FOR EACH CROP

crop name	TF-IDF		
	Precision	**Recall**	**F-score**
Wheat	0.849	0.855	0.852
Rice	0.841	0.856	0.848
Cotton	0.926	0.952	0.939
Local Bean	0.864	0.853	0.858
Tomato	0.893	0.866	0.879
Corn	0.867	0.899	0.883
Onion	0.955	0.941	0.948
Beet	0.902	0.895	0.898
Potato	0.925	0.915	0.921
Clover	0.822	0.856	0.839
Mango	0.785	0.795	0.793
Citrus	0.796	0.813	0.804
Grapes	0.773	0.784	0.778
Eggplant	0.899	0.875	0.887
Green Pepper	0.866	0.879	0.872
Cucumber	0.866	0.879	0.872
Zucchini	0.942	0.954	0.948
Orange	0.864	0.855	0.861
Garlic	0.941	0.948	0.938
Guava	0.796	0.813	0.804

The results show that the proposed text similarity LSA model using the TD-IDF method resolves the problem of the low recall of words in traditional semantic approaches well, and high the similarity performance of relevant words more than using only term frequency (TF) as shown in Tables VI, VII and VIII. The tables show the different measures for using TF-IDF, TF and POS for weeds, pests, diseases and irrigation categories.

TABLE VII. EVALUATION MEASURES FOR USING TF FOR EACH CROP

crop name	TF		
	Precision	**Recall**	**F-score**
Wheat	0.773	0.795	0.784
Rice	0.891	0.874	0.882
Cotton	0.902	0.896	0.899
Local Bean	0.881	0.854	0.867
Tomato	0.806	0.783	0.794
Corn	0.822	0.843	0.832
Onion	0.967	0.942	0.954
Beet	0.952	0.943	0.947
Potato	0.914	0.952	0.933
Clover	0.811	0.806	0.808
Mango	0.763	0.752	0.757
Citrus	0.752	0.767	0.759
Grapes	0.799	0.812	0.805
Eggplant	0.889	0.879	0.884
Green Pepper	0.853	0.861	0.857
Cucumber	0.856	0.879	0.789
Zucchini	0.921	0.926	0.923
Orange	0.864	0.831	0.847
Garlic	0.911	0.923	0.917
Guava	0.791	0.802	0.796

TABLE VIII. EVALUATION MEASURES FOR USING POS FOR EACH CROP

crop name	POS		
	Precision	**Recall**	**F-score**
Wheat	0.823	0.845	0.834
Rice	0.865	0.856	0.863
Cotton	0.897	0.881	0.889
Local Bean	0.832	0.889	0.862
Tomato	0.802	0.816	0.809
Corn	0.819	0.841	0.831
Onion	0.922	0.895	0.908
Beet	0.831	0.856	0.843
Potato	0.894	0.854	0.874
Clover	0.788	0.823	0.805
Mango	0.734	0.744	0.739
Citrus	0.744	0.723	0.733
Grapes	0.786	0.796	0.791
Eggplant	0.882	0.876	0.879
Green Pepper	0.869	0.856	0.862
Cucumber	0.855	0.879	0.788
Zucchini	0.909	0.911	0.914
Orange	0.823	0.822	0.822
Garlic	0.926	0.923	0.891
Guava	0.789	0.798	0.793

