ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 5, No. 4, December, 2024

Contextual Semantic Embeddings Based on Transformer Models for Arabic Biomedical Questions Classification

Ismail Ait Talghalit [1*], Hamza Alami [2], Said Ouatik El Alaoui [1]

[1] *Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, 14000, Morocco.*

[2] *LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, 30003, Morocco.*

Received 19 July 2024; Revised 23 November 2024; Accepted 26 November 2024; Published 01 December 2024

Abstract

Arabic biomedical question classification (ABQC) is a challenging task due to various reasons including, the specialized jargon expressed in Arabic language, complex semantics of Arabic vocabulary and the lack of specific datasets and corpora. When representing questions, only a few studies deal with ABQC by taking into account the word context. In this work, we propose a classification model designed for Arabic biomedical questions. We build vector representations capturing the contextual and semantic information of Arabic biomedical text, which presents numerous challenges, such as the derivational morphology of Arabic language, the specialized terminology of biomedical terms and the lack of capitalization in text. Our representation adapts the extensive knowledge encoded in BERT (Bidirectional Encoder Representations from Transformers) and other transformer models, to address the aforementioned challenges. Several experiments have been conducted on a dedicated Arabic biomedical dataset namely: MAQA, with well-known transformer models including BERT, AraBERT, BioBERT, RoBERTa, and DistilBERT fine-tuned for the classification task. Obtained results show that our method achieves remarkable performance with an accuracy of 93.31% and an F1-score of 93.35%.

Keywords: Arabic Question Classification; Biomedical Domain; Natural Language Processing; Transformers; BERT; Fine-Tuning; Question Answering Systems; Sentence Embedding.

1. Introduction

Arabic biomedical questions classification is the process of categorizing Arabic biomedical text, particularly biomedical questions, into predefined categories, such as diseases, medical procedure, symptoms, and specialties. This is particularly useful for healthcare applications [1], medical information retrieval systems, and decision support tools, since it allows these systems to better understand the type of information that user is seeking. As shown in Figure 1, our classification system takes an Arabic biomedical question as input, processes it, and assigns it to the appropriate category. By classifying questions into predefined categories, the system can refine the search process and retrieve more relevant information [2]. For instance, if a question is classified as a symptom, the system will focus resources that specifically related to symptoms, thus enhancing the efficiency of the retrieval process.

Biomedical question classification is essential in various biomedical applications, such as medical chatbots, decision support systems, and symptom checker applications. It enhances communication between patients and healthcare providers by allowing chatbots to understand and respond accurately to medical inquiries. In their work, Babu & Boddu

* Corresponding author: ismail.aittalghalit@uit.ac.ma

http://dx.doi.org/10.28991/HIJ-2024-05-04-011

[3] focused on developing a medical chatbot that uses BERT [4], to understand and classify biomedical natural language input. Tama & Lim [5] discussed the impact of various classifiers in clinical decision support systems and their importance in biomedical classification tasks. The article evaluates classifiers such as Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbours, and Neural Networks, highlighting their effectiveness in processing complex biomedical data. Hassan et al. [6] focused on using advanced language models to improve the classification of diseases from symptoms, exploring many pre-trained language models to process symptom description. These models are fine-tuned on medical datasets to improve their understanding of medical terminology and disease symptoms. The paper presents a set of experiments comparing traditional machine learning classifiers (SVM, and Decision Trees) with language models on disease classification tasks. The results show that language models outperform traditional methods, especially when dealing with complex or ambiguous symptom descriptions.

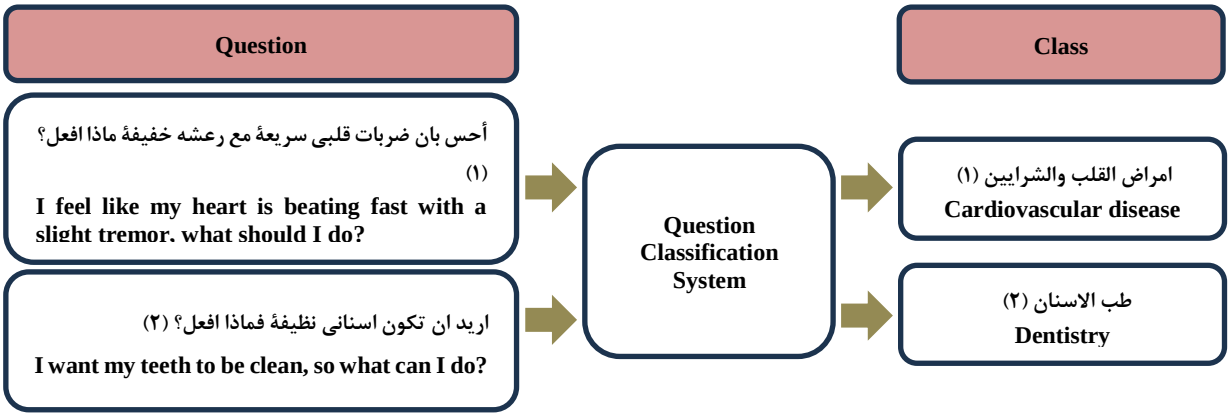

Figure 1. Arabic biomedical question classification system workflow

Question classification can be used in building open domain question answering systems. To provide the correct answer, this process is required to understand what type of information the question is seeking. As a result, a good classification narrows down the target data and makes passage retrieval component more efficient [7]. Alami et al. [8] introduced a new taxonomy for open-domain Arabic questions alongside a classification approach that leverages distributed word representations. Initially, the method involves creating word embeddings that encapsulate semantic relations between words, which is then followed by the application of machine learning methods to categorize questions into distinct classes. This innovative method has shown significant enhancements, achieving an accuracy of 90% in the classification of Arabic questions.

