Journal of Ambient Intelligence and Humanized Computing (2022) 13:1811–1827
https://doi.org/10.1007/s12652-021-02948-w

ORIGINAL RESEARCH

Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: a deep learning approach

Hossam Faris[1] · Maria Habib[2] · Mohammad Faris[2] · Alaa Alomari[2] · Pedro A. Castillo[3] · Manal Alomari[2]

Received: 4 August 2020 / Accepted: 1 February 2021 / Published online: 8 March 2021

Abstract
Automated question classification is a fundamental component of automated question-answering systems, which plays a critical role in promoting medical and healthcare services. Developing an automated question classification system depends heavily on natural language processing and data mining techniques. Question classification methods based on classical machine learning techniques face limitations in capturing the hidden relationships of features, as well as, handling complex languages and very large-scale datasets. Therefore, this paper proposes a deep learning approach for question classification, since deep learning methods have the powerful capability to extract implicit, hidden relationships and automatically generate dense representations of features. The proposed question classification model depends on unidirectional and bidirectional long short-term memory networks (LSTM and BiLSTM), which essentially developed to handle the Arabic language in the field of healthcare. The features are represented and created using a domain-specific word embedding model (Word2Vec) that is constructed by training around 1.5 million medical consultations from Altibbi company. Altibbi is a telemedicine company that is used as a case study and a source for curating and collecting the data. The proposed deep learning approach is a multi-class classification algorithm that automatically labels and maps the questions into 15 categories of medical specialities. The proposed deep learning model is evaluated using several evaluation metrics, including accuracy, precision, recall, and F1-score. Markedly, the proposed model achieved a superb classification capacity in terms of classification accuracy rate, which gained 87.2%.

Keywords Altibbi · BiLSTM · Deep learning · Long short-term memory · LSTM · Medical question classification · Word2Vec

✉ Pedro A. Castillo
pacv@ugr.es

Hossam Faris
hossam.faris@ju.edu.jo

Maria Habib
maria.habib@altibbi.com

Mohammad Faris
mohammad.faris@altibbi.com

Alaa Alomari
alaa.alomari@altibbi.com

Manal Alomari
manal@altibbi.com

[1] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan

[2] Altibbi, Amman, Jordan

[3] ETSIIT-CITIC, University of Granada, Granada, Spain

1 Introduction

Traditional medical services struggle to manage the dramatic increase in public health demands worldwide, such as Mayo clinic (Mayo 2020), Florentia clinic (Florentia 2020), and Novomed centers (Novomed 2020) (Longuenesse et al. 2012). Continuously-emerging crises and emergencies present an urgent need for advanced intelligent techniques to manage and meet people's needs in a timely and efficient manner. Artificial intelligence (AI) and data science are an emergent type of science and technology that has a promising future in transforming the present world into a smart world. In recent years, AI has had significant impacts on the digitization of medical services and the promotion of health informatics. AI has been utilized in many applications, including the development of intelligent models that assist clinicians immensely and reinforce the process of

decision-making in a highly accurate, fast, and effortless way.

Text mining is a sub-field of data mining and machine learning and is an essential component of AI due to the intensive, data-oriented nature of the web and digital environments. Text mining is carried out to extract useful information from raw data, which can then be used in advanced learning algorithms to build smart models. These models can learn and perform several tasks, including those related to prediction, classification, and clustering. These tasks can then be modified into useful, real-world applications. Text mining has an important place in the development of automatic question answering systems, where question answering frameworks adopt natural language processing techniques to handle the unstructured textual data and extract implicit features. The automated extracted features aid practitioners' understanding of the underlying semantics of textual questions so that they can extract accurate answers from them. Meanwhile, before a question can be answered, its type must be identified. Recognizing a question's type mainly depends on the type of answer, which is based on the corresponding application (i.e., yes/no, factoid, or summary questions).

Automated medical question answering systems can provide various benefits for healthcare services. For instance, they can save time and efforts of doctors in answering routine questions, save the effort of patients to continuously visit the clinic, yet, provide the doctors the ability to continually screen the patients. Even with the promising advantages of medical question-answering systems, it remains challenging to develop such tools, especially regarding the use of complex languages. This matter has attracted the attention of researchers who have undertaken serious efforts to process and analyze text. In this sense, Arabic is one of the richest morphological languages in the world and is spoken by more than 300 million people (Statista 2020). Dialectical Arabic (DA) is the most commonly used form of the language. This form of the language differs not only from country to country but from city to city within countries. The use of DA on the Internet makes the preprocessing stage of textual data mining extremely challenging. This challenge is primarily due to the considerable differences in the use of words and writing styles. These factors increase the complexity of the learning process performed by classifier algorithms, as they make it more difficult to extract informative syntactic and semantic features.

Medical question classification and answering in the Arabic context have drawn the attention of researchers from the MENA region (Faris et al. 2020). Several studies have addressed the question classification problem by implementing machine learning-based approaches (Faris et al. 2020; Hasan et al. 2018; Ahmed et al. 2017). In machine learning-based question classification methods, natural language processing techniques are used for feature extraction. However, when implementing these approaches, the performance of the classification models is considerably influenced by the set of features that is extracted. In other words, if irrelevant or redundant features are present, or if relevant features are lacking, the performance of the classifier is weakened. Therefore, natural language processing techniques are not optimal approaches for extracting the features and capturing the hidden semantics of a language, particularly the Arabic language. Alternatively, deep learning-based approaches have applied extensively in many research areas, including the medical informatics field.

Owing to the large amount of data produced by online telemedicine services, combined with exceptional advancements in the power of computational resources, deep learning techniques are invaluable resources for classifying medical questions in the Arabic language. Deep learning is a kind of machine learning. The architecture of this learning network involves many neural layers. Deep learning techniques indicate the ability of an algorithm to extract the hierarchical representations of features. They also have a significant capability to automate the processes of learning and producing models that can learn and automatically extract features.

This paper describes an automatic medical question classification method that depends on doctors' specialities. In this work, Altibbi (a telemedicine company)[1] curated and collected labeled medical questions. It is difficult to automatically classify the medical questions asked by patients according to doctors' specialities for several reasons. First, there is a lot of overlap between the characteristics of some specialities, and so the keywords contained in questions also overlap. Second, the Arabic language is very complex and contains many dialects with different phonological and morphological forms, variations in how words are spelled. The traditional approach to classifying questions is to manually label the questions with their corresponding speciality type. However, this wastes much time and effort, especially considering the large number of questions that are received by medical service firms.

This paper describes the creation of a deep learning-based medical question classification approach that automatically classifies the questions fielded by Altibbi into 15 classes based on doctors' specialities. The proposed approach is developed in three stages: cleaning the textual data, extracting sets of features, and training and building the classification model. The feature extraction phase is essential to ensuring a well-performing classifier. Several schemes have been proposed for extracting meaningful numerical features from textual data. Two prominent types of textual feature representation and vectorization are the statistical term

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

frequency (TF) (Luhn 1957) and term frequency-inverse document frequency (TF-IDF) (Jones 1972; Sammut and Webb 2010). The TF-IDF representation weighs the rare words found in a document more heavily than frequent words, which reduces the influence of irrelevant features. However, this type of feature representation is weak and cannot capture the meanings of words or identify relationships among words, which are both essential to gain a useful understanding of questions. Word embedding is an alternative feature representation scheme that is intended to quantify the semantic patterns between words. In our proposed approach, unlike most related previous works, which have utilized pre-trained word embeddings, we develop and train a domain-specific word embedding layer using Word2Vec neural architecture based on around 1.5 million unlabeled consultations collected from Altibbi. The Word2Vec model has been utilized due to several benefits, including the ability of the Word2Vec model to capture the semantics of words and generating denser representations of the features (Mikolov et al. 2013; Liu et al. 2018). The developed word embedding layer is integrated into various deep learning architectures for the deep extraction and representation of features. These models described are variants of the long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) and bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997) methods used in recurrent neural networks. Remarkably, the proposed deep neural approach performed better than previous question classification models employed by Altibbi.

The rest of this paper is divided into seven sections. Section 2 provides a review of related works that have examined the use of deep learning in health informatics and text classification applications, as well as the automated question answering frameworks. Section 3 describes the medical speciality classification problem and motivation. Section 4 describes the collected datasets that were used to build the proposed classification model and the embedding layer; this section also explains the preprocessing steps. Section 5 illustrates the deep learning model that was designed for question classification. Section 6 describes the evaluation measures that were used to assess the performance of the proposed model. Section 7 outlines the experimental settings and the results. Section 8 discusses the limitations and hypotheses. Finally, Sect. 9 concludes and summarizes the findings and provides directions for future work.

2 Literature review

Deep learning algorithms are promising types of neural learning that evolved from artificial neural networks (ANN) (Yegnanarayana 2009) by having many functional neural layers. The power of deep learning algorithms is impressive in many different AI applications. As such, they have gained a profound level of acceptance in the field of health and biomedical informatics. They have been employed by researchers and practitioners in a wide range of medical applications to ease the workload of clinicians and healthcare agents (Ryu et al. 2018; Zhu et al. 2020; Faris et al. 2020). This section reviews the application of deep learning techniques in healthcare frameworks and in medical question classification and answering tasks.

2.1 Deep learning in health informatics

In recent years, AI has advanced significantly in various medical domains, including medical imaging, drug discovery and its effects, clinical diagnosis, computer-aided decision support systems, and online health services (Dash et al. 2020).