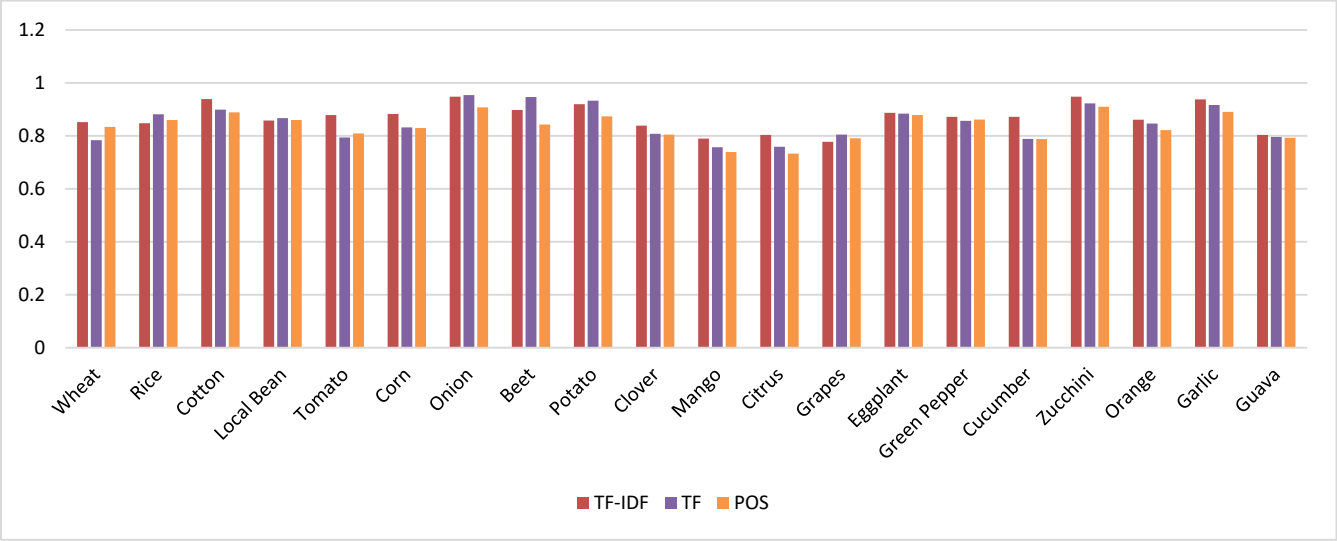

Fig. 3. F-score Values for different Metrics.

As a result, by evaluating the different experimental results of TF, TF-IDF and POS weight, we summarized that the results of LSA approach based on TF-IDF have the highest average F-measure as presented in Fig. 3.

V. Conclusion

Our work in this paper focused on building a semantic model for the available agricultural data, the design of interfaces and features for the system to ensure timely advice, easy access, consistency, and broadcasting service possible to farmers. We used MapReduce SVM classifier in Hadoop MapReduce to classify agricultural dataset into crops names. The performance of the system achieved better results than previous work. Also, we propose English semantic system for farmers' complaints that based on Latent Semantic Analysis depend on TF-IDF term to calculate similarity between user query and the complaints in the agriculture database. The results are tested on twenty different crops and also different complaint queries are applied on each crop. The system performed F-score with 0.939 using TF-IDF, then about 0.899 in the TF. The developed system with LSA based on TF-IDF achieved better results than the TF. The support provided by the system will be quickly and reliable not only for farmers but also for the 'research centers' and 'agricultural units' with minimal resources and training needs.

In the future work we will use different methods in semantic similarity process to enhance the system performance and also classify the dataset base on problem categories like pest, weed and irrigation.

Acknowledgment

This work has been supported by a Newton Institutional Links grant ID 347762518, under the Egypt Newton-Mosharafa Fund ID 30812 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] Chandrasekaran and V. Mago, "Evolution of Semantic Similarity - A Survey," arXiv, vol. 37, no. 4, 2020.

[2] G. Veeck, A. Veeck, and H. Yu, "Challenges of agriculture and food systems issues in China and the United States," Geogr. Sustain., vol. 1, no. 2, pp. 109–117, 2020, doi: 10.1016/j.geosus.2020.05.002.

[3] K. Al-Sabahi, Z. Zhang, J. Long, and K. Alwesabi, "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization," Arab. J. Sci. Eng., vol. 43, no. 12, pp. 8079–8094, 2018, doi: 10.1007/s13369-018-3286-z.

[4] W. Wali, B. Gargouri, and A. Ben Hamadou, "Sentence Similarity Computation based on WordNet and VerbNet," vol. 21, no. 4, pp. 627–635, 2017, doi: 10.13053/CyS-21-4-2853.

[5] "الهيئة العربية المصرية للاستشعار عن بعد." http://www.vercon.sci.eg/indexUI/uploaded/wheatinoldsoil/wheatinoldsoil.htm#r1 (accessed Sep. 01, 2021).

[6] "Apache Hadoop." http://hadoop.apache.org/ (accessed Sep. 01, 2021).

