Received June 22, 2021, accepted September 15, 2021, date of publication September 24, 2021, date of current version October 6, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3115617*

# AltibbiVec: A Word Embedding Model for Medical and Health Applications in the Arabic Language

**MARIA HABIB[1], MOHAMMAD FARIS[1], ALAA ALOMARI[1], AND HOSSAM FARIS[1,2,3]**

[1]Altibbi.com, Amman 11831, Jordan
[2]King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan
[3]School of Computing and Informatics, Al Hussein Technical University, Amman 11183, Jordan

Corresponding author: Hossam Faris (hossam.faris@ju.edu.jo)

**ABSTRACT** In recent years, the utilization of natural language processing (NLP) and Machine Learning (ML) techniques in clinical decision support systems have shown their ability in improving and automating the diagnosis process, and reducing potential clinical errors. NLP in the Arabic language is more intricate due to several limitations, such as the lack of datasets and analytical resources compared to other languages like English. However, a clinical decision support system in the Arabic context is of significant importance. A fundamental process in NLP is extracting features from text-based data via text embedding. Word embedding is a representation of words in a numeric format that encodes the statistic, semantic, or context information. Building a neural word embedding model requires hundreds of thousands of data instances to find hidden patterns of relationships within sentences. Essentially, extracting relevant and informative features promotes the performance of the learning algorithms. The objective of this paper is to propose an Arabic neural-based word embedding model in the medical and healthcare context (called ''AltibbiVec''). Around 1.5 million medical consultations and questions written in different dialects are obtained from Altibbi telemedicine company and used to train the embedding model. Three different embedding models are developed and compared, which are Word2Vec, fastText, and GloVe. The trained models were evaluated by different criteria, including the word clustering and the similarity of words. Besides, performing a specialty-based question classification. The results show that Word2Vec and fastText capture sufficiently the semantics of text more than GloVe. Hence, they are recommended for healthcare NLP-based applications.

**INDEX TERMS** Arabic, fastText, GloVe, healthcare, pre-trained, word embedding, Word2Vec.

## I. INTRODUCTION

One of the intriguing advances of NLP research is the development of word embedding. Word embedding is a technique of representing text-based data in numerical vectors that act as feature vectors of the words. Such advantages of word embedding are the ability to model the words syntactical attributes, and their semantic relationships in dense, low-dimensional representations [1]. Creating feature vectors of words has made the process of integrating text-based data into machine and deep learning models easier. Hence, several NLP applications have been flourished since then, such as sentiment analysis [2], question answering [3], information retrieval [4], and others [5]. Developing a word embedding model requires training on large datasets, where multiple patterns of representation can be captured. The sources of data for training such embedding model can be classified into external and internal. The former is concerned with building models for general purposes from general data sources, and also for a specific language. In contrast, the latter attempts to build models for a certain domain, where the data is just related to a specific type of knowledge [6]. Learning and creating word embedding can be performed either by training a new word embedding or using a pre-trained model. Arguably, this is sparked controversy in the NLP research community, especially when considering a specialized research area such as the medical and healthcare fields, due to mainly a lack of reproducibility [6], [7].

Biomedical and healthcare natural language processing have essential applications in clinical systems, where machine and deep learning methods are de facto approaches for medical, predictive systems, and knowledge extraction.

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

The objective of word embedding models is to evolve and facilitate NLP applications by healthcare workers or researchers in the medical domain in the Arabic context. Building word embedding models to promote the Arabic NLP community is essential. However, it is still immature due to the lack of needed large-scale corpora, especially for domain-specific NLP. To the best of our knowledge, there are no research studies to train word embeddings from domain-specific corpora for biomedical and healthcare NLP applications in the Arabic language. Therefore, this paper presents AltibbiVec, a domain-specific word embedding model for the health and medical-related content in the classical and dialectical Arabic.

The proposed embedding model is built based on approximately 1.5 million medical and health-related consultations and questions, which are obtained from Altibbi's databases. Altibbi [1] is a telemedicine company that provides the MENA region with simplified medical and health knowledge alongside telehealth services. The proposed AltibbiVec embedding models are developed based on three well-known word embedding techniques: Word2Vec [8], fastText [9], and GloVe [10]. The efficacy of the proposed embedding models is assessed based on different evaluation measures. The quantitative evaluation measures the quality of the embeddings by utilizing them in different NLP use-cases, such as question answering, named entity recognition, part-of-speech tagging, and many other supervised and unsupervised applications. In this paper, the embedding is assessed quantitatively by taking the text classification as a use case. Typically, the text classification is a specialty-based question classification that is performed using the bidirectional long short-term memory (BiLSTM) [11] network. Additionally, other measures assess the quality of the proposed word embeddings in encoding variants of syntactical and semantic features, this includes the words clustering based on their similarity and synonyms detection. Accordingly, the embedding models have shown promising performance in capturing the syntax and semantic features. Therefore, the best developed model will be applied in different use cases at Altibbi, such as the recommendation system and the search engine optimization. Figure 1 shows a conceptual representation of the proposed embeddings.

The rest of the article is structured into 8 sections. Section II presents the related works for word embeddings in the Arabic context. Section III represents the theory of used methods, i.e. Word2Vec, fastText, GloVe, and BiLSTM. Section IV provides the designed methodology, including the used dataset, models' development and experimental setup, and the evaluation criteria. Section V discusses the experimental results, while Section VI concludes the study remarks and points out the potential future work of expanding this research.

## II. RELATED WORKS

Distributed word representations had a considerable influence on the performance of learning models in NLP applications. Few studies are devoted to advance the research of NLP in the Arabic context. However, there is less awareness of biomedical NLP in the Arabic context. Generally, this section presents recent research studies in medical and health-related NLP, especially in the Arabic context.

One of the early implementations that consider the Arabic language is the "Polyglot" [12], which is a distributed word representation for multilingual language processing. It trained by using data from Wikipedia for one hundred languages. The authors claimed that it is competitive with state-of-the-art models regarding its performance. However, it was trained on general data from Wikipedia and not specifically for the medical domain. Besides, it is relatively old. Afterward, more robust embedding models were proposed to capture the semantic of words, such as the neural-based word embedding. In [13], the authors proposed a pre-trained word embedding model in the Arabic context using the Word2Vec model. The trained models are general, distributed word embeddings based on text-based data collected fundamentally from the Internet and social media platforms and are not devoted to a specific domain. In [14], the authors created a sentiment classification approach using different feature extraction methods, including word embedding. In their study, they collected a dataset from eight Arabic newsletters with 1.5 billion words. In essence, the performance results increased from 85% to 92%. Even that, the authors did not create a pre-trained public model.

Further, Grave *et al.* [15] proposed multilingual distributed word representations for 157 languages, including the Arabic language. The training of the embeddings relied on data from Wikipedia and Common Crawl project. Besides, the authors evaluated the proposed embeddings quantitatively in ten languages, for which the evaluation datasets were available. However, the objective was for generic domain embeddings. Additionally, Lachraf *et al.* [16] has proposed an Arabic-English word embedding model that trained on approximately 94 million sentences from the "Open Parallel Corpus Project". The results of the intrinsic and extrinsic evaluation demonstrated good performance. Where extrinsically, it achieved a correlation rate of 75.5% when the skip-gram (SG) and the random shuffling were used. Even though, the proposed embedding cannot adapt to the medical domain. Fouad *et al.* [17] designed "ArWordVec" that is a pre-trained word embedding model in the Arabic context. It is based on Twitter data and has shown better word similarity scores than the previous pre-trained models. Even though, the results of the extrinsic evaluation were not very satisfactory in comparison with other state-of-the-art models.