An extensive review of the use of deep learning in health informatics was presented in Kwak and Hui (2019). The authors highlighted the primary challenges related to converting the traditional healthcare and medical sector into a smart digitized sector. These challenges include the process of handling the high-dimensionality problem, the non-stationary and temporal nature of data, and the sparsity and heterogeneity. Moreover, they identified obstacles related to model development (e.g., the scalability and security of models). Additionally, Mulani et al. (2020) studied the utilization of a deep reinforcement approach to provide personalized health recommendations. This medical recommendation system performs several tasks (e.g., guiding patients to the correct doctor, suggesting medications making recommendations regarding nutrition and exercise). The authors explained that additional works are needed that encompass more recent deep learning algorithms and that consider the medical histories of patients. Furthermore, Kumar et al. (2020) used the convolutional neural network (CNN) algorithm to detect and classify malaria. Even though the proposed method achieved an accuracy rate of 96.8% by using three stacked convolutional layers and an Adam optimizer, it was not deployed in real-world applications. Akselrod-Ballin et al. (2019) proposed a deep learning-based model for predicting breast cancer by which they linked data from the electronic health records (EHRs) and mammograms. Their approach achieved good accuracy and was recommended by the Assuta Medical Center as a secondary tool for radiologists. However, the authors suggest incorporating more data to increase the model's performance. Shah et al. (2019) proposed a deep learning approach to evaluate the service quality of a healthcare platform based on a fusion of visual and textual data expressing patients opinions. The authors stated that a fusion model can improve the classification accuracy by 12% than relying only on textual opinions. In addition, Vidhya and Shanmugalakshmi (2020) proposed a

deep learning approach using the deep belief networks for the prediction of complications of type-2 diabetes. Where the proposed model achieved an accuracy of 81%.

In another study by Lauritsen et al. (2020), a deep learning model was developed to predict sepsis disease. The detection process relied on the analysis of raw sequences of events of electronic health records, which were collected from several Danish hospitals over seven years. However, the authors cited some challenges related to the lack of interpretability, bias, and reproducibility. Faes et al. (2019) created an automated deep learning approach to classify and diagnose medical images using Google Cloud AutoML. Five public datasets were used to assess the performance of the model, which achieved good accuracy. Further, the authors discussed the advantages and limitations of deploying such a tool. Estrada et al. (2020) built a FatSegNet tool, which is an automated deep learning model that recognizes adipose tissue based on abdominal MRI images. This tool utilizes two layers of 2D dense and fully connected convolutional networks, thus increasing the detection accuracy by 7% over previously proposed algorithms. Moreover, it was subsequently improved to handle 3D MRI images in approximately real-time. Edara et al. (2019) developed a deep learning approach using LSTM for analyzing and predicting the mood of patients who are affected by cancer. The authors conducted a sentiment analysis and categorization of tweets. Based on the findings, the authors stated that patients affected by cancer are more likely to have positive thinking.

Meanwhile, Liu et al. (2019a) explained suitable criteria for managing the narrative textual nature of the EHRs for promoting the clinical natural language processing and decision systems. The authors discussed key issues related to the extraction of useful information from real-world digitized data so it can be integrated into medical research. However, the real-world clinical applications for these criteria have not yet been fully interpreted and explored.

The integration of deep, machine and transfer learning methods has a profound impact on traditional health and medical frameworks. However, Zhang et al. (2019) conducted an extensive survey of the benefits and potential pitfalls of using deep and intelligent learning approaches for medical prognosis in health management systems. The authors also assert that the successful deployment of deep learning algorithms significantly depends on their application and the type of the used data.

2.2 Automated question answering

In Abdallah et al. (2020), the authors developed an automated question answering framework using different recurrent neural network (RNN) and encoder-decoder networks. The authors utilized data from WebMD, HealthTap, eHealthForums, and iCliniq, where the results of the model are very good regarding the BLEU score for evaluating the answers. Vu et al. (2020) proposed a visual question answering system in medical imaging, which fuses images and questions features by utilizing multimodal low-rank bilinear (MLB) model, convolutional neural network (CNN), and a gradient-weighted class activation mapping. Rawat et al. (2020) developed an entity-based deep neural model for question answering in clinical environments. The model proposed by training it on a large electronic medical records dataset, which showed improved performance than state of the art transformers models. In addition, Schmidt et al. (2020) created a transformer-based approach for question classification and answering by using the Stanford question answering dataset (SQuAD). Nonetheless, Ren et al. (2020) developed a deep neural model to automatically answering questionnaires in the medical context. It is developed using three Chinese medical datasets, which showed promising results in terms of f1-score. Liu et al. (2020) proposed a question answering system in the context of the Chinese language. The proposed model is based on the CNN, RNN, and the self-attention mechanism, which is trained by the cMedQA dataset and obtained up to 84% of accuracy. Further, (Mairittha et al. 2020) developed a question answering system based on the electronic health records and depending on variants of transfer learning (BERT) models. The purpose of the framework is to answer questions in discharge summaries by using data from the 2010 i2b2/VA workshop for NLP. The proposed model has shown improved performance than other algorithms.

Very few studies have proposed question-answering systems in the context of the Arabic language. However, Romeo et al. (2019) developed a community question answering system in the Arabic context using the Farasa Arabic textual processing tool for preprocessing and features extraction. However, tree kernels, word embeddings, and LSTM networks exhibited efficient text selection capabilities. Even though, the proposed approach did not target the medical field. Developing question answering frameworks in healthcare and in the Arabic context is essential, yet, poorly explored.

2.3 Deep learning for text classification

Based on our search, we have noticed that there are few papers that have studied the problem of mining textual data and especially textual question classification in healthcare informatics in the context of the Arabic language (Faris et al. 2020). However, several papers have been published in the context of different languages, including mostly the English language (Agrawal and Mishra 2019). Hence, this subsection provides a review of recent studies on medical text and question classification.

Liu and Guo (2019) designed a novel architecture of bidirectional LSTM, attention, and convolution layers for text classification that outperformed other models. In other work, Aydoğan and Karci (2020) interpreted several deep learning models and generated pre-trained word embedding representations for Turkish text classification. However, a multi-class classification problem arose when integrating Word2Vec (including the continuous BoW "CBoW" and skip-gram models) and Glove embedding with LSTM, CNN, RNN, gated recurrent unit (GRU) models among others. The GRU had the best performance. Moreover, Yilmaz and Toklu (2020) introduced an analytical study of deep learning models and word embedding representation for classifying Turkish text. The authors claimed that the different structures of Word2Vec embedding significantly influence the accuracy of various deep learning architectures.

Further, Liu et al. (2019b) created a novel deep learning architecture for an attention-based bidirectional GRU-CNN network for Chinese question classification. The authors constructed word vectors using the CBoW model, as this model was found to obtain better classification results than other models. Remarkably, Kim et al. (2020) examined the capacity of capsule networks to manage question classification tasks. The capsule network was integrated with a novel routing strategy and Glove pre-trained embedding with a dimension of (300). Implementing the experiments using seven benchmark datasets led to very promising and comparable results. Additionally, Jain et al. (2019) proposed a new architecture for binary text classification using a hybrid of bidirectional GRU, bidirectional LSTM, attention-layer, and capsule network layer. The network was trained using an Adam optimizer with a cyclic learning rate and logistic loss. This model achieved an outstanding F1-score of 97.8%. In other research, Zhang et al. (2018) formulated a bidirectional LSTM network for question classification. In their method, words were generated for embedding using part of speech tagging and word position. Interestingly, the proposed approach achieved an accuracy rate of approximately 92%.

Furthermore, Wang et al. (2019) proposed an approach for clinical text classification that utilizes weak-supervision and CNN, in which the weak-supervision strategy employs a rule-based NLP method to automatically label the training data. Compared with other machine learning algorithms, CNN showed superb results. The authors stated that the word embedding presentation performs better than the term frequency-inverse document frequency (TF-IDF). Moreover, Banerjee et al. (2019) developed a deep learning architecture for radiology text report classification. Using Glove word embedding, their architecture assessed the performance of CNN and attention-based hierarchical RNN networks, the latter of which yielded phenomenal results.

It is clear that there is a lack in studying either the text classification or question-answering in the medical or healthcare and the Arabic context. Therefore, the objective of this paper is to create a deep neural-based classification model of medical consultations and the Arabic language.

3 Problem description and motivation

Providing primary healthcare is one of the fundamental services of Altibbi Company for telemedicine services. Altibbi is a digital health platform that provides telehealth and telemedicine services in the MENA region by employing about 2000 medical doctors. In addition, it attempts to disseminate medical content in various forms, including glossaries, encyclopedias, medical articles, and consultations, as well as e-clinics, all in the Arabic context. One of the company's milestones is the "questions & answers" module, which answers patients' questions asynchronously with the aid of specialized doctors. A primary function of the "questions & answers" telehealth service is that it directs patients to the most suitable specialized doctor based on their queries. Traditionally, this process would be carried out by visiting a medical consultant and then a general practitioner, who would finally answer the patient's questions. This process can take up to one day. However, Altibbi encounters an exceptionally large number of questions that need to be answered asynchronously within 24 h. However, manually labeling questions into their corresponding specialities is cumbersome and time-consuming. Moreover, the manual labeling process is not completely accurate, and cannot be accomplished within the required time. There are many similarities and much overlap of keywords among different specialities, which diminishes the accuracy of the manual labeling process (i.e., vision loss might be a phenotype of diabetes or other eye-related diseases). Therefore, developing a more intelligent routing and labeling module is essential for automating the question classification process. Such a module must be capable of processing the natural language and applying advanced learning techniques to recognize informative patterns.