Building an Arabic biomedical question classification system is a challenging Natural Language Processing (NLP) task due to various factors including the complex morphology of Arabic language, the specialized terminology of biomedical domain, and the limited availability of biomedical Arabic corpora. The absence of diacritics in most Arabic texts creates ambiguity, as the same word can have multiple meanings depending on its context. Moreover, the lack of capitalization makes named entities recognition (Institutions, Renowned Researchers, Diseases, Health Organizations…) more difficult compared to other languages. Another obstacle is that Arabic's orthographic variations, where certain letters have different writing forms (e.g., "ة" vs. "ه" or "ي" vs. "ى"). These variations can complicate tasks like tokenization, stemming, and lemmatization in NLP. In addition to these challenges, the biomedical field is continuously evolving, and there is no original Arabic translation available for most diseases and biomedical terms because they are primarily named in English. The interdisciplinary nature of the biomedical domain, encompassing fields like medicine, biology, biochemistry, bioinformatics, and biotechnology adds complexity by leading to overlaps between classes. Unlike English language, which has sufficient datasets for medical texts, Arabic lacks datasets in this domain, which presents a significant challenge for NLP tasks.

Recently, text representation becomes a crucial process for any NLP application such as information retrieval [9], text summarization [10], text clustering [11], etc. Unlike classical bag-of words representation, Word Embedding including Word2vec [12], Glove [13], and FastText [14] have shown high performances in Text Mining [15]. However, most of these methods don't take into consideration both relationships between words and the context in which the word is used. This aspect should not be overlooked in Biomedical questions particularly because the Arabic language contains a lot of successive and composed words, where the meaning can change when the overall context of the sentence is not considered. For instance, the following two sentences have different meanings: "يعاني من عمى " ("He suffers from blindness") and "يعاني من عمى الألوان" ("He suffers from color blindness"). The first sentence refers to "blindness" or the state of being visually impaired, where a person is unable to see or has significant visual impairment, while the second one refers to "color blindness" or "color vision deficiency". A person under this condition has difficulty distinguishing between certain colors or fails to recognize colors correctly.

In some cases, static Word Embedding can be ineffective for text representation and questions will be misclassified. A minor adjustment to the sentence can completely change the class. Despite having almost, the same content, and differing by only one word, sentences can belong to different classes. For instance, the sentences presented in Table 1 stand as examples of this misclassification.

Table 1. Sentence variations leading to misclassification in biomedical questions

Classes translation	Classes	Sentences translation	Sentences
Gastrointestinal diseases	امراض الجهاز الهضمى	My father has a pain in the neck of his stomach	ابى يعانى من الم فى عنق المعده
Ear nose and throat	انف اذن وحنجرة	My father has neck pain	ابى يعانى من الم فى عنق
Ophthalmology	امراض العيون	My son suffers from photophobia	ابنى يشكو من رهاب الضوء
Psychiatric and neurological illnesses	الأمراض النفسية والعصبية	My son has a phobia	ابنى يشكو من رهاب

Contextual embedding models have demonstrated efficiency in terms of learning sentence representations for many NLP tasks. Lahbari & El-Alaoui. [16] introduced an Arabic question classification system integrated within a question answering system. The system operates in three steps: it begins by classifiying questions and forming queries, followed by retrieving relevant documents and passages, and finally, extracting answers. The proposed approach combines AraBERT [17] with a passage retrieval and query expansion to find and rank relevant passages. Tested on CLEF and TREC datasets, the proposed method achieves good results, achieving a 92% in terms of F1-score.

In this paper, we propose an Arabic biomedical question classification system. We build representations that capture contextual and semantic information of Arabic biomedical text, enhancing classification task performance. Our representations adapt the rich encoded information in language models such as BERT [4] and AraBERT [17] to address challenges specific to the Arabic biomedical domain. Trained on the large dataset encoded in BERT and other transformer models makes our representations robust. Our models use the encoder part of transformers architecture, along with the attention mechanism to capture both close and distant word relationships and extracting important information within the Arabic biomedical text. This enables the model to catch contextual information, making our representations powerful. For classification, we use cross-entropy loss as our objective function, to ensure efficient and accurate categorization of questions. To validate our approach, we use the MAQA dataset [18], a large biomedical dataset containing diverse biomedical questions and classes in Arabic language. Using our representation, we have built several models for question classification tasks including BERT [4], AraBERT [17], RoBERTa [19], BioBERT [20] taking into account the word context and the overall meaning of the sentence for effective question classification. Additionally, we include DistilBERT [21] model, a light BERT version, to prove its speed efficiency, which is crucial in biomedical NLP tasks. The main contributions of this work are as follows:

- We propose an Arabic biomedical question classification system.
- We build specific vector representations for Arabic biomedical text that catch semantic and contextual information. These representations adapt the extensive knowledge encoded in pretrained transformer models, namely BERT [4], AraBERT [17], RoBERTa [19], BioBERT [20], and DistilBERT [21].
- We carried out several experiments on MAQA [18] corpus using accuracy, F1 score, precision and recall metrics for training and testing.
- We compare computing time of the best models to prove the performance of the light version of multilingual DistilBERT [21].