Distributed word embedding is fundamental for performing various NLP tasks. Meanwhile, medical NLP is an emerging research community, where more attention is needed toward developing such biomedical word representations. In the literature, there are several research studies for medical word embedding models targeting the context of the English language. For example, Sh *et al.* [18] created biomedical embeddings depending on health-based reviews and the Word2Vec model, where the proposed approach evaluated

[1] https://www.altibbi.com/

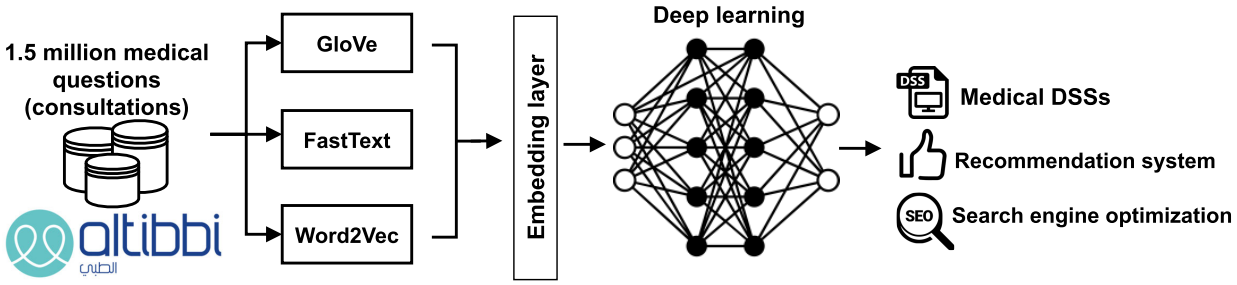



**FIGURE 1.** An abstract representation of the proposed embeddings.

extrinsically by implementing a named entity recognition task. Shen *et al.* [19] developed a Drug2vec word embedding model based on Word2Vec and depending on three types of data sources: pharmacological, and drug taxonomy information, in addition to textual descriptions. The developed Drug2vec is used for drug-drug interactions and similarity prediction. However, it was developed to model merely phrases in the English context. In [20], biomedical word vectors were created (denoted by "BioWordVec") that include subword information based on fastText. The proposed embeddings were evaluated based on several NLP downstream tasks. Also, it proposed to serve NLP in the English context. In another paper, Chen *et al.* [21] proposed "BioSentVec" that is a sentence embedding model for biomedical texts. These embeddings were trained using academic articles and clinical notes (i.e., PubMed, MeSH, and MIMIC-III) by utilizing the Sent2Vec model, which is assessed based on two similarity tasks. Also, Huang *et al.* [22] proposed "ClinicalBERT", a deep neural model for learning word representations of clinical texts to predict 30 days hospital readmission depending on discharge summaries. The ClinicalBERT model obtained better performance than the BiLSTM in terms of recall, as well as a higher correlation rate than fastText, and Word2Vec. However, it was developed for the English context. Also, Lee *et al.* [23] created a pre-trained biomedical word representation (known by "BioBERT") based on bidirectional encoders from transformers. The "BioBERT" model was evaluated on different tasks, including the biomedical named entity recognition and relation extraction as well as the question answering. Accordingly, "BioBERT" achieved 62% of f1-score in the named entity recognition, 86.51% of f1-score in the relation extraction, and 47.82% of Lenient accuracy in the question-answering task. Even though, the proposed "$BioBERT_{base}$" and "$BioBERT_{large}$" were just trained on PubMed abstracts.

To this end, it is noticeable that there is a lack of medical word embedding models in the Arabic context. However, Faris *et al.* [24] proposed an automatic medical question classification method by utilizing evolutionary-based support vector machines. In the proposed model, a word representation model was constructed depending on the term frequency-inverse document frequency by using Arabic medical consultations curated from Altibbi. The created model could achieve 85% of accuracy in classifying the medical questions into fifteen classes. Hence, the massive need for medical embeddings in Arabic has fostered our efforts for developing such a pre-trained medical and health-related embedding model.

## III. BACKGROUND

This section describes various embedding models, including the Word2Vec, fastText, and GloVe. Also, the theory of a variant of a recurrent neural network: the bidirectional LSTM.

### A. Word2Vec EMBEDDING

Word2Vec is a dense, low-dimensional text representation method, where each word (token) is represented by a real-valued vector. It was proposed by Mikolov for generating word embeddings, for which, the similar words that appear in similar contexts have similar representations. Word2Vec uses a neural network for creating the tokens' vectors. Its input is a set of unique tokens extracted from the input dataset, the hidden layers consist of several hidden neurons that form the embedding dimension and are activated by the Softmax function. Whereas, the output is of $n$ neurons, where $n$ equals the number of the input unique words.

Word2Vec has two structures of learning algorithms: the continuous bag-of-words (CBOW), and the SG. The CBOW predicts a target word from a set of continuously distributed context words. The input words are one-hot encoded vectors of size ($1 \times v$), which are multiplied by a weighting matrix ($v \times N$) and averaged before entering the hidden layer. The $v$ is the number of all tokens, and $N$ is the embedding size, as shown in Figure 2. Since the order of words, in this case, is not considered, it is called a bag-of-words. The output of the hidden layer is multiplied by a weighting matrix ($N \times v$) that encompasses the contextual semantics, to produce the final output vector ($1 \times v$). On the other hand, the SG predicts the context words from a given source word in the same strategy that CBOW does, where it predicts one context word at a time. The CBOW and SG models learn by backpropagation with the objective of minimizing a loss function, e.g. the Softmax function.

### B. fastText

The fastText embedding is a word representation that was developed by Facebook AI research [9]. It is an extension of the CBOW and SG structures of Word2Vec. The idea of

fastText is the use of subwords information. In other words, a bag of character n-grams to create the embeddings, where each word is represented by the sum of its n-gram vectors. Each word is surrounded by two symbols"<" and ">" to separate the affixes of the word from other words. For example, "where" can be represented by <wh, whe, her, ere, re>, this is described by Equation 1. In which, $v_w$ is the word's vector, $G_w$ is the group of n-grams of word $w$, $z_g$ is the vector representation of an n-gram $g$.

$$v_w = \frac{1}{|G_w|} \sum_{g \in G_w} Z_g \tag{1}$$

After creating the vector of the target word $w_t$, the original fastText model implements the skip-gram learning structure, where all context words are considered positive examples, and other sampled words from the corpus are negative examples. Thereby, the loss function is defined by Equation 2. In which, the $M_{t,c}$ is the set of negative examples from the corpus, and s(w(t),w(c)) is described by Equation 3.

$$loss = log(1 + e^{-s(w(t),w(c))}) + \sum_{m \in M_{t,c}} log(1 + e^{s(w(t),m)}) \tag{2}$$

$$s(w(t), w(c)) = \frac{1}{|G_{w(t)}|} \sum_{g \in G_{w(t)}} Z_g^T V_{w(c)} \tag{3}$$

Encompassing the vectors of subword in one vector embedding allows the fastText to encode the morphological features and be very efficient in representing the syntactical attributes more than the semantics. Moreover, it exhibited a better capability in handling rare words and dealing with the case of out of vocabulary [9].

### C. GloVe

The GloVe is an unsupervised and global representation of word embedding, which was created by Pennington from Stanford. The novel idea of GloVe is the ability to encode global statistics into embedding by relying on word co-occurrences and the ratios of co-occurrences probabilities. The GloVe can capture local and global features, since depending only on the global statistics of the dataset results in a weak performance in terms of the word analogy, and depending merely on the local statistics produces embeddings that are influenced only by the surrounding context of such words.

Primarily, the generation of GloVe embedding depends on the creation of the word-word co-occurrence matrix ($X$). Such as $X_{ij}$ denotes the number of times the word $j$ appears in the context of the word $i$. The $X_i = \sum_k X_{ik}$ denotes the number of times that any word occurs in the context of word $i$. The probability of word $j$ to occur in the context of word $i$ is $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$. Hence, GloVe uses the ratio of probabilities to infer the embeddings as in Equation 4. Where $w$ are word vectors, $\tilde{w}$ is distinct context word vectors, and $f$ is a weighting function.

$$f(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{4}$$

To illustrate, the probability of the word "arthritis" to appear with the word "fever" is 0.9, and the probability of "arthritis" to appear with the word "headache" is 0.3. Thus, the ratio of the two probabilities $0.9/0.3 = 3$, which is greater than 1; therefore, it is more likely that "arthritis" will appear with "fever" when considering the GloVe embedding.

### D. BIDIRECTIONAL LSTM (BiLSTM)

BiLSTM [25], [26] is a type of recurrent neural networks that can learn the long-term dependencies. The structure of the BiLSTM encompasses a chain of memory cells, where each cell consists of a cell state and three kinds of neural gates (i.e., input, output, and forget). The gate layers control the flow of information and decide what is relevant to the learning process to stay in the cell state or to be forgotten. LSTM is unidirectional, which means it can only learn from the historical information of the hidden states. Indeed, this limits its capacity to learn just from one direction of the textual context. Whereas, the bidirectional LSTM (BiLSTM), can learn from historical and future contexts, which makes it better in learning sequential data types. Figure 3 shows an overview of the BiLSTM.

## IV. METHODOLOGY

This section explains the procedure for building a word embedding model for healthcare and telemedicine services based on the Arabic language. This includes the description of the utilized data, the development of the approach, the evaluation criteria, and the experimental setup.

### A. DATA DESCRIPTION AND PREPROCESSING

The dataset is from Altibbi company. Creating neural-based embedding models requires analyzing large-scale text-based data to capture significant features that encompass deep implicit relations. Therefore, 1,464,411 unlabeled medical consultations were collected from Altibbi databases for training and learning the representations of medical terms. It is worth noting that another version of the medical consultations is available and labeled by the type of specialty. The consultations have more than 25 specialties. However, having the data labeled and unlabeled extends the potential of use-cases in supervised, unsupervised, and semi-supervised learning. All curated text-based consultations were written and asked mostly in colloquial Arabic with different specialties and dialects. Even that, many examples in the dataset are in the modern standard Arabic (MSA). Table 1 presents examples of such consultations and their translations in English.