The problem of classifying the questions requires a label c to be applied to each question q that represents a medical speciality. The problem is formulated as a classification function f, the domain of which is a set of textual questions Q and the range of which is the set of potential specialities C. This is expressed in Eq. 1.

$$f : Q \rightarrow C \tag{1}$$

Where Q represents a set of questions of length m ($\{q_1, q_2, \ldots, q_m\}$), and C is the set of all speciality classes ($\{c_1, c_2, \ldots, c_n\}$) of n number of classes. For instance, a question could be "العلوي للعين؟ما علاج جفاف الجفن" (which translates to "What is the treatment of the dry upper

eyelid?"), which is part of the "Ophthalmology & Eye Diseases" speciality. Accordingly, in the case of Altibbi medical data, there are a total of 15 speciality classes, which are as follows: "Diabetes," "Child Health," " Ear, Nose & Throat Problems," "Dental Medicine," "Nutrition," "Ophthalmology & Eye Diseases," "Dermatology," "Heart Diseases," "Tumors," "Psychiatric Diseases," "Urology & Venereology," "Digestive System Diseases," "Musculoskeletal Diseases," "Sexual Health," and "Gynecology & Women Diseases." Thus, each question from the input data should be processed and mapped to one of the 15 specialities.

Handling the questions in the context of the Arabic language is not trivial and faces many challenges. First, the Arabic language has two major forms of the language: the modern standard Arabic (MSA) and the DA. The MSA is the formal use of the language, while the DA is the common language especially on the Internet. However, the DA varies among Arabic countries, yet across cities. Second, the same word might differ in the spelling forms. For instance, the word مدرسه (meaning school) can be written also as مدرسة. Third, some relative pronouns or conjunctions have important meaning for the question, but might be considered as stopwords and removed, which causes a loss of information. For example, the negation word ما can be used also to indicate a question. Moreover, fourth, the verbs might be written in different forms depending on the context if it is singular or plural, feminine or masculine. To illustrate, when referring to a female that she likes traveling, it is written as (هي تحب السفر), while for a male it is written as (هو يحب السفر). Fifth, some names are obtained from adjectives and written in the same form, such as (Sami, Samia) for male, female, respectively that are written as (سامي، سامية).

The language is yet more complex and has plenty of variations grammatically, and morphologically, and further is the variable use of translated and transliterated keywords, especially when talking about diseases and drugs.

4 Dataset descriptions and preparation

4.1 Dataset descriptions

The datasets considered in this work were collected from Altibbi Company for telemedicine services. Two datasets were obtained. The first dataset is unlabeled and used for learning word embeddings, while the second is labeled and used to train, validate, and test the deep learning model. These datasets are described as follows.

- Dataset for learning word embeddings: Word embeddings are very effective in promoting the learning and its efficiency since they are built using predictive models instead of merely using descriptive statistical models. Therefore, they capture a wide range of fine-grained features. However, a huge textual corpus is needed (Li and Yang 2018) to efficiently learn the embeddings as is done in Word2Vec or GloVe (Pennington et al. 2014). Therefore, 1,464,411 unlabeled medical questions were retrieved from Altibbi's databases for training and learning such representations in the medical and health field. All collected textual questions were written in the Arabic language and were related to different speciality types.

  Depending on the created word embeddings, words with similar meanings that appear in similar contexts have similar vectors, and therefore, have a high similarity score (e.g., cosine similarity). For instance, the ten words that are the most similar to "وجع" (meaning pain) according to the cosine similarity are "ألم،" "ووجع،" "وحع،" "االم،" "والم،" "اللم،" "واللم،" "نخز،" "بوجع،" "واجع،" as their similarity scores are 0.788, 0.763, 0.726, 0.695, 0.694, 0.658, 0.653, 0.651, 0.648, and 0.647, respectively. Meanwhile, all of these terms are related to the same meaning.
- Model development dataset: This dataset contains 75,000 medical questions, each of which is grouped by a team of specialized doctors from Altibbi into one of the 15 medical specialities mentioned earlier. The violin plot in Fig. 1 illustrates the distribution and probability density of the 15 medical specialities in terms of their length. The figure shows that all specialities have very similar medians values. However, the shape of the distribution varies between specialities. The lengths of many specialities (e.g., Ear, Nose & Throat problems, and Tumors) are heavily concentrated around the third quartile. Meanwhile, for other specialities (e.g., Nutrition), the opposite is true. Additionally, bimodal distributions are displayed by other specialities (e.g., Diabetes and Dermatology). It can be noticed also that some specialities like Psychiatric Diseases and Sexual Health have more outliers (represented by the outside upper thin line in Fig. 1) in terms of question length than other classes.

4.2 Dataset processing

The first step of the algorithmic design involves data collection, preprocessing, processing, and analysis using learning or mining models. Data preprocessing is a fundamental stage that significantly affects the performance of the entire trained model. The main components of this stage are cleaning and feature extraction. In the text domain, the online retrieved data demands additional cleaning phases, including the removal of stopwords (e.g., articles and prepositions), punctuation, and symbols (including the numerals). The text also needs to be normalized and denoised. Normalization or denoising refer to unifying the writing style of specific characters into one unique form (i.e., "أ" will be "ا"). Figure 2

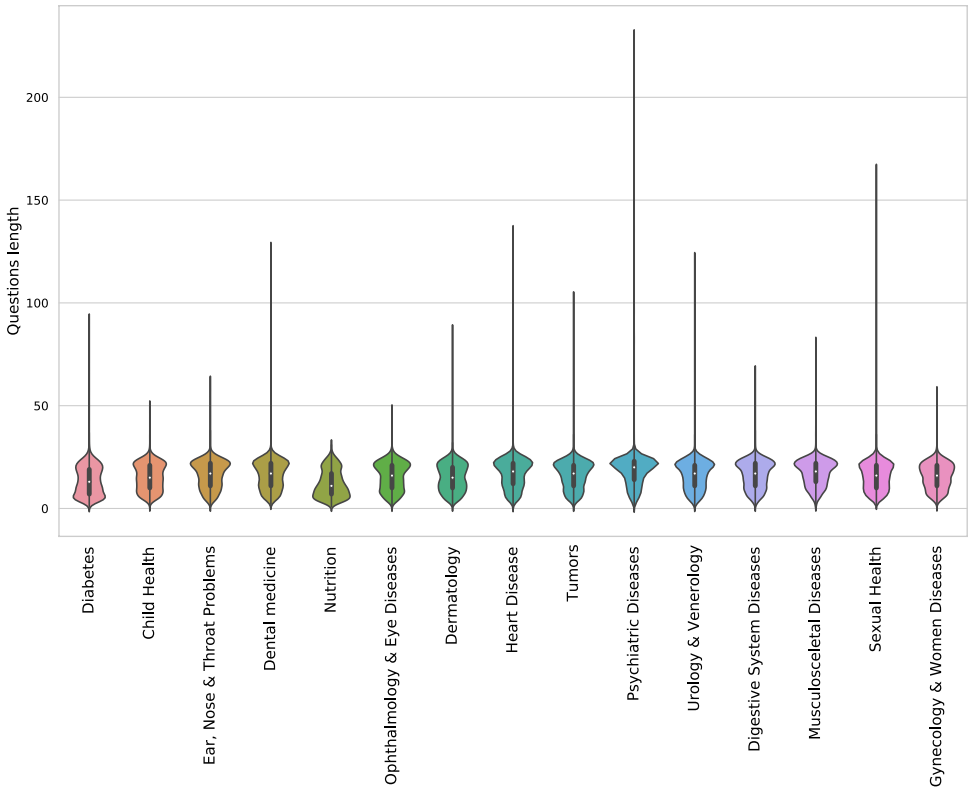

Fig. 1 Violin plots show the distribution of medical specialities/classes in terms of question length

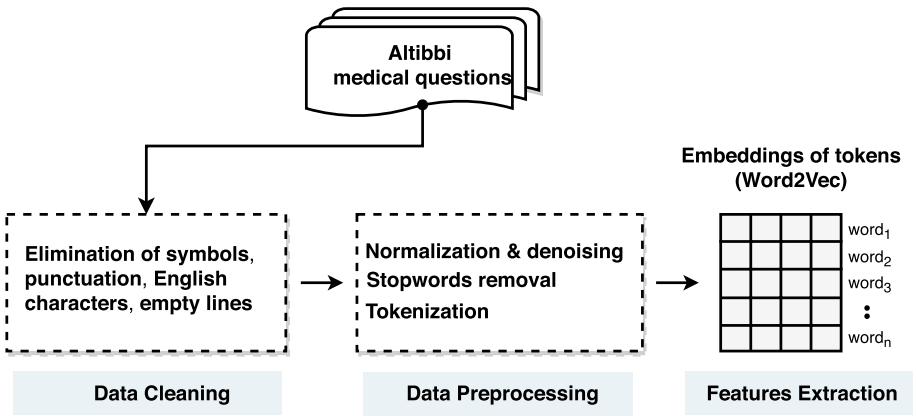

Fig. 2 The fundamental questions' processing stages, including data cleaning, data preprocessing, and feature extraction

summarizes the steps for preparing the collected questions for classification.

At this stage, the cleaned questions consist of only meaningful terms that can be used to identify specialities. These terms are now tokenized, while the tokenized questions are padded to create the embedding matrix. The tokenized questions are padded by the length of the longest question, which is 502 of tokens. The creation of this embedding matrix depends heavily on the learned word embedding process, which is based on 1.5 million questions that comprise 570,883 unique tokens. In the embedding matrix, if a word is not found in the learned word embeddings, then it is initialized to zero. Eventually, the resultant embedding matrix is used to initialize the embedding layer in the deep learning model.