It is noteworthy that finding Arabic biomedical classification systems using contextual semantic embeddings are rare. Our work addresses this gap by experimenting with contextual semantic embeddings based on transformer models.

The remainder of the paper is organized as follows. Section 2 presents related work. In Section 3 we describe our devised approach. The experimental results and discussion are detailed in Section 4. Section 5 concludes and outlines future work.

2. Related Works

Deep learning techniques have been widely explored in many NLP applications and have shown big performances due to recent developments in Transfer learning approach. The later has become increasingly popular with the introduction of Transformers models [22]. Fine tuning a pretrained transformer model on a specific task has been demonstrated to be performant; it achieves state-of-the-art in NLP tasks, including text classification, question answering, and text translation.

Mutabazi et al. [23] proposed a novel deep learning model to enhance the classification of medical questions on forums. This model utilized on Word2Vec for word embedding, CNN for feature extraction, and BiLSTM for

classification. After training and testing on two benchmark datasets, the CNN-BiLSTM model demonstrated superior performance in capturing semantic and syntactic features. Results show that the combined model surpasses baseline methods (CNN and BiLSTM) on both datasets, achieving an accuracy of 57.73% on the ICHI dataset and 100% on the MedQuAD dataset.

Vihikan & Trisna [24] explored many deep learning techniques and baseline methods for the classification of health questions in Indonesian language. Authors developed a multi-classification model, capable of categorizing questions into different health related categories. The deep learning methods evaluated in this study include GRU, LSTM, Bidirectional LSTM and GRU (BiLSTM and BiGRU), CNN, BiGRU-CNN, BiLSTM-CNN, and Transformer-based models. The majority of the deep learning models tested in the study outperformed SVM, which is used as a baseline method.

Recently, various Transformers models have been proposed to enhance NLP tasks, such as BERT [4], which is a language model based on Transformer architecture that was launched by google in 2018. The initial BERT model available for Arabic was the multilingual BERT (mBERT). Devlin et al. [4] developed this model, which was pre-trained on a large corpus supporting 104 languages including Arabic language. The multilingual BERT uses two unsupervised learning tasks which are masked language modeling (MLM) and next sentence prediction (NSP). To optimize BERT approach, Facebook researchers created XLM-RoBERTa [19] which removes Next Sentence Prediction (NSP) and uses instead a different approach named Byte Pair Encoding (BPE). Mansour et al. [25] applied multilingual BERT fine-tuned to identify Arabic dialect from Arabic tweets. The authors compared BERT model with machine learning models including Support Vectors Machine, Naive Bayes, and Voting Classifier. Experiments have shown that BERT model is more efficient in language understanding, and has achieved the best F1 score compared with other machine learning models.

More recently, several pretrained models have been researched in the context of Arabic NLP tasks. ABioNER [26] is an extended BERT model that have been developed to identify Named Entity Recognition in Arabic biomedical text. The authors demonstrated that their model outperformed both AraBERT and multilingual BERT with 85% in terms of F1 score. However, authors have tested the model only on two types of entities ("Disease or Syndrome" and "Therapeutic or Preventive Procedure").

The KIMedQA system [27] is an enhanced medical question-answering system that employs knowledge graphs with pre-trained language models. The enhanced system combines query and context representations with the pruned knowledge network to produce precise answers. The results of the conducted research on the datasets of both MASH-QA and COVID-QA datasets reveal that KIMedQA outperforms ChatGPT in terms of F1 Score and adequacy.

To tackle the lack problem of Arabic Biomedical data sets, Hammoud et al. [28] produced a novel medical dataset for diseases classification by collecting multiple Arabic medical websites, as well as the Arab Medical Encyclopedia. After the fine-tuning of pretrained models BERT, AraBERT, and ABioNER on this Arabic medical corpus, authors obtained good results for a classification task of two thousand medical documents, containing 10 classes.

The fusion of transformer models with other deep learning methods in biomedical question classification has proven highly effective. Al-Smadi [29] has introduced a new performed model designed for classifying Arabic medical questions into multiple classes. The model integrates DeBERTa for extracting contextual embeddings and BiLSTM for comprehensive feature extraction and representation. Evaluated on a dataset of COVID-19-related questions from the Altibbi platform, the study demonstrates the potential of combining pre-trained language models with BiLSTM for effective multi-label classification. The model's performance surpasses that of other baseline methods, achieving a hamming loss of 0.042 and a micro-F1 score of 0.84.

Yu et al. [30] proposed TinyBERT-CNN model for intent classification in the in the Chinese text "Treatise on Febrile Diseases" using a fusion of TinyBERT and CNN. It is based on TinyBERT for embedding and encoding global text information, followed by CNN for extracting local features, achieving high accuracy of 96.4%.

In pursuit of speed and efficiency, researchers have been motivated to develop lighter and faster versions of models. For instance, DistilBERT [21] is a light and a fast BERT version, these two models share the general architecture, except that DistilBERT has 40% less trainable parameters and is suited for situations with restricted computer resources. Similar to other transformers architectures this model can be used for text classification. Akpatsa et al. [31] evaluated the performance of Distil-BERT model and other text classifiers on a Covid-19 online news binary classification dataset. Despite having fewer trainable parameters than the BERT-based model, the DistilBERT model achieved an accuracy of 0.94 on the validation set after only two training epochs.