The creation of embeddings starts with data collection, data preprocessing, and the generation of distributed representations. The preprocessing of the data is an essential step that affects considerably the quality of the embeddings (i.e., features). Typically, the preprocessing phase includes data

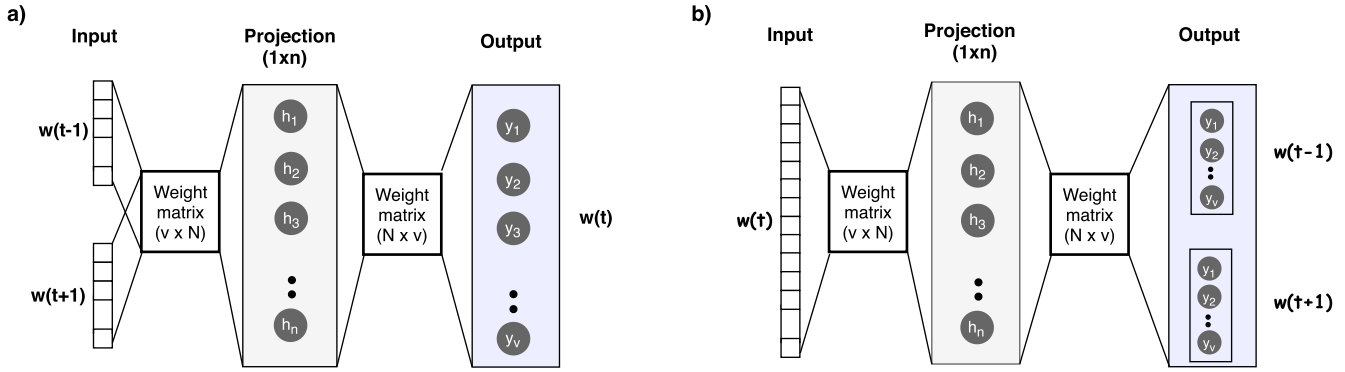



**FIGURE 2.** The CBOW and SG algorithms of Word2Vec, subfigure (a) shows the CBOW, and subfigure (b) presents the SG.

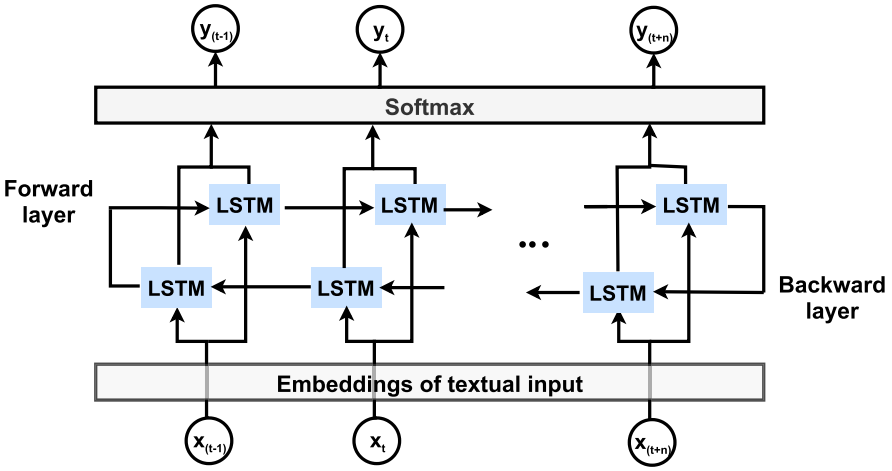



**FIGURE 3.** A description of BiLSTM structure, and classification framework, where *t* denotes a time step.

cleaning and feature extraction. In the case of text-based data, the curated data from the Internet demands further cleaning stages, including stopwords elimination, the removal of articles, prepositions, punctuation, symbols, and numbers. Also, the normalization or denoising, which refers to changing the writing format of the ''Alef'' (A) character to be in one consistent form (i.e., ''أ'' will be ''ا''). In addition, it includes the elimination of diacritics.

Then, the preprocessed questions were stemmed from the ISRIStemmer [27]. Besides, tokenized and padded by the length of the longest question, which is 502 tokens. Even though there are a variety of stemmer tools designed to process Arabic, such as Khoja [28], and Madamira [29], but the ISRIStemmer is still able to perform well. For instance, in [30], the authors compared various stemmers on the Saudi Dialectal Arabic, where the ISRIStemmer was performing the best. Similarly is in [31], where ISRIStemmer was outperforming. Whereas, in [32], [33], other stemmers were showing outperforming results. In this context and for simplicity, the ISRIStemmer is utilized.

### *B. MODEL DEVELOPMENT*

Learning embedding models is an unsupervised process. Around 1.5 million consultations were utilized for training three different embedding models, where the consultations are categorized into fifteen classes (specialty types). Figure 4 shows a general overview of the proposed methodology. Essentially, the collected consultations were cleaned, stemmed, and tokenized to be ready for the embedding modeling. The embedding models were generated based on three different embedding algorithms (i.e., the Word2Vec, fastText, and GloVe). Hence, since the extracted embeddings represent potential features, they were evaluated by using extrinsic and intrinsic methods. Extrinsically, the embedding models are evaluated by performing a text classification problem. The objective of text classification is to classify the set of consultations into suitable specialty types, where they can be classified broadly into 15 types. The specialities are ''Diabetes'', ''Child Health'', ''Ear, Nose & Throat'', ''Dental Medicine'', ''Nutrition'', ''Ophthalmology Eye Diseases'', ''Dermatology'', ''Heart Disease'', ''Heart Disease'', ''Tumors'', ''Psychiatric Diseases'', ''Urology & Venereology'', ''Digestive System Diseases'', ''Musculoskeletal Diseases'', ''Sexual Health'', and ''Gynecology & Women Diseases''. Moreover, they were evaluated intrinsically by performing word clustering and word similarity. The word clustering aims to categorize the similar words together based on their embeddings and using the k-means algorithm.

**TABLE 1.** Examples of altibbi received questions translated into English.

| Question | Translated Question |
|---|---|
| انا فتاة عمرى ١٨ سنة عندى ضعف عام وانميا واميبا علما بانى نحيفةجدا فماذا افعل حتى يزيد وزنى؟؟ | I am an 18 years old girl I have generalized weakness anemia and amoeba knowing that I am very thin so what should I do to increase my weight ?? |
| ماهي حقيقة علاج السكري بالخلايا الجذعية وهل زراعة الخلايا الجذعية تسبب السرطان؟ | What is the reality of treating diabetes with stem cells and does stem cell transplantation cause cancer? |
| ما انواع الطعام التي تحتوي الحديد بنسبه كبيرة؟ | What types of food contain iron in a large percentage? |
| وقعت علي رجلي وورمت وزرقا جدا ودكتور قالي في وريد نزف .. دلوقتي هيا حرقاني وبتزرق اكتر وانا خايفه جدا | I fell on my leg, it is swollen and very blue-colored, and my doctor said there is a vein is bleeding .. Now, I feel it burning and it is becoming more in blue, and I'm very afraid. |

The results of word clustering were visualized using the t-Distributed Stochastic Neighbor Embedding (t-SNE) [34], which drops the high-dimensionality of the embeddings into two-dimensional embeddings for more tangible representations. Furthermore, the generated clusters were evaluated using the purity and the v-measure. Besides, they were evaluated based on their similarity by calculating the cosine similarity score.

The training of each model (i.e., Word2Vec, fastText, and GloVe) corresponds to different configuration settings. In the case of Word2Vec, the NLTK library [35] was used to tokenize the words, where the generated tokens were fed into the Word2Vec model from the Gensim library [36] to create the embeddings. Primarily, the skip-gram training model was considered, where the *window* size was set to three. The window parameter specifies the maximum distance between the current token and the predicted one. The maximum number of epochs (*iter*) was five, and the $min_count$ was three to ignore the words of frequency less than that. Regarding fastText, the CBOW model was used. For which, the minimum count *minCount* was three, the size of the context window *ws* was set to three, and the number of training epochs was five. Whereas, for the GloVe model, the NLTK library was used for the tokenization, where the number of epochs was set to five.

The generation of embeddings as representational features is an integral phase toward utilizing the learned embeddings into algorithms that learn and perform specific tasks. The proposed medical embeddings were used to build a medical specialty classification model. The objective is to automate the routing process of questions towards the correct doctors based on their specialties. Hence, to build the model, the utilized dataset consists of a set of consultations and their respective labels (i.e., the specialty type). Thus, the BiLSTM classifier is utilized to identify the specialty type of consultations. The structure of the BiLSTM classifier involves an embedding layer. The BiLSTM units, which are 30 with a dropout rate of 20%, and recurrent-dropout rate of also 20%, while the activation function is the Sigmoid "tanh" function. The parameters were set based on a previous study [37]. Also, a fully connected layer with a softmax activation. The problem of the question classification can be represented as a mapping module that receives questions and maps them to labels from a set of different fifteen classes of medical specialties.