5 Model architecture and procedure

This section describes the deep learning model designed for medical speciality classification. Figure 3 depicts an overview of the question classification approach, which encompasses three sequential modules. The first module is the creation of the embedding layer, which is accomplished based on a massive, unsupervised corpus of 1.5 million medical consultations. The questions of supervised and unsupervised

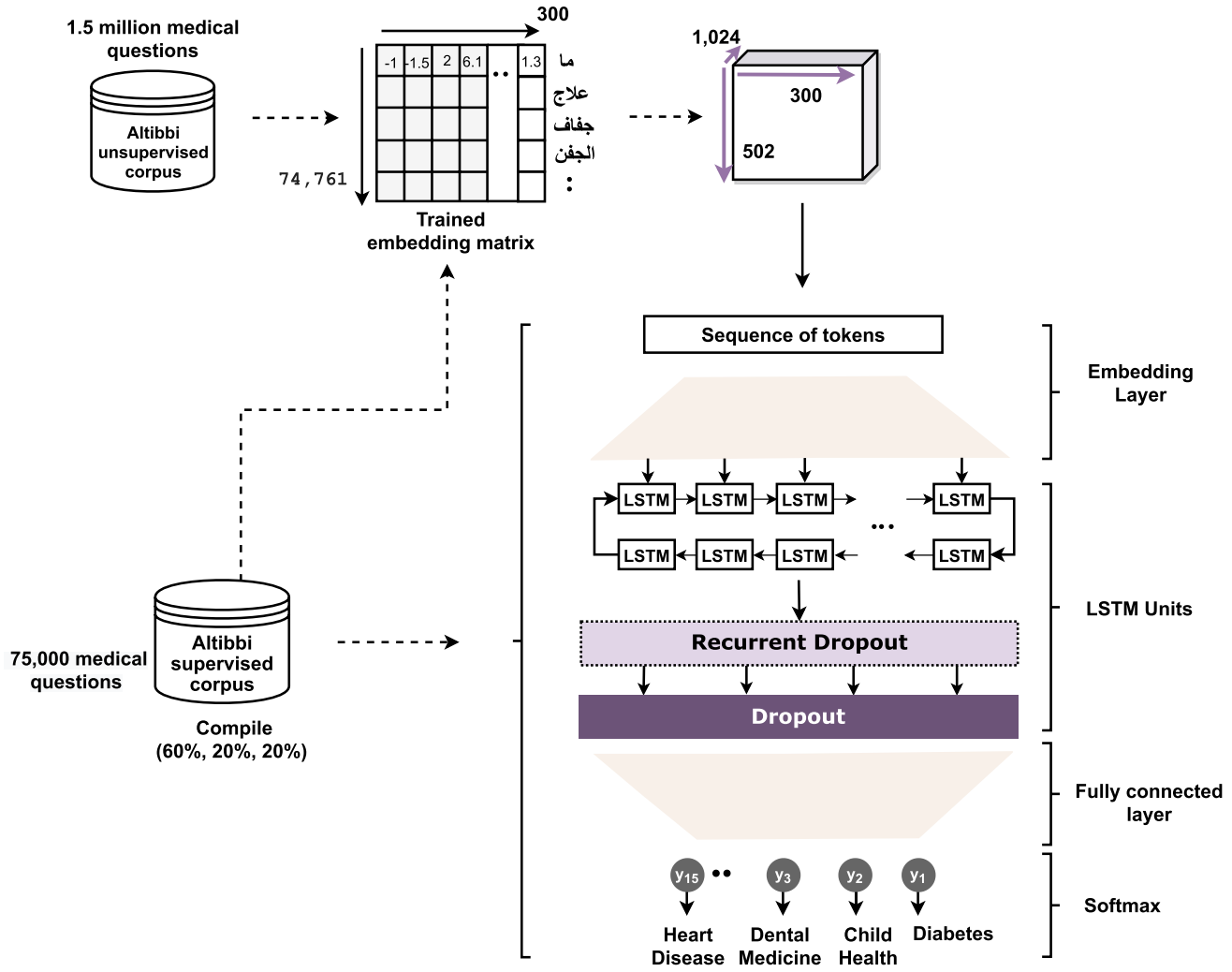

Fig. 3 A schematic representation of the designed deep learning model based on Altibbi learned word embeddings and BiLSTM neural networks

corpora are used to train the embedding matrix, which is then fed into the model in batches. The second module contains the LSTM or BiLSTM deep models, which are compiled and trained using training and validation subsets of data. The third module is a fully connected dense layer with a softmax activation function, which generates a probability distribution of the 15 classes. The details of each module are presented in the following subsections.

5.1 Embedding layer

Machine learning algorithms cannot handle textual data that is represented in a string format unless it encoded in a convenient representation that is understandable by the algorithm. Various text representation methods have been proposed, such as the frequency-based methods (e.g., TF-IDF), and the prediction-based methods (e.g., Word2Vec). In the literature, prediction-based methods have shown superior ability in capturing the meaning of words better more than the frequency-based methods. Representing the textual data numerically is a fundamental preprocessing phase for the learning step.

Word embedding is a word vectorization mechanism by which words are represented in numerical vectors and in which embedding involves encoding the words as dense vectors. For example, a word embedding representation of the word "network" can be illustrated as in the following vector: [0.23, 2, − 1, 0.9, 1.5, − 11.7, 88, 75.5, 1, 0], where the length of the vector is a predefined parameter. Using shallow neural networks (Soltanolkotabi et al. 2019) is one option for creating word embeddings. In these networks, the weights of the hidden layers represent the embedding vectors of words. The word embedding technique has revolutionized the learning process by producing smarter features that express the implicit meanings of texts (i.e., questions). This is achieved by neural networks that

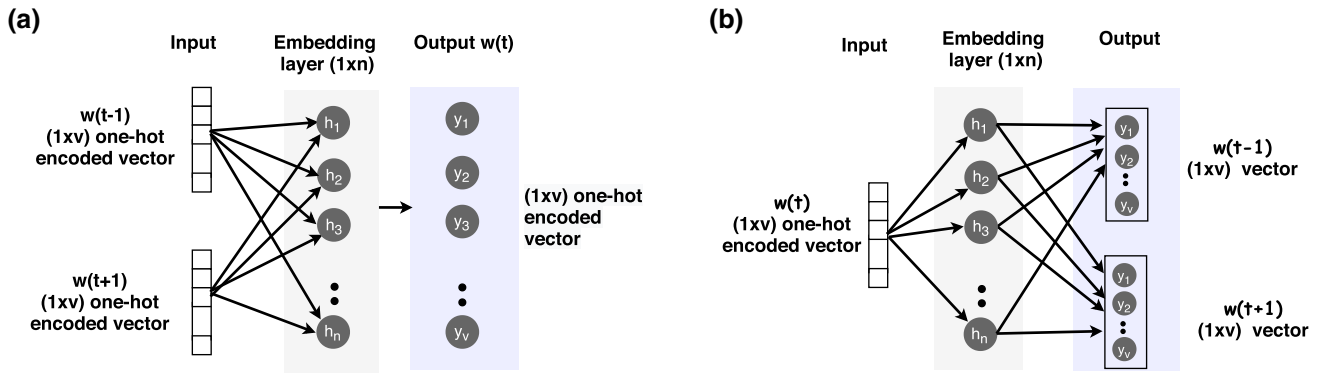

Fig. 4 The **a** CBOW and **b** SG models of Word2Vec

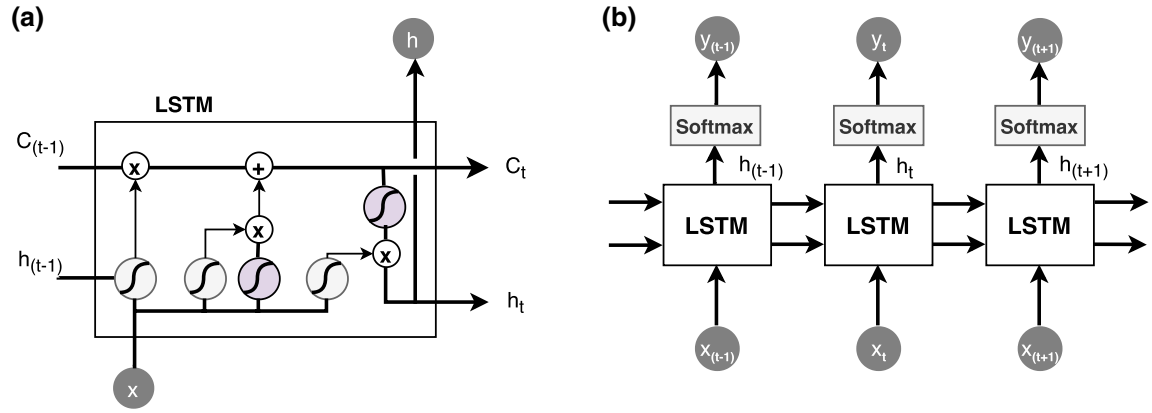

Fig. 5 A depiction of An LSTM neural network, including **a** the internal structure, where the purple node is the "tanh" layer and the grey is the Sigmoid layer, and **b** a chain of LSTM units

can tune embedding vectors so that they are similar when the words that comprise them have similar meanings. In other words, similar words used in the same context will have similar vector representations and, consequently, high similarity scores. Word2Vec is a popular neural model for word embedding. Word2Vec is a two-layer neural network invented by Mikolov et al. (2013) that is designed to create word embeddings. Its inputs are word vectors that represent words' indexes (i.e, words are represented in one-hot encoded vectors). The hidden layer of this shallow network contains several hidden neurons (referred to as the embedding size) that have a softmax activation and that learn using backpropagation. In the output layer, the number of output neurons is the same as the number of input words (neurons).

Conventionally, Word2Vec can be implemented and learned based on one of two structures: continuous bag-of-words (CBOW), and skip-gram (SG) (Fig. 4). In the former, the network learns by considering a set of neighboring (context) words determined by a window-sized parameter for identifying the potential middle word. Differently, the latter learns by depending on one word to predict the surrounding context words.