In contrast to Multilingual BERT, AraBERT [17] was developed especially for Arabic language and its dialects, news in Modern Standard Arabic, taken from various Arabic media outlets constitute pretraining dataset. The final model is trained using over 3 billion Arabic tokens and 70 million phrases. Aftan & Shah [32] used AraBERT for

customer satisfaction classification based on tweets in Saudi Arabia Telecom Companies. They compared AraBERT to two Deep Learning algorithms including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Using Mobily and STC datasets, AraBERT achieved the best prediction accuracy.

El-Alami et al. [33] explored the AraBERT [17] model for contextual text representation in two ways: as a transfer learning model and as a feature extractor. They fine-tuned AraBERT parameters on the OSAC datasets to improve its efficacy in Arabic text classification. The effectiveness of AraBERT is further evaluated as a feature extractor by integrating it with various classifiers, including LSTM, Bi-LSTM, MLP, SVM, and CNN. Comparative experiments are conducted between two BERT models, AraBERT and multilingual BERT. The results have shown that the fine-tuned AraBERT model achieves a high performance, reaching an F1-score and accuracy of up to 99%.

BioBERT [20] is a biomedical language representation model created for biomedical applications. It was trained on a range of biomedical datasets, such as PMC full-text articles and PubMed abstracts. Houssein et al. [34] explored fine-tuned transformer models, such as BioBERT, RoBERTa, BERT, XLNet, and BioClinicalBERT, for heart disease detection and extraction of related risk factors from clinical notes using the i2b2 dataset. These fine-tuned transformer models have demonstrated high performance in extracting semantic information and identifying disease risk factors.

Based on the aforementioned works, despite the proven effectiveness of fine-tuned transformer models in many NLP applications, few efforts have been devoted to the Arabic biomedical domain due to many challenges, such as the specific jargon of the biomedical domain and the ambiguity present in the Arabic language. In this paper, we exploit the powerful capabilities of contextual semantic embedding with the fine-tuning of transformer models for Arabic biomedical question classification. Furthermore, we conducted several experiments on the MAQA [18] dataset to show that various pretrained transformer models are highly effective in classifying Arabic biomedical questions.

3. Research Methodology

We handle the Arabic biomedical question classification task by proposing transformers models as transfer learning models and fine-tuning their parameters on a large medical dataset. Our method includes several steps, including (1) Text preprocessing and tokenization, (2) Sentence representation, (3) Fine-tuning pretrained models and question classification. Figure 2 illustrates the architecture of our proposed system.

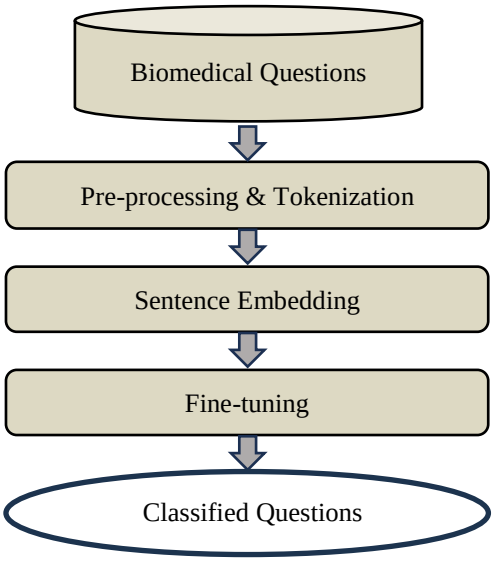

Figure 2. The architecture of the proposed system

3.1. Preprocessing and Tokenization

We apply a preprocessing stage for cleaning, tokenizing, and preparing textual data. It consists of removing extra words and strings including punctuation, stop words, and links, etc. Truncation is a sub operation which aims to reduce the inflectional forms of each word to a common base or root. After cleaning the data, we divide the question which is a short text into tokens using a reserved tokenizer for each pretrained Transformer model.

Table 2 displays the findings from applying different tokenizers to a sample Arabic biomedical question. We note that AraBERT tokenizer is more effective for Arabic because it takes into consideration the Arabic morphology by using Farasa Segmenter [35].

Table 2. Results of applying different tokenizers on an Arabic biomedical question

Question	ما هى الاعراض الاوليه لمرض السل؟
AraBERT Tokenizer	'ما', 'هى', 'الاعراض', 'الاولى', '##ه', 'لمرض', 'السل', '؟'
mBERT Tokenizer	'ما', 'هى', 'ال', '##ع', '##راض', 'ال', '##ول', '##يه', 'لم', '##رض', 'ال', '##سل', '؟'
mDistilBERT Tokenizer	"ما', 'هى', 'ال', '##ع', '##راض', 'ال', '##ول', '##يه', 'لم', '##رض', 'ال', '##سل', '؟'
mRoBERTa Tokenizer	'ما', 'ـــهى', 'ـــالا', 'عراض', 'ـــالاول', 'يه', 'ـــل', 'مرض', 'ـــالسل', '؟ـــ'
BioBERT Tokenizer	'م', '##ا', 'ه', '##ى', 'ا', '##ل', '##ا', '##ع', '##ر', '##ا', '##ض', 'ا', '##ل', '##ا', '##و', '##ل', '##ى', '##ه', 'ل', '##م', '##ر', '##ض', 'ا', '##ل', '##س', '##ل','؟'

In general, the input question is first tokenized into individual tokens as follows:

$$\{\mathrm{Tok}_1, \mathrm{Tok}_2, \ldots, \mathrm{Tok}_N\} \quad (1)$$

3.2. Sentence Embedding

In this step, we produce a representation of questions using pre-trained transformer models. The transformer encoder takes a sequence of words that constitute the Arabic question as input. All input questions must have the same length in terms of tokens; hence, padding is applied to have a single constant length. Real tokens and padding tokens will be distinguished using the attention mask. Lastly, these tokens are embedded into vectors before being processed in the neural network. The transformer output is also a sequence of vector, each vector corresponds to the input token with the same index.