### C. EVALUATION STRATEGY

The quality of word embedding models can be assessed by qualitative and quantitative approaches. Qualitative methods express the capacity of the embedding model in encoding the semantics and syntactic features of words. The quantitative methods reflect the ability to use the embeddings as features in supervised machine learning tasks, which are assessed by various evaluation metrics (e.g., accuracy and precision).

In this paper, two different assessing criteria are used: the words clustering, and the similarity of the words. Typically, such measures demonstrate the relatedness and coherence of created embeddings. Words categorization or clustering is a quality evaluation measure that groups similar words in clusters. Words clustering indicates the capacity of an embedding model in representing similar words that appear in the same context in similar vectors. The k-means algorithm [38] is a well-known machine learning method for unsupervised clustering. In a repetitive process, it places the nearest points to a known (pre-computed) center together in one cluster. However, visualizing such high-dimensional embeddings in clusters requires reducing their dimensionality. A popular method is the t-SNE, which is a non-linear dimensionality reduction method that projects similar points that are close to each other in a two or three-dimensional space.

The word similarity is measured by the cosine similarity, which is the angular distance between two vectors of the same length. It is defined as the inner-product, as given by Equation 5. Where $\theta$ is the enclosed angle between a and b, $(a.b)$ corresponds to the dot product and equals to $\sum_{i=0}^{m}(a_i \times b_i)$, while $||a||$ is the Euclidean norm of vector $\overrightarrow{a}$ and equals to $\sqrt{\sum a_i^2}$.

$$cosine\ similarity(\overrightarrow{a}, \overrightarrow{b}) = cos(\theta) = \frac{a.b}{||a||.||b||} \quad (5)$$

Smaller angles mean a higher similarity score of the words, while having the angle $(\theta) \geq 90$ means no similarity.

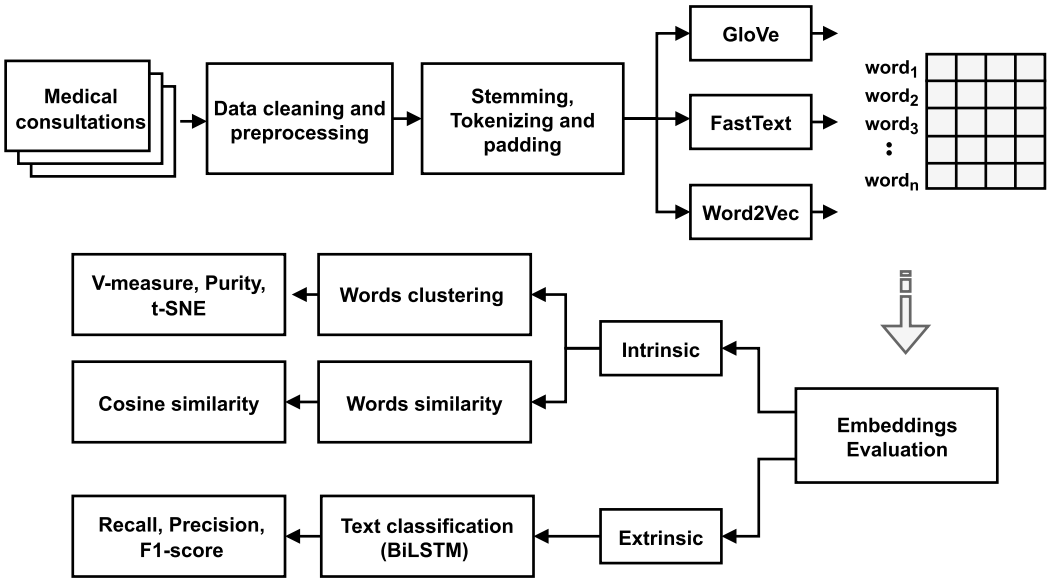



**FIGURE 4.** An overview of the structured methodology.

Quantitatively, the evaluation of the trained embeddings is carried out by implementing and evaluating a use-case of question classification, where the questions were classified depending on the specialty type. The question classification is performed by the BiLSTM neural classifier. The final evaluation of the question classification depends on three performance measures: precision, recall, and f1-score.

The macro-precision ($Precision_m$) calculates the average precision throughout all classes. In this case, the precision is defined by the proportion of correctly identified positive questions over the actual number of positive questions (as in Equation 6). Where ''positive'' corresponds to the class of interest.

$$Precision_m = \frac{1}{|L|}\sum_{l\in L} P(y_l, \hat{y}_l), \quad P(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|y_l|} \tag{6}$$

The macro-recall ($Recall_m$) is the mean of the recall of each class. The recall denotes the ability of the model to recognize the instances of the class of interest. The macro-recall is defined by Equation 7. Where, $(L)$ is the number of classes, $(y_l)$ presents the questions that are labeled by class $l$, and $\hat{y}_l$ represents the instances that have true labels.

$$Recall_m = \frac{1}{|L|}\sum_{l\in L} R(y_l, \hat{y}_l), \quad R(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|\hat{y}_l|} \tag{7}$$

The macro f1-score ($F1 - score_m$) represents the unweighted average of the f1-scores of all classes. The F1-score is the harmonic mean of precision and recall, where it represents the level of balance between them. The $F1 - score_m$ is given by Equations. (8-9), where $\beta$ is a weighting parameter.

$$F1 - score_m = \frac{1}{|L|}\sum_{l\in L} F_\beta(y_l, \hat{y}_l) \tag{8}$$

$$F_\beta(y_l, \hat{y}_l) = \left(1 + \beta^2\right) \frac{P(y_l, \hat{y}_l) \times R(y_l, \hat{y}_l)}{\beta^2 P(y_l, \hat{y}_l) + R(y_l, \hat{y}_l)} \tag{9}$$

### D. EXPERIMENTAL SETUP

The experiments are implemented by using Python version (3.7.3) on Ubuntu-1804-bionic-64 cloud server, the RAM is 64 GB, and the processor is Intel(R) Core(TM) i7-7700 with a speed of 3.6 GHz, while the GPU is GeForce GTX 1080 of 8 GB.

The experiments were implemented using Keras deep-learning framework [39] that is developed on top of TensorFlow 2.0 [40]. The constructed BiLSTM classifier is a sequential model of a non-trainable embedding layer, a BiLSTM layer, and a fully connected dense layer of 15 neurons. The number of the BiLSTM units was set to 30 based on previous findings [41], and the dropout was 0.2. The Adam optimizer was used with a learning rate of 0.0001, while the loss was the categorical cross-entropy. The batch size in the case of Word2Vec and GloVe, was 128, for the fastText, it was 512, and the maximum number of epochs was 100. The data was balanced and divided into 60% for training, 20% for validation, and 20% for testing.

The influence of the dimensionality parameter of the embeddings is dramatic, yet interpreting its behavior on the performance of word embeddings is of great importance. The word clustering was implemented at three embedding dimensions (100, 200, and 300) to investigate the performance of the models. Accordingly, based on the best-obtained results, the rest of the experiments were implemented when the dimension is 100.

For the k-means and t-SNE algorithms, they were implemented using the Scikit-learn library [42], where the number of clusters was set to 12, based on a random-drawn sample with known labels. For t-SNE, the number of components was

set to 2, the number of iterations was randomly set to 2500, and the embedding was initialized by the principal component analysis (PCA) method, which is the default criterion.

## V. RESULTS

This section presents the evaluation of the word embeddings depending on the clustering and the similarity of words and based on the question classification that is observed by the macro average of recall, precision, and f1-score.

### A. WORDS CLUSTERING

This subsection investigates the performance of the three embedding models in clustering similar words. Table 2 presents the words clustering using the k-means clustering algorithm, regarding the purity and v-measure. Purity indicates how good is the clustering algorithm in assigning the data points to their correct clusters, having the purity values close to 1 means all points are clustered correctly and assigned to their true class labels. The V-measure [43] is an external entropy-based measure implemented by the Scikit-learn library, which is the harmonic mean between homogeneity and completeness. A higher v-measure denotes a better clustering. From the table, it is clear that Word2Vec at an embedding dimension of 100 performed the best by having a purity of 66.7%, and a v-measure of 73.3%. At dimension (200), even that Word2Vec and fastText obtained the same purity of 59.6% but Word2Vec achieved a better v-measure of (64%). However, at dimension (300), Word2Vec performed the best clustering considering the purity by having 61.4%, while in terms of the v-measure, fastText performed the best by having 66.9%. Overall, the three models showed better abilities in categorizing similar words together at dimension 100. Therefore, the subsequent experiments were executed when the size of the embedding dimension is 100.