The SG structure of Word2Vec has various advantages over the CBOW structure. For instance, since the CBOW makes predictions based on context words, it fails to represent infrequent words that are rarely found in a given context. Meanwhile, the SG predicts the context based on a word, and, therefore, it can represent infrequent words much better than the CBOW, especially in the case of the Arabic context (Naili et al. 2017).

5.2 LSTM and BiLSTM layers

LSTM networks are RNN networks that can learn long-term dependencies of data in specific cases, as previously proposed by Hochreiter and Schmidhuber (1997). LSTM involves a chain of connected memory units, with each unit containing a cell state and three neural layers (gates).

Figure 5 shows a typical structure of an LSTM network. The cell state is the horizontal line at the top of the unit where the information flows during the processing stage,

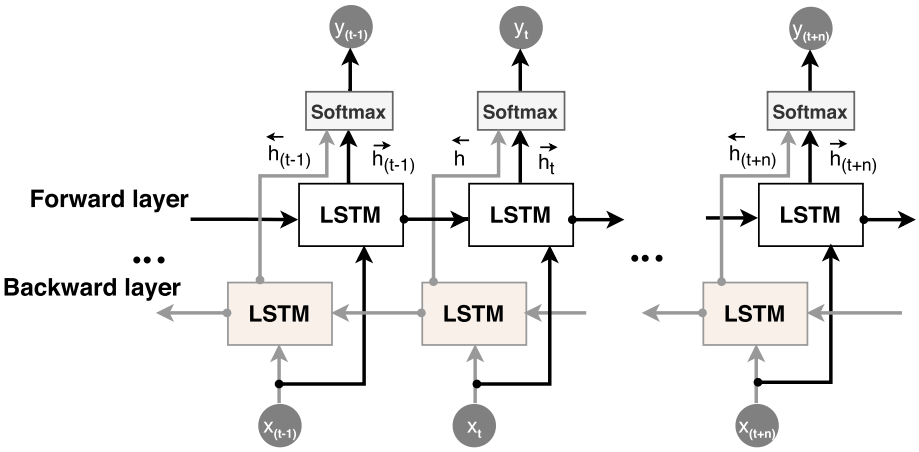

Fig. 6 An illustration of a BiLSTM neural network consisting of a forward layer and a backward layer

regulated by the gated layers underneath. A gate layer controls the flow of data and determines which information is relevant and kept and which information is irrelevant and removed from the state during the learning process. The gate structure consists of an element-wise multiplication operator and a Sigmoidal neural network layer. When the output of the Sigmoid layer is "0", the flow of information is prohibited since it is multiplied by 0 and, therefore, has been forgotten; conversely, when this layer is "1", the data flows.

LSTM has three types of gates: the forget gate, the input gate, and the output gate. The input of the forget gate is the data from the previous hidden state h_{t-1} and the current input x_t, which enter a "σ" neural layer, the output of which determines whether the previous information C_{t-1} is kept or forgotten. The output of the forget gate is given by Equation 2, in which σ is the Sigmoid function, W is the weight, and b is the bias.

$$f_t = \sigma\left(W_f.\left[h_{t-1}, x_t\right] + b_f\right) \tag{2}$$

Two neural layers (a Sigmoid layer (input gate) and a "tanh" layer (the hyperbolic tangent function)) are considered to decide which data should be written on the memory cell. Both layers take h_{t-1} and x_t as inputs. The input gate layer determines which information to update, while the "tanh" layer assists in regulating the learning process since its output lies within the range of -1 to 1. The output of the input gate i_t and the "tanh" ($\tilde{C}_t$) are defined in Eqs. (3, 4).

$$i_t = \sigma\left(W_i.\left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

$$\tilde{C}_t = tanh\left(W_C.\left[h_{t-1}, x_t\right] + b_C\right) \tag{4}$$

Until this point, the memory cell has updated its state depending on the previous operations of the forget and input gates. Therefore, the new state C_t is described by Eq. 5, where $\otimes$ is an element-wise multiplication.

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \tag{5}$$

The final stage of the LSTM network is computing the output o_t if it is relevant to the learning task, This step results in determining the next hidden state h_t (see Eqs. 6 & 7).

$$o_t = \sigma\left(W_o.\left[h_{t-1}, x_t\right] + b_o\right) \tag{6}$$

$$h_t = o_t \otimes tanh(C_t) \tag{7}$$

LSTM networks have different structuring variants, one of which is the bidirectional LSTM (BiLSTM) network. While the LSTM (the unidirectional) was proposed to capture the context from the previous hidden states, the BiLSTM is proposed to consider the contexts from previous (historical) and next (future) states. This is achieved by combining two different layers (i.e., the feedforward LSTM and the backward LSTM layers). Doing this gives access to forward and backward information at each time step and successfully promotes the performance of sequential data prediction (Liu and Guo 2019; Zhou et al. 2019). Figure 6 is a graphical representation of a BiLSTM network.

5.3 Softmax layer

The softmax layer is one of the output layers in deep neural networks; it is often the last layer in the structure. Usually, the neurons compute the weighted sum z of the input a (as in Eq. 8, where k is the number of neurons from the previous layer and j is the number of neurons in the current layer L). This weighted sum is then applied to a non-linear activation function (e.g., a Sigmoid function). However, the softmax layer does not use non-linear activation to generate the final output but instead employs a softmax function.

$$z_j^L = \sum_k w_{jk}^L \cdot a_k^{L-1} + b_j^L \tag{8}$$

The softmax activation function is represented by Eq. 9. From the equation, it can be concluded that increasing the value of z_j^L of an output neuron increases the output a_j^L, while the outputs of all other neurons are decreased so that the sum of all neurons is $1\left(\sum_j a_j^L = 1\right)$.

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} \tag{9}$$

The softmax layer predicts the probability distribution of the classes, as each output neuron produces an output $a_j^L \in [0, 1]$ that acts as the probability of the prediction of its respective class.

6 Evaluation measures

The evaluation of the multi-class classification model relies on four measures (namely, the accuracy, the macro-average of recall, precision, and f1-score). Accuracy is considered the correct speciality predictions divided by the total number of questions (m), given as a percentage. This is illustrated in Eq. 10, where y is the actual label of question (i), and $\hat{y}$ is the predicted label.

$$\texttt{Accuracy}(y, \hat{y}) = \frac{1}{m}\sum_{i=0}^{m-1} 1\ (\hat{y}_i = y_i) \tag{10}$$

The macro-recall ($Recall_m$) computes the mean of the recall of each class. The recall indicates how well the model can recognize examples of the class of interest. Macro-recall is calculated using Eq. 11, where (L) is the total number of classes, (y_l) represents the questions that were assigned to the predicted label l, and $\hat{y}_l$ denotes the number of samples that have true labels.

$$Recall_m = \frac{1}{|L|}\sum_{l\in L} R(y_l, \hat{y}_l),\quad R(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|\hat{y}_l|} \tag{11}$$

Macro-precision ($Precision_m$) is the average precision across all classes. In this case, the macro-precision is the number of correctly identified positive questions in proportion to the actual number of positive questions. It is calculated using Eq. 12, where "positive" refers to the class of interest.

$$Precision_m = \frac{1}{|L|}\sum_{l\in L} P(y_l, \hat{y}_l),\quad P(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|y_l|} \tag{12}$$

Macro f1-score ($F1 - score_m$) is the unweighted average of all f1-scores from all classes. The f1-score is the harmonic mean of precision and recall, as it expresses the level of balance between them. $F1 - score_m$ is given by Eqs. (13, 14).

Table 1 The settings used to build the Word2Vec model and the classification deep learning model

Word2Vec		LSTM/BiLSTM	
Parameter	Value	Parameter	Value
Structure	SG	Loss function	Categorical_crossentropy
Word vector dimension	300	Optimizer	Adam
Window size	3	Learning rate	0.001
Iterations	5	Epochs	100
Workers	4	Batch size	1024
Minimum count	3	Recurrent_dropout	0.2
Maximum input length	502	Dropout	0.2
		Return_sequences	False

$$F1 - score_m = \frac{1}{|L|}\sum_{l\in L} F_\beta\left(y_l, \hat{y}_l\right) \tag{13}$$

$$F_\beta(y_l, \hat{y}_l) = \left(1 + \beta^2\right)\frac{P(y_l, \hat{y}_l) \times R(y_l, \hat{y}_l)}{\beta^2 P(y_l, \hat{y}_l) + R(y_l, \hat{y}_l)} \tag{14}$$

7 Experiments and results

7.1 Experimental setup

The proposed approach has two main parts: the feature extraction model, which depends on the word embedding representation (Word2Vec), and the deep classification model. Table 1 shows the experimental configurations of these two parts. The Word2Vec model is based on the SG structure, which is utilized to construct word embeddings using the Gensim library (Řehůřek and Sojka 2010). In this case, the embedding dimension was set to 300, the window size was 3, the minimum count was 3, the workers were 4, and the number of iterations was 5. Furthermore, questions were expanded so that they all had the same length (502), with shorter questions filled by zeros to accomplish this. Meanwhile, the deep LSTM and BiLSTM models were developed using a sequential stack of layers. These models were trained using 60% of the data and were validated and tested using two different subsets, each of which represented 20% of the data. The training process was based on the "Categorical_crossentropy" function, which served as the loss function for multi-class classification. The weights of the network were learned using an adaptive learning optimizer, which represents the adaptive moment estimation ("Adam"). Its learning weight parameter was set to a fixed learning rate of (0.001). This learning rate was chosen based on multiple

Table 2 A summary of the parameters and output shapes of different layers of the best models

Word2Vec		LSTM/BiLSTM	
Parameter	Value	Parameter	Value
Output shape	(1024, 502, 300)	Output shape	BiLSTM: (1024, 60) LSTM: (1024, 40) Dense layer: (1024, 15)
No. of parameters	22,428,300	No. of parameters	BiLSTM: 79,440 LSTM: 54,560
		Parameters of the dense layer	BiLSTM: 915 LSTM: 615

experiments of a step-decaying learning rate starting at 0.01 and then decreasing (see Eq. 15). The best results were obtained when the learning rate was 0.001 at 100 epochs. Also, the batch training strategy was used with a size of 1024. In addition, the "dropout" and "recurrent_dropout" for LSTM/BiLSTM were set to 0.2, and "return_sequences" was not used. The total numbers of trainable and non-trainable parameters of LSTM were 55,175, and 22,428,300, respectively. For the BiLSTM, the values were 80,355, and 22,428,300, respectively. Table 2 summarizes the proposed and best-obtained models.

$$learning\ rate = learning\ rate \cdot 0.5^{(\frac{1+epoch}{2})} \tag{15}$$

Regarding system and hardware settings, the used system was Windows, the development platform was Google's Colaboratory, the processor of which was Intel(R) Xeon(R) CPU @ 2.00GHz, and the memory of which was 27 GB. The deep learning framework employed was Keras (Chollet et al. 2015) based on the Tensorflow (Abadi et al. 2015) backend.