Building vector representations that capture the contextual and semantic information of Arabic biomedical questions requires addressing the linguistic challenges posed by both the Arabic language and biomedical domain. In this context, transformer models are essential due to their ability to produce contextual embeddings that account for these linguistic complexities.

Transformers models use a self-attention mechanism, enabling each token in a sequence to assesses the relevance of all other tokens. This approach helps models to capture both nearby and distant relationships within a sentence, enhancing the understanding of context across the entire sequence.

Mathematically, the self-attention mechanism calculates attention scores α_{ij} between each token i and every other token j in the sequence, using their query Q_i and key K_j vectors, which determines how much influence token j has on token i at that particular layer. This can be expressed as:

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) \quad (2)$$

where d_k is the dimensionality of the key vectors, The attention scores are used to compute the final output representation of token i by taking a weighted sum over the value vectors V_j of all tokens:

$$Z_i = \sum_j \alpha_{ij} V_j \quad (3)$$

This process allows the model to generate contextual embeddings that adapt to the context of each token in the sentence, making it especially suitable for handling the semantic complexity of Arabic biomedical text. These embeddings are further fine-tuned through the model's pre-training on vast Arabic biomedical text corpora, allowing the transformer to leverage both its general Arabic linguistic understanding and biomedical domain knowledge.

Embeddings are extracted using the [CLS] strategy by adding the [CLS] token at the beginning of the text to generate sentence embeddings. This special classification token must be added to the beginning of every Arabic biomedical question, when the model is used for classification tasks. The [CLS] token is significant since each layer of the model inputs a list of word embeddings and produces the same number of embeddings in the output.

Another special token that is used in this strategy is the separator token [SEP]. This token is added to the tokenized sequence, and is useful for performing multi sentence tasks where the model is given many sentences as input and asked to perform a certain task. The initial and final input tokens in the sequence must be special tokens. The token sequence becomes:

$$[CLS, Tok_1, Tok_2, \ldots, Tok_N, SEP] \quad (4)$$

The resulting sequence is then given as input to the input layer.

Let ECLS be the embedding for the [CLS] token, and Ei be the embedding for the i-th token. Each token in the sequence is transformed into its corresponding embedding, capturing the semantic and contextual information. The sequence of embeddings then becomes:

$$\{E_{CLS}, E_1, E_2, \dots, E_N, E_{SEP}\} \quad (5)$$

The token embeddings are fed into the transformer encoder layers. Each layer consists of multi-head self-attention and feed-forward sub-layers. The transformer encoder produces a contextual representation for each token in the sequence. Let H^l be the hidden state at the $l-th$ transformer layer. The initial state is the input embeddings. The operation at each layer can be formulated as:

$$H^l = \text{TransformerEncoder}^l(H^{l-1}) \quad (6)$$

The final hidden states from the last transformer layer, H^l, represent the contextualized embeddings.

Sentence representations encapsulate meaning and contextual nuances by capturing the interrelations between words and sentences within a single compact representation. To capture robust representations, we make use of five different BERT models, including multilingual BERT, XLM-RoBERTa, DistilBERT, BioBERT, and AraBERT.

3.3. Fine-Tuning

The fine-tuning is a transfer learning method where we freeze a part of the layer weights of a pretrained model. This approach speeds up the training process and improves performance on the target task. In this work, we explore the aforementioned pretrained models, and then we fine-tune them on Arabic biomedical questions dataset.

Figure 3 illustrates the architecture of the proposed fine-tuned Transformer models for Arabic Biomedical Question Classification.

As shown in Figure 3, we assign h as the final hidden vector of the special [CLS] token and T_i as the final hidden vector for the $i-th$ input token. To obtain the probability distribution over the predicted output category c, we use the final hidden state h - which refers to the entire text - as an input for the feed-forward layer with Softmax classifier. This probability is calculated using the following formula [36]:

$$p(c|h) = \text{Softmax}(Wh) \quad (7)$$

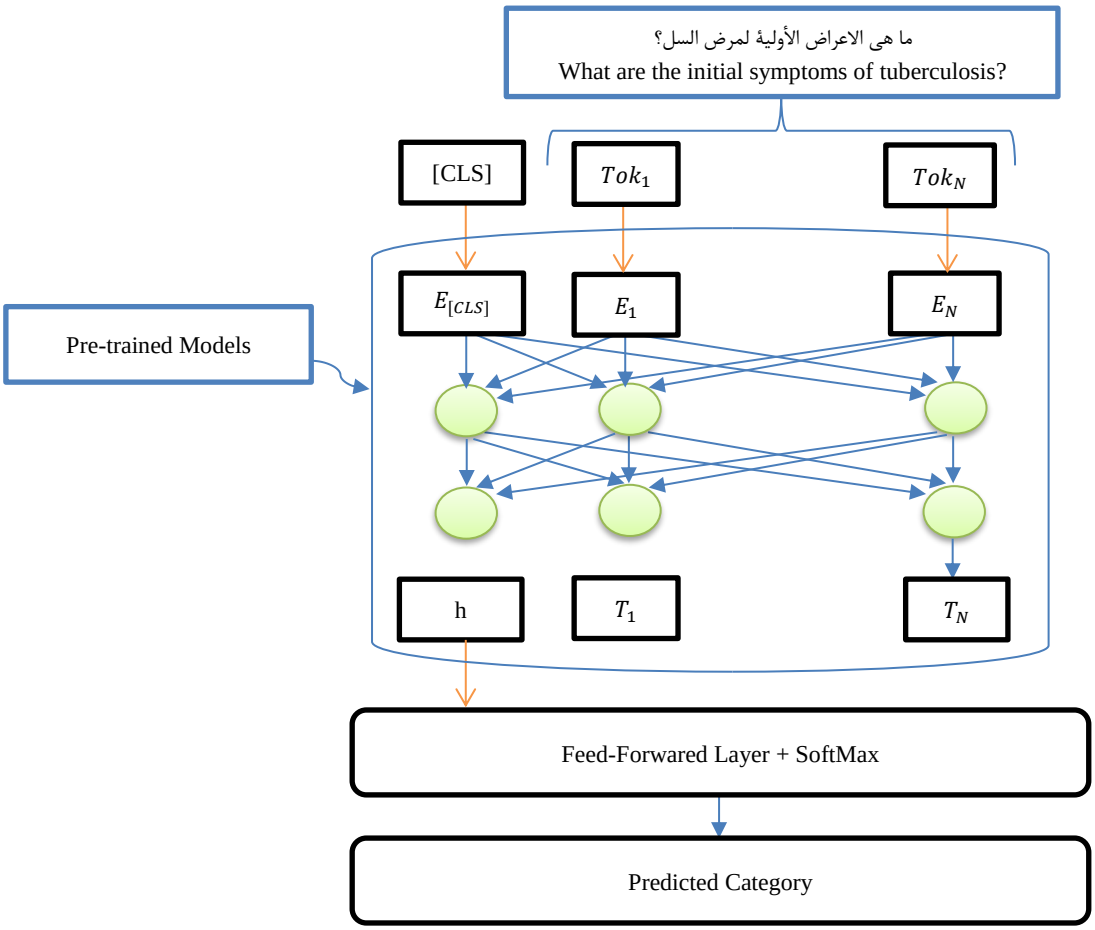

Figure 3. The architecture of the proposed fine-tuned Transformer models for Arabic Biomedical Question Classification

In order to maximize the log-probability of the correct category, all parameters from the pretrained models are utilized and jointly trained during the fine-tuning process.

4. Experimental Results

We conduct a series of experiments to evaluate the performance of the fine-tuned Transformer models for the Arabic biomedical questions classification. In this section, we explore the efficacy of fine-tuned BERT models using MAQA dataset. We use accuracy, F1-measure, precision, and recall as evaluation metrics to measure the performance of used models.

4.1. Dataset

All experiments are conducted using the medical Arabic dataset MAQA. It is the largest Arabic healthcare dataset, collected from many websites, including altibbi.com (70%), tbeeb.net (20%), and cura.healthcare (10%). It contains about 430K questions distributed into 20 biomedical classes [37]. The choice of the MAQA dataset can be justified by its exceptional size, diversity, and suitability for the Arabic biomedical domain. As the largest available dataset in this field, containing a wide variety of classes relevant to Arabic biomedical questions, its extensive size and diversity make it an ideal choice for training and evaluating models. In order to fine-tune the pretrained models, we use 247,763 questions classified into 10 biomedical Arabic classes. The dataset is split into 80% for training, with the rest reserved for testing. Table 3 shows the number of questions per class.

Table 3. Questions number per class

Number of questions	Category name	
103,683	Gynecology diseases	امراض نسائية
33,050	Musculoskeletal and joint diseases	امراض العضلات والعظام والمفاصل
22,373	Gastrointestinal diseases	امراض الجهاز الهضمى
21,773	Sexually transmitted diseases	الامراض الجنسية
20,207	Dentistry	طب الاسنان
15,368	Cardiovascular disease	امراض القلب والشرايين
14,439	Ophthalmology	امراض العيون
13,933	Ear nose and throat	انف اذن وحنجرة
1,596	Plastic surgery	جراحة تجميل
1,341	Blood Diseases	امراض الدم

4.2. Experimental Setup

We use a 10-core Apple M1 Pro processor chip and a 16-core Apple M1 Pro graphics card. For all models, we use the Adam optimizer with a learning rate of 3e-5. We add a layer of 10 nodes, which is the number of our classes, with the Softmax activation function and the Cross Entropy Loss function. The models are trained for 10 epochs with a batch size of 96. This training setup is selected after many trials, and it gives the best results for all models. The dataset is split into 80% for training and 20% for testing. Further, Table 4 shows the different hyperparameters used in our experiments.

Table 4. Experiment hyper-parameters values for transformers model

Hyper-parameter	Value
Optimizer	Adam
Learning rate	3e-5
Number of epochs	10
Loss function	Cross Entropy
Dropout rate	0.3

4.3. Performance Evaluation

Tables 5 and 6 summarize the results of the experiments during training and testing phases in terms of accuracy, F1 score, precision, and recall.