In Figure 5-(a), the Word2Vec embeddings were clustered and visualized using t-SNE, where the x-axis shows the first component of t-SNE (Dimension 1), and the y-axis is the second component (Dimension 2). It is clear from the figure that Word2Vec can efficiently cluster correctly most of the data points. For instance, it categorized ''اسنان'' (*Teeth*), ''تقويم'' (*Orthodontics*), ''حشوة'' (*Dentalfilling*) together in a cluster but failed to classify ''اللسان'' (*Tongue*) with them. Moreover, it could group food-related terms that were colored with dark blue in one cluster, but incorrectly classified ''تاجي'' that means coronary with the group of food terms. Also, it classified correctly the coffee ''قهوة'', and the blood pressure ''ضغط'' together in one cluster colored in red. However, it incorrectly classified diabetes with them. In sub-figure (b), the plot shows the algorithm's ability in capturing the synonym words in different dialects. For example, it returned the words ''غرب'' and ''الكريب'' from the Syrian dialect, which means cold. Whereas, sub-figure (c) shows phrases that are syntactically relevant by having different structures, which refer to nouns, adjectives, and verbs.

Figure 6 presents the clusters produced by the fastText model and visualized by the t-SNE, as well as the synonyms and syntactical words. Sub-figure (a) shows the k-means clusters. It is obvious that fastText fails to group correctly most of the terms, in other words, it classified ''تخثر'' that means thrombosis with the group of food terms, which colored in red. Also, it created clusters of one data point at the terms cold ''رشح'', and dryness ''جفاف'' even that they are belonging to other groups. Also, the cluster of the eye and blurry that colored in pink belongs to another cluster, which includes the terms dryness (''جفاف''), eyes (''عيون''), myopia (''قصر''), and aberration (''انحراف''). Sub-figures (b) and (c) show the efficiency of the model in capturing several relevant synonyms and syntactical words.

**TABLE 2.** A comparison of the clustering quality for Word2Vec, fastText, and GloVe in terms of purity and v-measure.

| Model | Measure | Dimension | | |
|---|---|---|---|---|
| | | 100 | 200 | 300 |
| Word2Vec | Purity | **0.667** | **0.596** | **0.614** |
| | V-measure | **0.733** | **0.640** | 0.655 |
| fastText | Purity | 0.632 | **0.596** | 0.596 |
| | V-measure | 0.664 | 0.622 | **0.669** |
| GloVe | Purity | 0.544 | 0.544 | 0.474 |
| | V-measure | 0.598 | 0.593 | 0.504 |

Furthermore, Figure 7 represents the word clusters of the GloVe model, in addition to the synonyms and syntactical words. Sub-figure (a) demonstrates the created clusters based on the GloVe embeddings. It is apparent that GloVe poorly categorizes the terms into their correct clusters. It failed to group ''نزيف'' that colored in dark purple, ''ضباب'' (with pink color), and ''تخثر'' (in dark blue) to their correct clusters, thus, it considered each of them as a single cluster. For the synonyms and syntactical words, GloVe can represent various phrases correctly as demonstrated in sub-figures (b) and (c).

To sum up, all of the embedding models can perform well in clustering similar words, however, the Word2Vec model obtained the best results among fastText and GloVe.

### B. WORDS SIMILARITY

Figure 8 shows the tenth similar words of ''دوار'', which means dizziness at the three embedding models (i.e., Word2Vec, fastText, and GloVe) when the embedding dimension is 100. Regarding the Word2Vec, the similar words are at the left, while at the right are the similarity scores. It is noticeable that Word2Vec can capture semantically and syntactically similar words. For instance, semantically, it returned ''دوخة'', ''صداع'', ''غثيان'', ''اتزان'', ''نعاس'', where ''دوخة'' had the highest similarity score (83.7%). Whereas, syntactically, it considered ''ودوار'', ''بدوار'', and ''دوران'' as the most relevant words. On the other hand, fastText could capture efficiently the syntactically similar words, where ''فدوار'' obtained the highest similarity score (94.1%). However, fastText failed to catch semantically similar words.

Regarding the GloVe model, it is clear that it had higher ability in representing the similar words semantically more than syntactically also with high similarity scores.

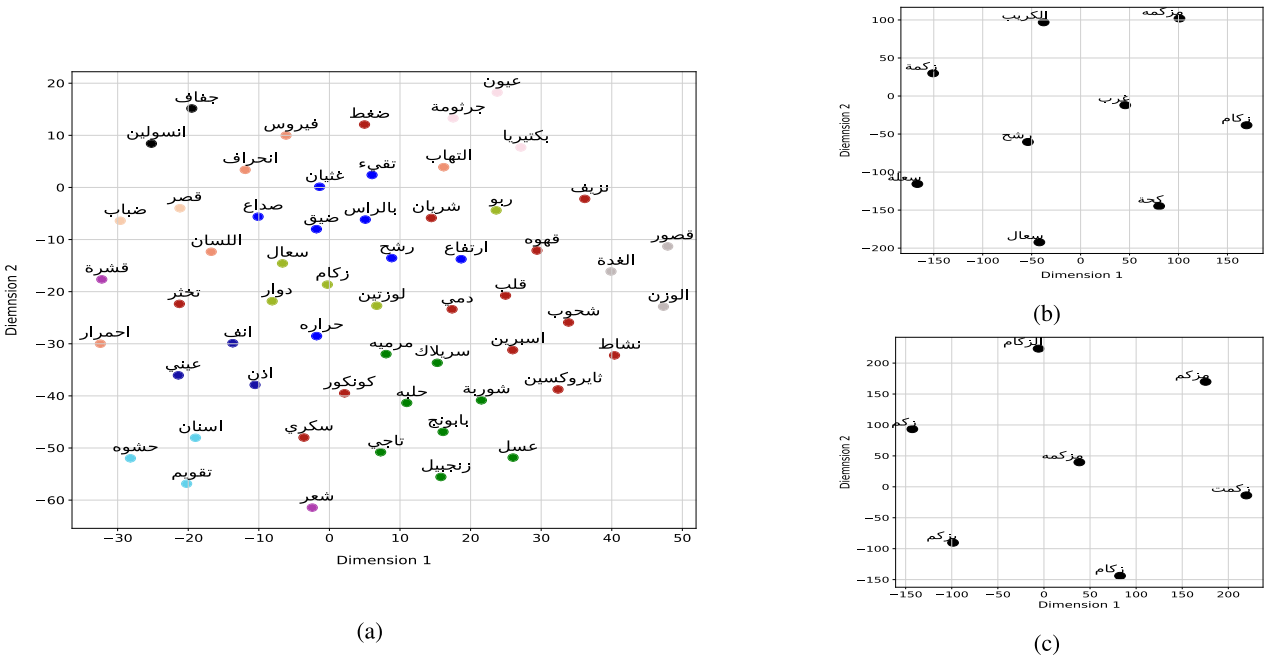



**FIGURE 5.** (a) The created clusters of Word2Vec (b) synonym words (c) syntactical words.

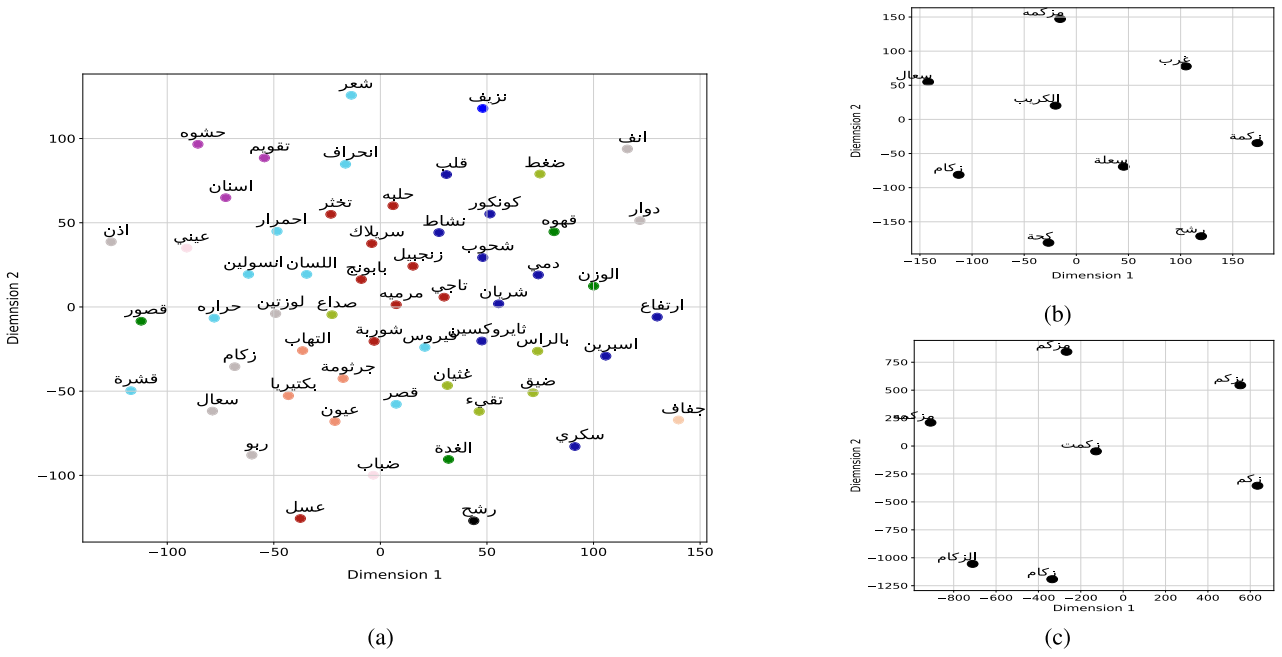



**FIGURE 6.** (a) The clusters produced by fastText (b) fastText synonym words (c) fastText syntactical words.

In other words, it could capture "دوخة" (90.9%), "دوخه" (90.8%), "صداع" (90.3%), "ودوخه" (89.1%), "غثيان" (88.7%), "راس بال-" (87.8%), "ودوخه" (87.1%), and "وصداع" (86.5%).