7.2 LSTM and BiLSTM results

This subsection presents the performance results of different LSTM and BiLSTM models that were structured using various numbers of units.

Table 3 shows a comparison of the performances of the LSTM and BiLSTM frameworks based on their classification accuracies at training, validation, and testing. The comparison also considers the effect of different numbers of units (10-50). For the LSTM and BiLSTM, the accuracy increased when the number of LSTM units was increased throughout the training and validation processes. Conversely, it slightly decreased during the testing phase. During the training stage, the highest accuracy rates for LSTM (50) and BiLSTM (50) were 88.8% and 88.7%, respectively. Meanwhile, during the validation stage, LSTM (50) and BiLSTM (50) achieved accuracy rates of 87.3%, 87.4%, respectively. These results indicate the model's resistance to not overfitting and underfitting. During testing, both models gradually exhibited a slight increase in the accuracy. For instance, the accuracy of the LSTM model peaked at 87.1% at LSTM (40) but then decreased by 0.002 at LSTM (50). Meanwhile, the accuracy of the BiLSTM model reached its peak of 87.2% at BiLSTM (30) before decreasing by 0.003, and 0.004

Table 3 A comparison of the accuracy measure at training, validation, and testing of different LSTM and BiLSTM configurations

Model	Training accuracy	Validation accuracy	Testing accuracy
LSTM (10)	0.837	0.849	0.864
LSTM (20)	0.864	0.868	0.869
LSTM (30)	0.875	0.869	0.870
LSTM (40)	0.882	0.873	0.871
LSTM (50)	**0.888**	0.873	0.869
BiLSTM (10)	0.838	0.853	0.863
BiLSTM (20)	0.861	0.867	0.871
BiLSTM (30)	0.874	0.872	**0.872**
BiLSTM (40)	0.881	**0.874**	0.869
BiLSTM (50)	0.887	**0.874**	0.868

The best results have been highlighted in bold

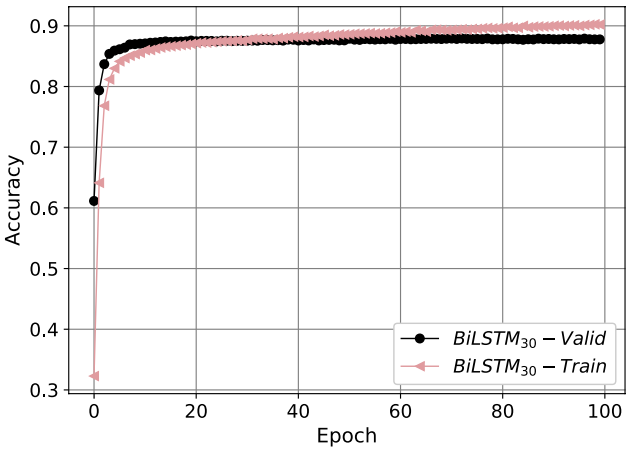

Fig. 7 The accuracy of the BiLSTM (30) during training and validation

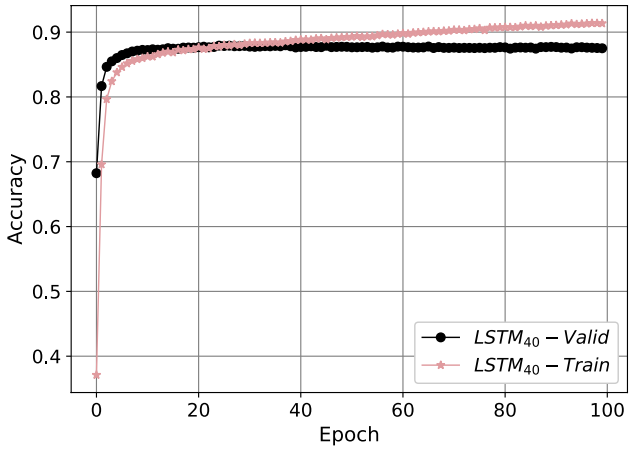

Fig. 8 The accuracy of the LSTM (40) during training and validation

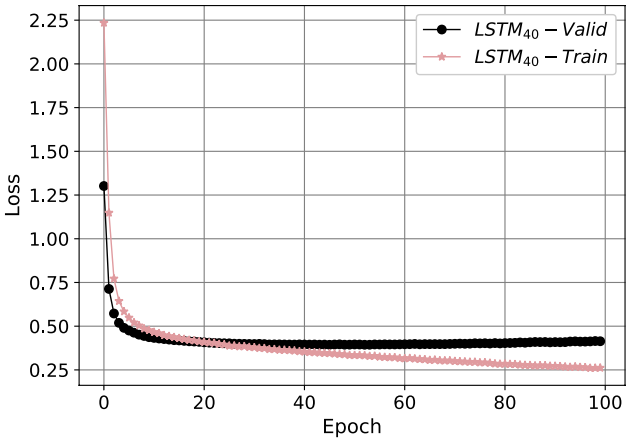

Fig. 10 The loss at training and validation for the LSTM (40)

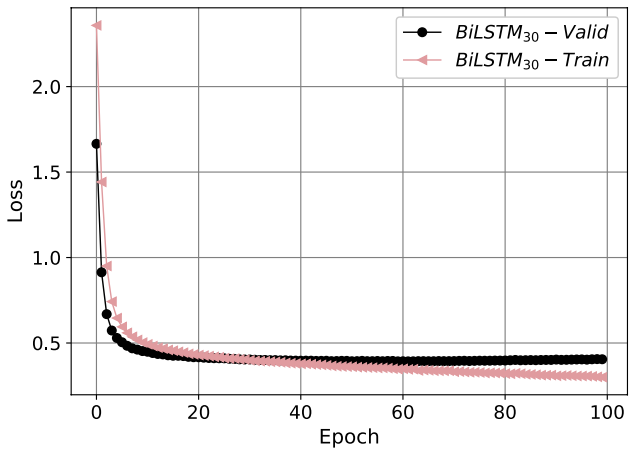

Fig. 9 The loss at training and validation for the BiLSTM (30)

at BiLSTM (40), and BiLSTM (50). Overall, LSTM (40) and BiLSTM (30) were associated with the best testing accuracy rates.

Figures 7 and 8 depict the convergence of the best models that were obtained (i.e., BiLSTM (30), and LSTM (40)) based on the classification accuracy during training and validation over 100 epochs of learning. It is clear that the curves show a smooth increase toward the optimal values of the accuracy after ten epochs of training, which means it took 10 epochs to converge from random initialization toward better near-optimal values. In addition, the converging trends experience a slight difference between the training and validation curves, which indicates that the model does not overfit.

Furthermore, Figs. 9 and 10 show the convergence of the best models based on the loss function (objective function) and across all epochs. During the training stage, the curves of BiLSTM (30) and LSTM (40) converge steadily where the loss values decrease as the number of epochs increases. However, during the validation process, the curves converge to their minimum values at epoch 30, after which they remain relatively constant. At 100 epochs, the LSTM (40) is more sensitive and susceptible to overfitting than BiLSTM (30) since there is a higher difference between the training and validation curves.

Table 4 provides a comparison between LSTM (40) and BiLSTM (30) based on their precision, recall, and f1-measure values for all classes, as well as the overall macro-average values of the metrics. The precision values of BiLSTM (30) ranged from 83% to 95%. Minimal precision was recorded for "Musculoskeletal Diseases" and "Gynecology & Women Diseases," perhaps because the percentage of women than men experience musculoskeletal diseases (Nakua et al. 2015; Worell 2001). Meanwhile, for "Child Health," "Dental Medicine," and "Urology & Venereology" the model achieved precision rates of 94.9%, 90.2%, and 91.7%, respectively. Meanwhile, for LSTM (30), the precision values ranged from 81.9 to 94.1%. The least precision was associated with "Ear, Nose & Throat" (81.9%) and "Gynecology & Women Disease" (82.7%). The highest precision rates were recorded for "Child Health" (94.1%), "Dental Medicine" (90.9%), "Urology & Venereology" (92.2%), and "Digestive System Diseases" (90.2%).