Classification accuracy which is the percentage of correctly predicted questions out of the total number of questions is one of the most widely used evaluation metrics. Precision refers to the percentage of correctly predicted classes among all positive classification predictions. Recall, on the other hand, is the proportion of correctly predicted positive classes out of the actual positive instances. The F1-score is the harmonic mean of recall and precision. Given a number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the following definitions apply for precision, recall, and F1-score:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

$$\text{F1-Score} = \frac{2\times(TP)}{2\times(TP)+FP+FN} \quad (11)$$

For the testing data, all of the fine-tuned models provide results ranging between 88% and 93%. AraBERT achieves the best results in terms of accuracy (93.31%), F1-measure (93.35%), Precision (93.42%), and recall (93.28%). Its high recall reflects its efficiency in identifying as many true positives as possible during training and testing phases, which is crucial in medical applications. One key reason for AraBERT's superior results is that it has been pre-trained on a vast and diverse corpus of Arabic data, enabling it to capture more nuanced semantic and contextual relationships.

Table 5. The training results of used models

Model	Accuracy	F1 score	Precision	Recall
AraBERT	98.91%	98.92%	98.96%	98.88%
XLM-RoBERTa	96.76%	96.78%	96.98%	96.59%
Multilingue DistilBERT	97.82%	97.84%	97.97%	97.71%
mBERT	97.35%	97.36%	97.53%	97.18%
BioBERT	93.18%	93.31%	94.51%	92.16%

Table 6. The testing results of used models

Model	Accuracy	F1 score	Precision	Recall
AraBERT	93.31%	93.35%	93.42%	93.28%
XLM-RoBERTa	92.50%	92.55%	92.83%	92.28%
Multilingue DistilBERT	91.94%	91.98%	92.20%	91.78%
mBERT	91.79%	91.87%	92.09%	91.65%
BioBERT	88.09%	88.46%	89.62%	87.34%

Since the multilingual DistilBERT and Multilingual BERT share the same architecture, they achieve nearly the same performances in terms of accuracy and F1-measure, outperforming the BioBERT model. However, DistilBERT is a light version of BERT, trained with fewer parameters, making it advantageous in biomedical applications where the speed of answering is crucial. Additionally, for all models, the F1 measure remains high, confirming that the models remain robust in the face of class imbalance. Despite BioBERT achieving good results, it shows the lowest performance among the models. This may be due to the tokenizer it employs, which splits sequences into individual characters, as shown in Table 2. This tokenization process fails to capture meaningful word or subword units.

The results confirm that transformer-based models are better for Arabic biomedical question classification because of their capacity to collect complex contextual data of text. Furthermore, model selection depends on resource availability and requirements because some models, such as DistilBERT, offer faster inference speed without sacrificing much performance.

Table 7 illustrates the predicted class for many questions that belong to "Blood Diseases امراض الدم" using various models. As demonstrated in the table, AraBERT outperforms the other models, always predicting the correct class. In contrast, the other models occasionally fail to identify the correct category.

Table 7. The predictions for different Arabic biomedical questions using the studied models

Sentences translation	Sentences	AraBERT	XLM-RoBERTa	mBERT	mDistilBERT	BioBERT
How can blood clotting disorders be treated?	كيف يمكن التعامل مع اضطرابات تخثر الدم؟	Correct	Incorrect	Incorrect	Correct	Incorrect
What are the causes of blood clotting?	ماهى اسباب تخثر الدم؟	Correct	Incorrect	Incorrect	Incorrect	Correct
What are the early symptoms of leukemia?	ما هى الاعراض المبكرة لسرطان الدم؟	Correct	Correct	Correct	Correct	Correct

The average computing time for the four top fine-tuned models while predicting the class of Arabic biomedical questions is shown in Table 8. Multilingual DistilBERT proves its performance in terms of execution time speed,

requiring only 0.37 seconds. It performs well in situations where fast response time is crucial. This efficiency can be attributed to the fact that distilBERT is a compressed version of BERT, trained with few parameters.

Table 8. The average computing time across top fine-tuned models

Model	Computing time (Second)
AraBERT	0.83
XLM-RoBERTa	0.71
mBERT	0.71
DistilBERT multilingual	0.37

After the evaluation of used models, we use the confusion matrices in Figure 4 and the classification report in Table 9 to detect the classes where the models exhibit confusion and identify those where the fine-tuned models fail. All models show notable confusion between "Sexually transmitted diseases الامراض الجنسية" and "Gynecology diseases امراض نسائية". This overlap can be attributed to the intersection of these two classes, as many sexually transmitted diseases are also addressed within the field of gynecology. The shared terminology and thematic overlap create additional challenges for the model, making it difficult to clearly distinguish between the two categories.