[7] V. N. Phu, V. T. N. Chau, and V. T. N. Tran, "SVM for English semantic classification in parallel environment," Int. J. Speech Technol., vol. 20, no. 3, pp. 487–508, 2017, doi: 10.1007/s10772-017-9421-5.

[8] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," Comput. y Sist., vol. 20, no. 4, pp. 647–665, 2016, doi: 10.13053/CyS-20-4-2506.

[9] E. Moatez et al., "Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences To cite this version : HAL Id : hal-01683494 Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences," 2018.

[10] X. Jin, S. Zhang, and J. Liu, "Word Semantic Similarity Calculation Based on Word2vec," ICCAIS 2018 - 7th Int. Conf. Control. Autom. Inf. Sci., pp. 12–16, 2018, doi: 10.1109/ICCAIS.2018.8570612.

[11] K. Al-Sabahi and Z. Zuping, "Document Summarization Using Sentence-Level Semantic Based on Word Embeddings," Int. J. Softw. Eng. Knowl. Eng., vol. 29, no. 2, pp. 177–196, 2019, doi: 10.1142/S0218194019500086.

[12] R. Chandrathlake, L. Ranathunga, S. Wijethunge, P. Wijerathne, and D. Ishara, "A Semantic Similarity Measure Based News Posts Validation on Social Media," 2018 3rd Int. Conf. Inf. Technol. Res. ICITR 2018, pp. 1–6, 2018, doi: 10.1109/ICITR.2018.8736136.

[13] 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, doi: 10.1109/ACCESS.2019.2925006.

[14] S. Zhang, Z. Liang, and J. Lin, "Sentence Similarity Measurement with Convolutional Neural Networks Using Semantic and Syntactic

Features," Comput. Mater. Contin., vol. 63, no. 2, pp. 943–957, 2020, doi: 10.32604/cmc.2020.08800.

[15] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4.5 algorithm for english emotional classification," Evol. Syst., vol. 10, no. 3, pp. 425–451, 2019, doi: 10.1007/s12530-017-9180-1.

[16] N. Yang et al., "Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification," Int. J. Speech Technol., vol. 20, no. 1, pp. 27–41, 2017, doi: 10.1007/s10772-016-9364-2.

[17] A. Haque and K. S. Rao, "Modification of energy spectra, epoch parameters and prosody for emotion conversion in speech," Int. J. Speech Technol., vol. 20, no. 1, pp. 15–25, 2017, doi: 10.1007/s10772-016-9386-9.

[18] G. S. Victor, P. Antonia, and S. Spyros, "CSMR: A scalable algorithm for text clustering with cosine similarity and MapReduce," IFIP Adv. Inf. Commun. Technol., vol. 437, pp. 211–220, 2014, doi: 10.1007/978-3-662-44722-2_23.

[19] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., vol. 12, no. 1, pp. 1–7, 2010.

[20] "Cloud Translation | Google Cloud." https://cloud.google.com/translate/ (accessed Sep. 01, 2021).

[21] D. A. Said, N. M. Wanas, N. M. Darwish, and N. H. Hegazy, "A Study of Text Preprocessing Tools for Arabic Text Categorization," Second Int. Conf. Arab. Lang. Resour. Tools, no. January 2009, pp. 230–236, 2009.

[22] Y. Liu, C. Sun, L. Lin, and X. Wang, "yiGou: A Semantic Text Similarity Computing System Based on SVM," no. SemEval, pp. 80–84, 2015, doi: 10.18653/v1/s15-2014.

[23] S. Alowaidi, M. Saleh, and O. Abulnaja, "Semantic Sentiment Analysis of Arabic Texts," vol. 8, no. 2, pp. 256–262, 2017.

[24] A. Voutilainen, "Part-of-Speech Tagging," Oxford Handb. Comput. Linguist., vol. 9780199276, no. June 2018, pp. 1–16, 2012, doi: 10.1093/oxfordhb/9780199276349.013.0011.

[25] V. Batanović and D. Bojić, "Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity," Comput. Sci. Inf. Syst., vol. 12, no. 1, pp. 1–31, 2015, doi: 10.2298/CSIS131127082B.