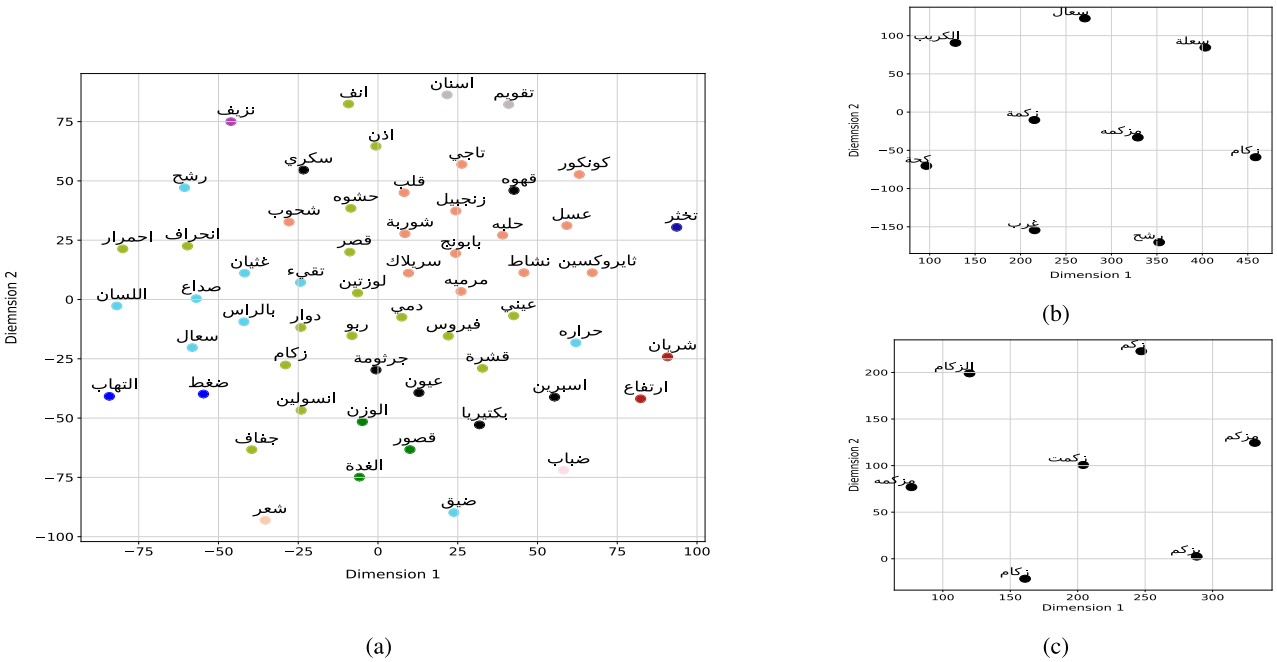
Figure 9 shows a depiction of the similarity of the most commonly used medical terms in the most actively used specialties, i.e. gynecology, dermatology, psychology, internist diseases, and urology. As well as, the specialty of the respiratory system (Ear/Nose and throat). The similarity is calculated based on the Euclidean distance between the tokens' vectors. The t-SNE method is used to reduce the dimensionality of the vectors to two for visualization. In the figure, it can be seen that Word2Vec categorizes the tokens based on the specialty in a clearer way more than fastText, and GloVe as in Figure 9-(a). In sub-figure (a), tokens related to dermatology, psychology, and respiratory system are clustered closely, whereas, tokens related to urology, gynecology, and internist diseases are overlapping since

**TABLE 3.** Medical symptom words in Arabic and translated to English.

| Specialty | Medical Term (Arabic) | Medical Term (English) |
|---|---|---|
| Gynecologist | تكيس | Polycystic |
| | عقم | Infertility |
| | تليف | Fibrosis |
| | داء سكري | Diabetes |
| | قصور | Cervical Insufficiency |
| Dermatology | داء الصدفية | Psoriasis |
| | جرب | Scabies |
| | بثور | Warts |
| | بهاق | Vitiligo |
| | أكزيما | Eczema |
| Psychiatrist | الانتحار | Suicide |
| | الاكتئاب | Depression |
| | الانفصام | Schizophrenia |
| | الهلوسة | Hallucinations |
| | الذهان | Psychosis |
| Urology | البواسير | Hemorrhoids |
| | حصوة المرارة | Gallstone |
| | الإمساك | Constipation |
| | الدوالي | Varicose |
| | السرطان | Cancer |
| Internist | حازوقة | Hiccup |
| | انتفاخ البطن | Flatulence |
| | عسر الهضم | Indigestion |
| | فتاق | Hernia |
| | الارتجاع | Reflux |
| Ear/ Nose Throat Diseases | زكام | Cold |
| | رشح | Runny |
| | الانفلونزا | Flu |
| | الربو | Asthma |
| | الجيوب الأنفية | Nasal Sinuses |

**TABLE 4.** The precision, recall, and f1-score results of the BiLSTM (30 units) for Word2Vec, fastText, and GloVe across all speciality classes, when the dimension size is 100.

| The class of speciality | Word2Vec | | | fastText | | | Glove | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\text{Precision}_m$ | $\text{Recall}_m$ | $\text{F1-score}_m$ | $\text{Precision}_m$ | $\text{Recall}_m$ | $\text{F1-score}_m$ | $\text{Precision}_m$ | $\text{Recall}_m$ | $\text{F1-score}_m$ |
| Diabetes | **0.880** | 0.880 | 0.880 | **0.880** | **0.898** | **0.889** | 0.869 | 0.878 | 0.874 |
| Child Health | 0.773 | 0.840 | 0.805 | **0.800** | 0.848 | **0.823** | 0.707 | **0.855** | 0.774 |
| Ear, Nose & Throat | 0.793 | 0.772 | 0.782 | **0.807** | **0.808** | **0.808** | 0.763 | 0.700 | 0.730 |
| Dental Medicine | 0.920 | 0.878 | 0.899 | **0.928** | **0.896** | **0.912** | 0.741 | 0.721 | 0.731 |
| Nutrition | 0.811 | 0.800 | 0.805 | **0.813** | **0.828** | **0.820** | 0.695 | 0.756 | 0.724 |
| Ophthalmology Eye Diseases | 0.716 | **0.778** | 0.746 | **0.791** | 0.755 | **0.772** | 0.649 | 0.544 | 0.592 |
| Dermatology | **0.712** | 0.740 | 0.726 | 0.689 | **0.774** | **0.729** | 0.542 | 0.642 | 0.588 |
| Heart Disease | 0.793 | 0.792 | 0.792 | **0.805** | **0.829** | **0.817** | 0.753 | 0.767 | 0.760 |
| Tumors | 0.799 | 0.816 | 0.808 | **0.815** | **0.820** | **0.818** | 0.685 | 0.693 | 0.689 |
| Psychiatric Diseases | 0.770 | 0.761 | 0.766 | **0.785** | **0.776** | **0.780** | 0.614 | 0.625 | 0.619 |
| Urology & Venereology | 0.665 | 0.664 | 0.665 | **0.743** | **0.728** | **0.735** | 0.663 | 0.479 | 0.556 |
| Digestive System Diseases | 0.757 | 0.684 | 0.719 | **0.788** | **0.711** | **0.748** | 0.645 | 0.558 | 0.598 |
| Musculoskeletal Diseases | 0.766 | 0.782 | 0.774 | **0.782** | **0.799** | **0.790** | 0.663 | 0.718 | 0.689 |
| Sexual Health | 0.771 | 0.708 | 0.738 | **0.790** | **0.724** | **0.755** | 0.588 | 0.623 | 0.605 |
| Gynecology & Women Diseases | 0.781 | 0.801 | 0.791 | **0.817** | **0.824** | **0.820** | 0.757 | 0.759 | 0.758 |
| **Macro average** | 0.781 | 0.780 | 0.780 | **0.802** | **0.801** | **0.801** | 0.689 | 0.688 | 0.686 |