In terms of recall, values for BiLSTM (30) ranged from 82 to 95%, and values for LSTM (40) ranged from 80 to 96%. The highest recall for BiLSTM (30) (94.5%) was obtained for the "Child Health" class and the lowest (82.4%) was obtained for "Ear, Nose & Throat" and "Gynecology & Women Diseases." The highest recall for LSTM (40) (96.5%) was recorded for "Child Health", the lowest (80.3%) was recorded for "Gynecology & Women Diseases." The range of f1-scores was similar to those of precision and recall values. Furthermore, for BiLSTM (30), the best f1-score (94.7%) was obtained for "Child Health," while the lowest f1-scores (approximately 83%) were found for "Musculoskeletal Diseases," "Gynecology & Women

Table 4 Comparison between BiLSTM (30) and LSTM (40) based on precision, recall, and f1-scores for all classes

Class	BiLSTM (30)			LSTM (40)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Diabetes	**0.859**	0.854	**0.857**	0.850	**0.861**	0.855
Child Health	**0.949**	0.945	0.947	0.941	**0.965**	**0.953**
Ear, Nose & Throat	**0.844**	0.824	**0.834**	0.819	0.824	0.822
Dental Medicine	0.902	**0.922**	**0.912**	**0.909**	0.912	0.911
Nutrition	0.859	**0.847**	**0.853**	**0.860**	0.842	0.851
Ophthalmology & Eye Diseases	0.841	**0.869**	0.855	**0.868**	0.861	**0.864**
Dermatology	0.853	**0.888**	**0.870**	**0.859**	0.873	0.866
Heart Disease	**0.860**	0.873	0.867	0.855	**0.879**	0.867
Tumors	**0.873**	0.855	0.864	0.872	**0.862**	**0.867**
Psychiatric Diseases	**0.888**	0.860	**0.874**	0.865	**0.875**	0.870
Urology & Venereology	0.917	**0.953**	0.935	**0.922**	0.949	0.935
Digestive System Diseases	0.890	0.875	0.883	**0.902**	**0.880**	**0.891**
Musculoskeletal Diseases	0.833	**0.824**	**0.829**	**0.837**	0.813	0.825
Sexual Health	0.878	0.868	0.873	**0.880**	**0.870**	**0.875**
Gynecology & Women Diseases	**0.832**	**0.824**	**0.828**	0.827	0.803	0.815
Macro-average	**0.872**	**0.872**	**0.872**	0.871	0.871	0.871

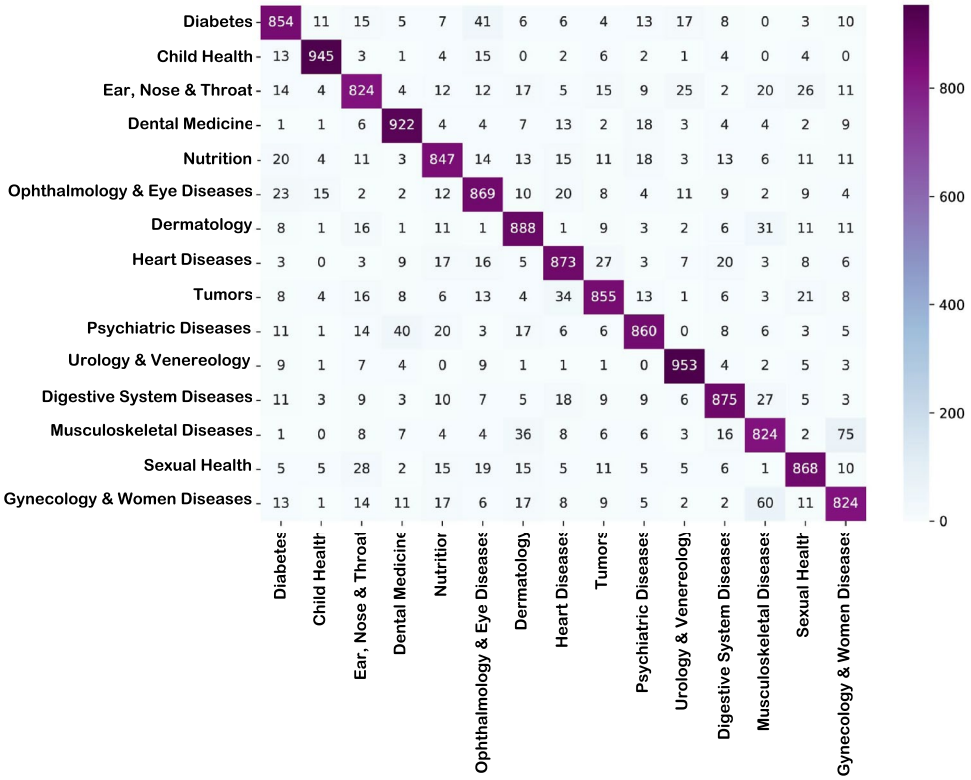

Fig. 11 A heatmap of the confusion matrix for the BiLSTM with 30 units. The vertical axis presents the true labels, and the horizontal axis represents the predicted labels

Diseases," and "Ear, Nose & Throat." For the LSTM (40), the best f1-score was also obtained for "Child Health," and the lowest was obtained for "Gynecology & Women Diseases". Moreover, referring to Fig. 1, it can be noticed that both models achieved very good recall rates (80–96%) across the classes although the classes have different distributions and some of them have outliers

Overall, neither the BiLSTM (30) nor the LSTM (40) performs better than the other, as each performs better than the other for certain metrics and classes.

In the confusion matrix of the best deep learning model (BiLSTM (30)) (Fig. 11), the horizontal axis shows the predicted labels while the vertical represents the true labels. The "Diabetes" class contains 41 instances that were

Table 5 A comparison of the performance of the best model (BiLSTM (30)) using different word embedding methods

Model		Accuracy			Precision	Recall	F1-score
		Training	Validation	Testing			
Altibbi-Embedding	300	0.8738	**0.8721**	**0.8721**	**0.8720**	**0.8721**	**0.8719**
AraVec (witter)	300	0.7926	0.8031	0.8125	0.8124	0.8125	0.8122
AraVec (wiki)	300	0.7791	0.7893	0.8071	0.8072	0.8071	0.8069
Keras-Embedding	300	**0.9867**	0.8059	0.7872	0.7878	0.7872	0.7872

misclassified as "Ophthalmology & Eye Diseases," This is logical, as some of the symptoms of diabetes are related to vision (Gong and Cormack 2020). There were also many misclassifications between "Musculoskeletal Diseases" and "Gynecology & Women Diseases." Specifically, 75 "Musculoskeletal Diseases" were classified as "Gynecology & Women Diseases" and 60 were misclassified in the opposite way. Such a high number of misclassified examples between medically related classes reveals the challenging nature of the problem.

7.3 Comparison with other neural networks architectures

This subsection provides a comparison of the created embedding layer based on Altibbi's dataset while using different embedding layer structures, for the best model (i.e., BiLSTM (30)).

Table 5 presents a comparison of the embedding model used by Altibbi with different models proposed by AraVec (Soliman et al. 2017), which are based on Twitter or Wikipedia at 300 dimensions. Additional comparisons are made by a randomly initialized and trainable embedding matrix with a dimension of 300. The table shows the performance based on the accuracy at training, validation, and testing, as well as the precision, recall, and f1-scores. BiLSTM (30) with Altibbi's embedding model yielded the best results when all performance measures were considered. Of all the models, randomly initialized Keras embedding performed the worst with accuracy rates of 98.7%, 80.6%, and 78.7% for training, validation, and testing, respectively. Even though it achieved high training accuracy, it experienced severe overfitting. "AraVec_twitter" (300) produced the second-best results after Altibbi embeddings, as it was 81.3% accurate for precision, recall, and f1-score.

It is worth noting that even the proposed approach is built while considering the context of the Arabic language, but this does not mean it is just customized for the Arabic language. The Arabic language has its special preprocessing steps to prepare the data in an appropriate and acceptable format for any machine and deep learning algorithm. Evaluating the classification model in the context of the English language is expected to perform better since the structure of the Arabic language is more complex and challenging for the learning algorithm to understand the meanings and the hidden relationships between words.

8 Assumptions and limitations

This section provides a description of the limitations and delimitations of the proposed approach. Developing a question answering or classification model using electronic medical or health records is of significant importance. However, creating such models is challenging for some languages that are weakly studied in the literature and have a lack of resources in NLP. This paper is kind of a preliminary study to create a question classification framework in the Arabic language in healthcare. Hence, this section pinpoints the challenges and limitations faced during the development of the model.

In the real-world, data is much complex than expected in theories, since it is rife with noisy, missing, and exceptional scenarios. Yet the amount of available correctly labeled training data in the health and medical fields is relatively small to build a state of the art models that need much larger amounts of data. Therefore, building the bidirectional encoder representation from transformers (BERT) is a further research direction to explore and implement as the size of the data increases. Furthermore, the proposed model ingests consultations and classifies them into appropriate specialities out of 15 predefined classes, but the actual number of specialities can be much more than this number, where many of them are overlapping in their characteristics causing the identification process to be more challenging. Moreover, the distribution of the data per class varies highly. Certainly, the presence of imbalanced data degrades the performance of the learning algorithm and urges handling such cases by special techniques like over-sampling or under-sampling. This is critical since it is related to sensitive real-case medical scenarios where the precision and the accuracy of made decisions are important.

9 Conclusions and future works

Automatic question answering is essential to advancing medical informatics, including cyber telemedicine and telehealth services. Question classification has a considerable influence

on the accuracy of the question-answering process. Therefore, this work assessed various elements of medical question classification. Deep neural network is useful in extracting deep representations of features. Thus, they improve the question answering process when a large amount of data is involved. This was demonstrated using Altibbi Company for telemedicine services as a case study. Altibbi proposed a medical speciality classification approach based on deep neural networks. As part of this, Altibbi curated and collected approximately 1.5 million health-related questions, which were subsequently used to construct word embeddings. Two types of recurrent deep neural models (namely, LSTM and BiLSTM deep networks) were used with their behaviors interpreted at different numbers of units. The LSTM (40 units) and BiLSTM (30 units) performed very well on various evaluation measures and showed smooth convergence over all epochs. Remarkably, the BiLSTM (30) achieved a maximum classification accuracy of 87.2%. Further, a comparison between the embeddings proposed by Altibbi and randomly created embeddings (as well as different AraVec embedding models) revealed that the best results were obtained when using LSTM and BiLSTM models.