(a) (b) (c)

**FIGURE 7.** (a) The clusters of the GloVe model (b) the synonym words (c) the syntactical words.

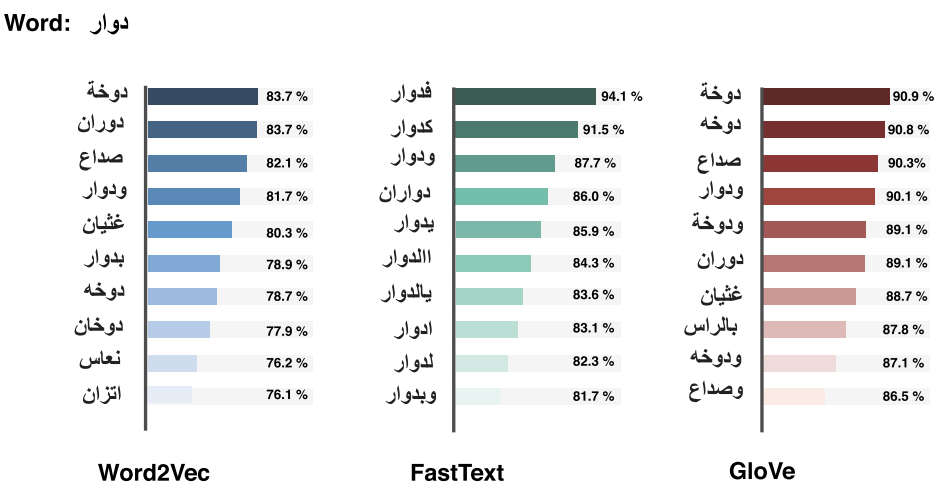


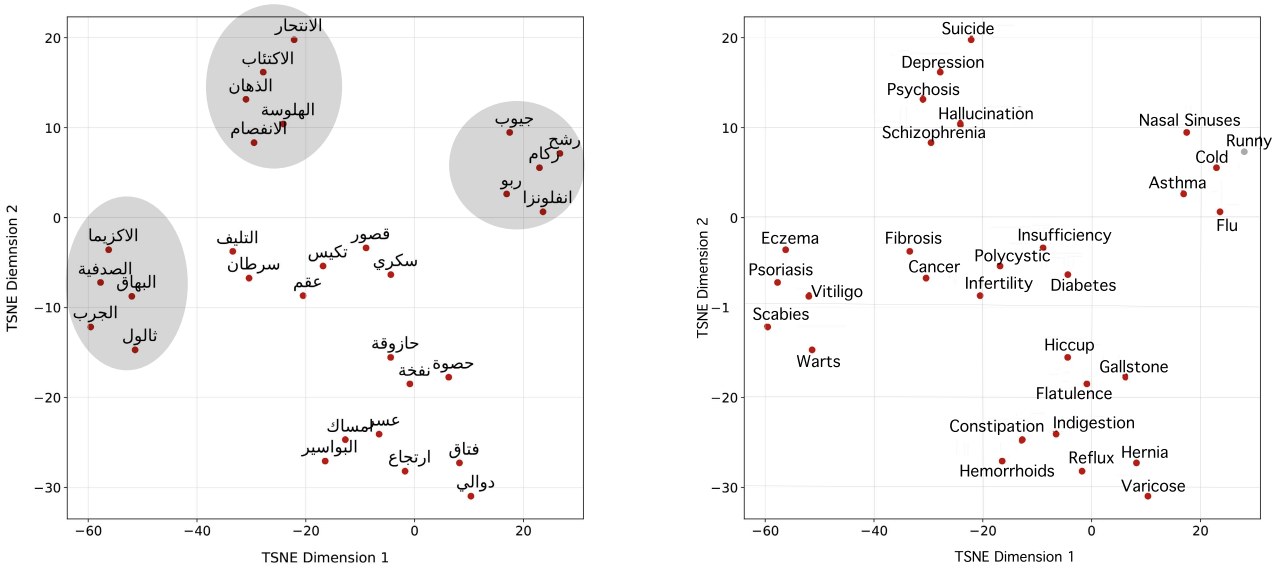

**FIGURE 8.** A comparison of the similar words and similarity scores for Word2Vec, fastText, and GloVe.

they share various symptoms. Figure 9-(b) shows sub-figure (a) translated to English. Moreover, it can be seen that fastText (Figure 9-(c)) can also group some of the context words in dermatology and Ear/Nose and throat specialties. Whereas, GloVe (Figure 9-(d)) fails obviously to catch the contextual similarity. Table 3 shows the used terms in Figure 9 in Arabic and their translation in English.

### C. QUANTITATIVE ANALYSIS

The performance of the three embedding models was compared when the length of the embedding vector is 100. Table 4 presents the precision, recall, and f1-score at Word2Vec, fastText, and GloVe. It is clear from the table that fastText achieved the best results regarding the precision, recall, and f1-score at more than 90% of the classes. Where the macro average of precision, recall, and f1-score were 80.2%, 80.1%, and 80.1%, respectively. fastText achieved the best performance in class ''Dental Medicine'', which obtained 92.8 % in terms of precision. However, it is noticeable that the Word2Vec model accomplished slightly close performance in comparison with fastText by obtaining 78.1%, 78.0%, and 78.0% in terms of precision, recall, and f1-score, respectively. On the other hand, the GloVe model failed to achieve good performance in comparison with Word2Vec and fastText.

Even that word embeddings had revolutionized the textual representations and showed merits over the previously proposed count-based vectorization methods, but distributed




(a) Word2Vec

(b) Word2Vec (English Translation)

(c) fastText

(d) GloVe

**FIGURE 9.** A projection of common medical tokens in Altibbi based on the best obtained models: a) Word2Vec, b) Word2Vec in English translation, c) fastText, and d) GloVe.

word embedding models still suffer from various limitations. Word2Vec, fastText, and GloVe are different methods for creating word embeddings, however, each one has advantages and drawbacks. The Word2Vec model is a window-based model that does not capture information of the whole document, hence, it cannot handle out of vocabulary words. Nonetheless, one of the key drawbacks of Word2Vec is the multi-sense disambiguation, which means that the same word that can be existing in different contexts, will be represented with the same embeddings [44]–[46]. So, Word2Vec poorly captures the contextual information. The problem of the out of vocabulary words can be resolved by fastText or GloVe since fastText encodes the subwords information using the average of the character n-gram, while the GloVe uses global measures of the ratios of co-occurrence probabilities to represent the contextual information. In contrast,

fastText and GloVe cannot solve the disambiguation problem. This opens the directions for additional research studies, where the problem of multi-sense disambiguation was addressed by embeddings from the language models (ELMO), which are deep representational models that mainly tackle the words polysemy problem [47]. However, implementing contextualized word representations (e.g., ELMO) demands large amounts of data [48]. In the case of Altibbi, the number of received consultations has been increased dramatically, where lately it is exhibiting an average growth rate of 58%. Therefore, contextualized word representations are planned to be utilized when larger amounts of data are available.

## VI. CONCLUSION

This paper proposed medical and health-related pre-trained word embedding models in the classical and dialectical Arabic. Medical text-based embeddings are a stepping-stone for various NLP-powered applications in medical or healthcare situations. This paper developed "AltibbiVec" embedding at different dimensions and based on Word2Vec, fastText, and GloVe. AltibbiVec is trained on a massive amount of data collected from Altibbi company. The embedding models were evaluated by relying on different evaluation approaches, including word clustering, word similarity, and synonyms identification. Besides, they were assessed by a question classification task using the BiLSTM classifier, which was evaluated by precision, recall, and f1-score. The proposed embedding models have shown promising performance. Regarding the similarity and clustering of words, the trained models based on Word2Vec or fastText performed the best. Whereas, the fastText based models obtained superior performance in text classification. The objective of developing embedding models is to serve the research community in the medical NLP in Arabic, where they are pre-trained embeddings of words in the medical context. Even that word embedding can perform very well in modeling the semantics, but it fails to encode the contextual information. Therefore, this paper can be extended further by using the bidirectional encoder representations from Transformers (BERT) that reshaped the perspectives of the NLP research community toward a more promising analysis. However, such models require massive amounts of training data. Hence, developing an Arabic medical BERT model is of significant importance, however, their applications are more conceivable when much larger training data is available. Additionally, the model's training and inference times are important to investigate further in future works.

## REFERENCES


[1] Z.-L. Ye and H.-X. Zhao, "Syntactic word embedding based on dependency syntax and polysemous analysis," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 4, pp. 524–535, Apr. 2018.

[2] A. Rajput, "Natural language processing, sentiment analysis, and clinical analytics," in *Innovation in Health Informatics*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 79–97.

[3] Z. Yin, C. Zhang, D. W. Goldberg, and S. Prasad, "An NLP-based question answering framework for spatio-temporal analysis and visualization," in *Proc. 2nd Int. Conf. Geoinform. Data Anal.*, Mar. 2019, pp. 61–65.

[4] M. Parmar, N. Jain, P. Jain, P. J. Sahit, S. Pachpande, S. Singh, and M. Singh, "NLPExplorer: Exploring the universe of NLP papers," in *Proc. Eur. Conf. Inf. Retr.* Lisbon, Portugal: Springer, 2020, pp. 476–480.

[5] D. Suleiman and A. Awajan, "Comparative study of word embeddings models and their usage in Arabic language applications," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2018, pp. 1–7.

[6] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, "A comparison of word embeddings for the biomedical natural language processing," *J. Biomed. Informat.*, vol. 87, pp. 12–20, Nov. 2018.

[7] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *J. Biomed. Informat.*, vol. 100, 2019, Art. no. 100057.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, L. Sutskever, and G. Zweig. *Word2Vec* Accessed: Mar. 22, 2013. [Online]. Available: https://code.google.com/p/word2vec

[9] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.Zip: Compressing text classification models," 2016, *arXiv:1612.03651*. [Online]. Available: http://arxiv.org/abs/1612.03651

[10] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[12] R. Al-Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual NLP," 2013, *arXiv:1307.1662*. [Online]. Available: http://arxiv.org/abs/1307.1662

[13] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic word embedding models for use in Arabic NLP," *Proc. Comput. Sci.*, vol. 117, pp. 256–265, Jan. 2017.

[14] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Improving sentiment analysis in Arabic using word representation," in *Proc. IEEE 2nd Int. Workshop Arabic Derived Script Anal. Recognit. (ASAR)*, Mar. 2018, pp. 13–18.

[15] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–5.

[16] R. Lachraf, Y. Ayachi, A. Abdelali, and D. Schwab, "ArbEngVec: Arabic-English cross-lingual word embedding model," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, 2019, pp. 40–48.

[17] M. M. Fouad, A. Mahany, N. Aljohani, R. A. Abbasi, and S.-U. Hassan, "ArWordVec: Efficient word embedding models for Arabic tweets," *Soft Comput.*, vol. 24, pp. 8061–8068, Jun. 2019.

[18] M. Z. Sh, E. V. Tutubalina, and A. E. Tropsha, "Identifying disease-related expressions in reviews using conditional random fields," *Comput. Linguistics Intell. Technol.*, vol. 1, no. 16, pp. 155–166, 2017.

[19] Y. Shen, K. Yuan, Y. Li, B. Tang, M. Yang, N. Du, and K. Lei, "Drug2 Vec: Knowledge-aware feature-driven method for drug representation learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 757–800.

[20] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Sci. Data*, vol. 6, no. 1, pp. 1–9, Dec. 2019.

[21] Q. Chen, Y. Peng, and Z. Lu, "BioSentVec: Creating sentence embeddings for biomedical texts," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2019, pp. 1–5.

[22] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019, *arXiv:1904.05342*. [Online]. Available: http://arxiv.org/abs/1904.05342

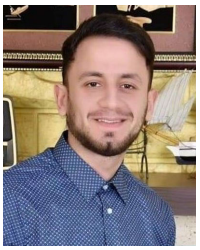
[23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[24] H. Faris, M. Habib, M. Faris, M. Alomari, and A. Alomari, "Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines," *J. Biomed. Informat.*, vol. 109, Sep. 2020, Art. no. 103525.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Jun. 2019.
[27] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in *Proc. Int. Conf. Inf. Technol., Coding Comput. (ITCC)*, vol. 1, 2005, pp. 152–157.
[28] M. N. Al-Kabi, "Towards improving Khoja rule-based Arabic stemmer," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Dec. 2013, pp. 1–6.
[29] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. LREC*, vol. 14, 2014, pp. 1094–1101.
[30] G. Alwakid, T. Osman, and T. Hughes-Roberts, "Towards improved Saudi dialectal Arabic stemming," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Apr. 2019, pp. 1–5.
[31] T. Sakakini, S. Bhat, and P. Viswanath, "Fixing the infix: Unsupervised discovery of root-and-pattern morphology," 2017, *arXiv:1702.02211*. [Online]. Available: http://arxiv.org/abs/1702.02211
[32] A. El Kah and I. Zeroual, "The effects of pre-processing techniques on Arabic text classification," *Int. J.*, vol. 10, no. 1, pp. 1–12, 2021.
[33] M. El-Defrawy, Y. El-Sonbaty, and N. A. Belal, "A rule-based subject-correlated Arabic stemmer," *Arabian J. Sci. Eng.*, vol. 41, no. 8, pp. 2883–2891, Aug. 2016.
[34] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[35] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL Interact. Presentation Sessions*, 2006, pp. 63–70.
[36] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. Workshop New Challenges NLP Frameworks*, Valletta, Malta, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en
[37] H. Faris, M. Habib, M. Faris, A. Alomari, P. A. Castillo, and M. Alomari, "Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: A deep learning approach," *J. Ambient Intell. Hum. Comput.*, vol. 4, pp. 1–17, Mar. 2021.
[38] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.
[39] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io
[40] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, and Z. Chen, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2015, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467
[41] Le, Ho, Lee, and Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, Jul. 2019.
[42] F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, and O. Grisel, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[43] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 410–420.
[44] E. Huang, R. Socher, C. Manning, and A. Ng, "Improving word representations via global context and multiple word prototypes," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2012, pp. 873–882.
[45] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, "Problems with evaluation of word embeddings using word similarity tasks," 2016, *arXiv:1605.02276*. [Online]. Available: http://arxiv.org/abs/1605.02276
[46] T. Ruas, W. Grosky, and A. Aizawa, "Multi-sense embeddings through a word sense disambiguation process," *Expert Syst. Appl.*, vol. 136, pp. 288–303, Mar. 2019.
[47] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[48] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction," 2020, *arXiv:2005.12833*. [Online]. Available: http://arxiv.org/abs/2005.12833

**MARIA HABIB** received the bachelor's degree in computer engineering from the Faculty of Engineering and Technology, The University of Jordan, and the master's degree in web intelligence from the Department of Information Technology, King Abdullah II School of Information Technology. She was a Research Assistant with The University of Jordan. She is currently a Data Scientist and a Researcher at Altibbi, Amman, Jordan. She is a Former Graduate Research Trainee in bioinformatics and big data analysis with the Bioinformatics Lab, supervised by Jianguo (Jeff) Xia at the Parasitology Department, McGill University. She is a member of the (Evo-ML.com) Research Group.



**MOHAMMAD FARIS** graduated (Hons.) in computer information systems from Al Albayt University, Jordan. He is currently a Data Scientist at Altibbi Telemedicine Company. His main technical skills include Python, TensorFlow, Flask, and PHP.

**ALAA ALOMARI** received the B.Sc. degree in in computer science from Yarmouk University, Jordan, and the M.Sc. degree in computer science from Jordan University of Science and Technology, Jordan. He is currently the Chief Information Officer (CIO) and the Product Director of Altibbi (telemedicine platform for MENA region), where he is leading the development, planning, and administration of an innovative, robust, and secure information technology environment throughout the platform and systems. His primary responsibilities for Alomari as a CIO are to setup a technical strategic plan that covers and governs policies, resource allocation, information technology protocols, and security compliances. Prior to working as a CIO, he spent 15 years in information technology roles in managing and administering web-based projects and servers in different areas like cloud communication, NGO, online games, media, online recruitment, and others. Most recently, he is the Technical Director at Unifonic Cloud Communication. He received some other technical and management certificates, like ZCE, CMDBA, and AWS Solution Architect.

**HOSSAM FARIS** received the B.A. degree in computer science from Yarmouk University, Jordan, in 2004, the M.Sc. degree in computer science from Al-Balqa' Applied University, Jordan, in 2008, and the Ph.D. degree in e-business from the University of Salento, Italy, in 2011. In 2016, he worked as a Postdoctoral Researcher with the GeNeura Team, Information and Communication Technologies Research Center (CITIC), University of Granada, Spain. He co-founded The Evolutionary and Machine Learning (Evo-ML.com) Research Group. He is currently the Chief Data Science Officer at Altibbi. He is a Professor with the School of Computing and Informatics, Al Hussein Technical University, and the Information Technology Department, King Abdullah II School for Information Technology, The University of Jordan, Jordan. His research interests include applied computational intelligence, evolutionary computation, knowledge systems, data mining, semantic web, and ontologies. He was awarded a Full-Time Competition-Based Scholarship from the Italian Ministry of Education and Research to pursue his Ph.D. degree in e-business at the University of Salento.