However, additional research is required to further investigate the relationships between different hyperparameters settings that are related to either the construction of the embedding layer or the deep learning model. Further, the lack of sufficient data is a major weakness when training deep neural networks. Therefore, implementing transfer learning techniques is of huge advantage in decreasing the training time and boosting the performance without the need for big training data. Moreover, the Word2Vec model faces some limitations, such as multi-sense disambiguation, where a word might have different meanings depending on its context. However, the Word2Vec model creates one representation for such a word in all contexts which degrades the algorithm's performance. Hence, the application of recently developed language models that take the context of the words into consideration like the BERT model needs to be explored based on much larger domain-specific datasets.

Acknowledgements This work has been supported in part by: Ministerio español de Economía y Competitividad under project TIN2017-85727-C4-2-P (UGR-DeepBio).

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M et al (2015) Tensorflow: large-scale machine learning on heterogeneous systems

Abdallah A, Kasem M, Hamada M, Sdeek S (2020) Automated question answer medical model based on deep learning technology. arXiv:200510416

Agrawal S, Mishra N (2019) Question classification system for health care: a review. In: Proceedings of the Third International Conference on Advanced Informatics for Computing Research, Association for Computing Machinery, New York, NY, USA, ICAICR '19, 10.1145/3339311.3339341

Ahmed W, Ahmed A, Babu AP (2017) Web-based arabic question answering system using machine learning approach. Int J Adv Res Comput Sci 8:1

Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, Barkan E, Herzel E, Naor S, Karavani E et al (2019) Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292(2):331–342

Aydoğan M, Karci A (2020) Improving the accuracy using pre-trained word embeddings on deep neural networks for turkish text classification. Phys A Stat Mech Appl 541(123):288

Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, Chapman B, Amrhein T, Mong D, Rubin DL et al (2019) Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. Artif Intell Med 97:79–88

Chollet F et al (2015) Keras. https://keras.io

Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A (2020) Deep learning techniques for biomedical and health informatics. Springer, Berlin

Edara DC, Vanukuri LP, Sistla V, Kolli VKK (2019) Sentiment analysis and text categorization of cancer medical records with lstm. J Ambient Intell Hum Comput 2019:1–17

Estrada S, Lu R, Conjeti S, Orozco-Ruiz X, Panos-Willuhn J, Breteler MM, Reuter M (2020) Fatsegnet: a fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. Magn Reson Med 83(4):1471–1483

Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, Back T, Chopra R, Pontikos N, Kern C et al (2019) Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Dig Health 1(5):e232–e242

Faris H, Habib M, Faris M, Alomari M, Alomari A (2020) Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines. J Biomed Informatics 2020:103525

Florentia (2020) Florentia clinic

Gong JW, Cormack TG (2020) Re: vision loss as a presenting symptom of type ii diabetes mellitus. Br J Gener Pract 2020:5

Hasan AM, Rassem TH, Noorhuzaimi M et al (2018) Combined support vector machine and pattern matching for arabic islamic hadith question classification system. In: International conference of reliable information and communication technology, Springer, pp 278–290

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Jain DK, Jain R, Upadhyay Y, Kathuria A, Lan X (2019) Deep refinement: capsule network with attention mechanism-based system for text classification. Neural Comput Appl 2019:1–18

Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. J Document 1972:5

Kim J, Jang S, Park E, Choi S (2020) Text classification using capsules. Neurocomputing 376:214–221

Kumar A, Sarkar S, Pradhan C (2020) Malaria disease detection using cnn technique with sgd, rmsprop and adam optimizers. In: Deep learning techniques for biomedical and health informatics. Springer, pp 211–230

Kwak GHJ, Hui P (2019) Deephealth: Deep learning for health informatics. arXiv:190900384

Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B (2020) Early detection of

sepsis utilizing deep learning on electronic health record event sequences. Artif Intell Med 2020:101820
Li Y, Yang T (2018) Word embedding for understanding natural language: a survey. In: Guide to big data applications. Springer, pp 83–104
Liu F, Weng C, Yu H (2019a) Advancing clinical research through natural language processing on electronic health records: traditional machine learning meets deep learning. In: Clinical Research Informatics. Springer, pp 357–378
Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. Neurocomputing 337:325–338
Liu HI, Ni CC, Hsu CH, Chen WL, Chen WM, Liu YT (2020) Attention based r&cnn medical question answering system in chinese. In: 2020 International conference on artificial intelligence in information and communication (ICAIIC), IEEE, pp 341–345
Liu J, Shang W, Lin W (2018) Improved stacking model fusion based on weak classifier and word2vec. In: 2018 IEEE/ACIS 17th international conference on computer and information science (ICIS), IEEE, pp 820–824
Liu J, Yang Y, Lv S, Wang J, Chen H (2019b) Attention-based bigru-cnn for chinese question classification. J Ambient Intell Hum Comput 2019:1–12
Longuenesse E, Chiffoleau S, Kronfol N, Dewachi O (2012) Book: Public health in the arab world section: the context of public health chapter: Public health, the medical profession and state building–a historical perspective. HAL multidisciplinary open archive
Luhn HP (1957) A statistical approach to mechanized encoding and searching of literary information. IBM J Res Dev 1(4):309–317
Mairittha T, Mairittha N, Inoue S (2020) Improving fine-tuned question answering models for electronic health records. In: Adjunct Proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM International Symposium on Wearable Computers, pp 688–691
Mayo (2020) mayo clinic
Mikolov T, Chen K, Corrado G, Dean J, Sutskever L, Zweig G (2013) word2vec. https://www.codegooglecom/p/word2vec22
Mulani J, Heda S, Tumdi K, Patel J, Chhinkaniwala H, Patel J (2020) Deep reinforcement learning based personalized health recommendations. In: deep learning techniques for biomedical and health informatics. Springer, pp 231–255
Naili M, Chaibi AH, Ghezala HHB (2017) Comparative study of word embedding methods in topic segmentation. Procedia Comput Sci 112:340–349
Nakua EK, Otupiri E, Dzomeku VM, Owusu-Dabo E, Agyei-Baffour P, Yawson AE, Folson G, Hewlett S (2015) Gender disparities of chronic musculoskeletal disorder burden in the elderly ghanaian population: study on global ageing and adult health (sage wave 1). BMC Musculoskel Disord 16(1):204
Novomed (2020) Novomed centers
Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
Rawat BPS, Weng WH, Raghavan P, Szolovits P (2020) Entity-enriched neural models for clinical question answering. arXiv :200506587
Řehůřek R, Sojka P (2010) Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, pp 45–50
Ren J, Liu N, Wu X (2020) Clinical questionnaire filling based on question answering framework. Int J Med Informatics 141(104):225
Romeo S, Da San MG, Belinkov Y, Barrón-Cedeño A, Eldesouki M, Darwish K, Mubarak H, Glass J, Moschitti A (2019) Language processing and learning models for community question answering in arabic. Inf Process Manag 56(2):274–290
Ryu JY, Kim HU, Lee SY (2018) Deep learning improves prediction of drug-drug and drug-food interactions. Proc Nat Acad Sci 115(18):E4304–E4311
Sammut C, Webb GI (eds) (2010) TF–IDF, Springer US, Boston, MA, pp 986–987. https://doi.org/10.1007/978-0-387-30164-8_832
Schmidt L, Weeds J, Higgins J (2020) Data mining in clinical trial text: transformers for classification and question answering tasks. arXiv:200111268
Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
Shah AM, Yan X, Shah SAA, Mamirkulova G (2019) Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. J Ambient Intell Hum Comput 2019:1–18
Soliman AB, Eissa K, El-Beltagy SR (2017) Aravec: a set of arabic word embedding models for use in arabic nlp. Procedia Comput Sci 117:256–265
Soltanolkotabi M, Javanmard A, Lee JD (2019) Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. IEEE Trans Inf Theory 65(2):742–769
Statista (2020) The world's most spoken languages
Vidhya K, Shanmugalakshmi R (2020) Deep learning based big medical data analytic model for diabetes complication prediction. J Ambient Intell Hum Comput 2020:1–12
Vu MH, Löfstedt T, Nyholm T, Sznitman R (2020) A question-centric model for visual question answering in medical imaging. IEEE Trans Med imaging 2020:8
Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H (2019) A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 19(1):1
Worell J (2001) Encyclopedia of women and gender, two-volume set: sex similarities and differences and the impact of society on gender, vol 1. Academic Press, Cambridge
Yegnanarayana B (2009) Artificial neural networks. PHI Learning Pvt, New York
Yilmaz S, Toklu S (2020) A deep learning analysis on question classification task using word2vec representations. Neural Comput Appl 57:1–20
Zhang L, Lin J, Liu B, Zhang Z, Yan X, Wei M (2019) A review on deep learning applications in prognostics and health management. IEEE Access 7:162,415–162,438
Zhang Q, Mu L, Zhang K, Zan H, Li Y (2018) Research on question classification based on bi-lstm. In: Workshop on Chinese Lexical Semantics, Springer, pp 519–531
Zhou J, Lu Y, Dai HN, Wang H, Xiao H (2019) Sentiment analysis of chinese microblog based on stacked bidirectional lstm. IEEE Access 7:38,856–38,866
Zhu Y, Li L, Lu H, Zhou A, Qin X (2020) Extracting drug-drug interactions from texts with biobert and multiple entity-aware attentions. J Biomed Informatics 2020:103451