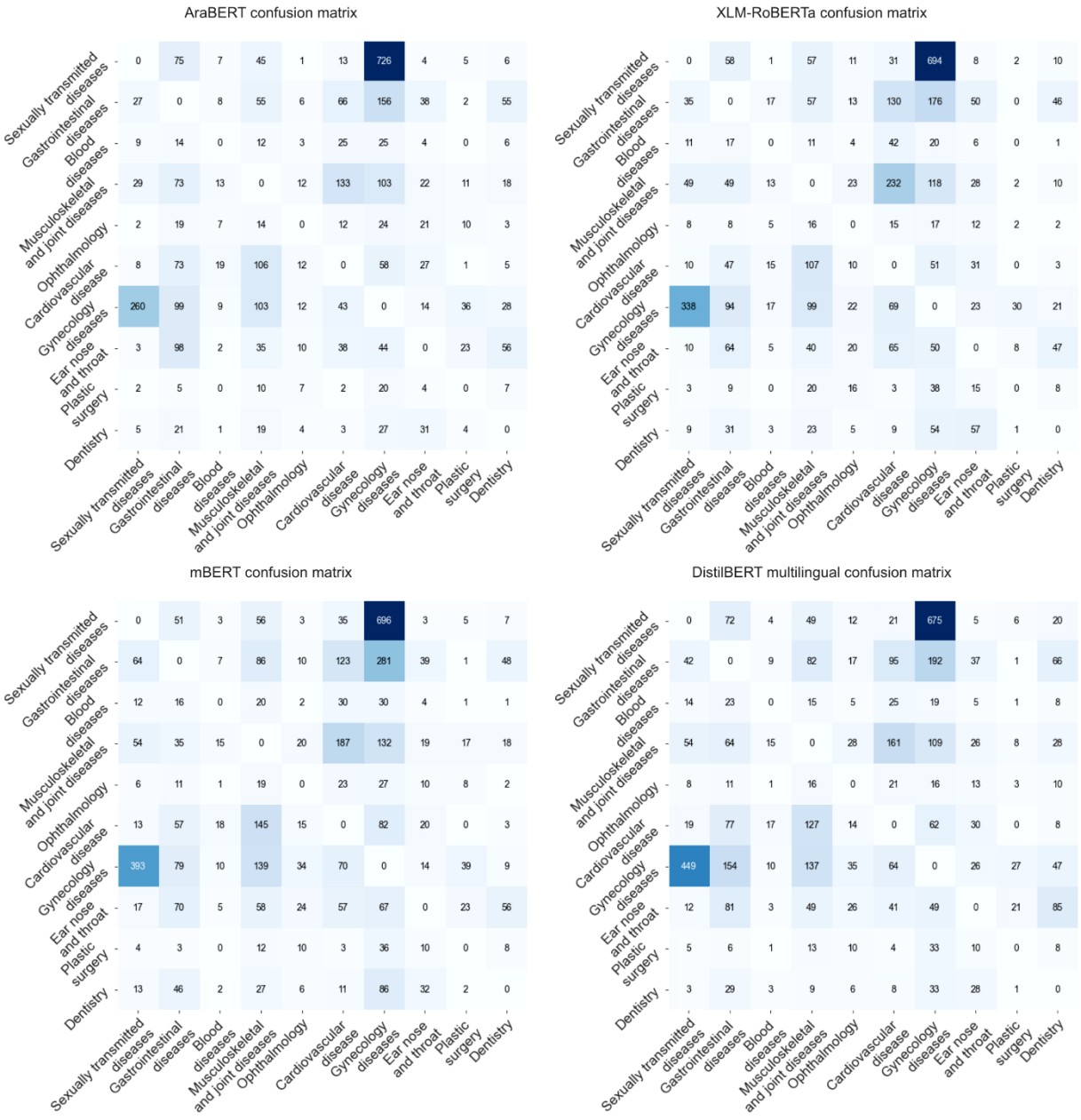

Figure 4. Confusion matrices of the four best models

Table 9. Classification report of top four models

Category translation	Category	AraBERT		XLM-RoBERTa		mBERT		mDistilBERT	
		Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Sexually transmitted diseases	الامراض الجنسية	0.91	0.80	0.88	0.80	0.86	0.80	0.85	0.80
Gastrointestinal diseases	امراض الجهاز الهضمي	0.90	0.91	0.92	0.89	0.91	0.86	0.89	0.88
Blood Diseases	امراض الدم	0.72	0.64	0.68	0.59	0.72	0.57	0.71	0.57
Musculoskeletal and joint diseases	امراض العضلات والعظام والمفاصل	0.94	0.94	0.93	0.92	0.92	0.93	0.93	0.93
Ophthalmology	امراض العيون	0.98	0.96	0.96	0.97	0.96	0.96	0.95	0.97
Cardiovascular disease	امراض القلب والشرايين	0.89	0.90	0.82	0.91	0.83	0.88	0.86	0.88
Gynecology diseases	امراض نسائية	0.94	0.97	0.94	0.97	0.93	0.96	0.94	0.95
Ear nose and throat	انف اذن وحنجرة	0.94	0.89	0.92	0.89	0.94	0.87	0.93	0.87
Plastic surgery	جراحة تجميل	0.73	0.81	0.81	0.63	0.69	0.71	0.76	0.70
Dentistry	طب الاسنان	0.95	0.97	0.96	0.95	0.96	0.94	0.93	0.97

As presented in classification report in Table 9, most errors occur when predicting the "Blood Diseases امراض الدم" class. For instance, the AraBERT model achieves a precision of 72% when classifying questions of this class; 172 out of the 270 were correctly classified. In contrast, it demonstrates a higher precision of 98% to classify the "Ophtalmology امراض العيون" questions; of the 2843 questions, 2731 were correctly classified. The errors committed when predicting certain classes can be justified given that although a question belongs to a specific discipline, it may contain words from other biomedical subfields. For example, in the case of "Blood Diseases امراض الدم" and "cardiovascular disease امراض القلب والشرايين" many terms are shared between these two classes, leading to potential misclassifications. Additionally, the lower performance in classifying "Blood Diseases امراض الدم" can be attributed to the relatively small number of samples available for this class, making it more complicated for the model to learn accurate representations.

Figure 5 presents F1 scores of different classes using the top four fine-tuned models. As evidenced by the classification report, all models demonstrate strong performance in predicting questions related to Ophthalmology, achieving high accuracy rates. However, the classification of Blood Diseases remains a challenging task for these models.

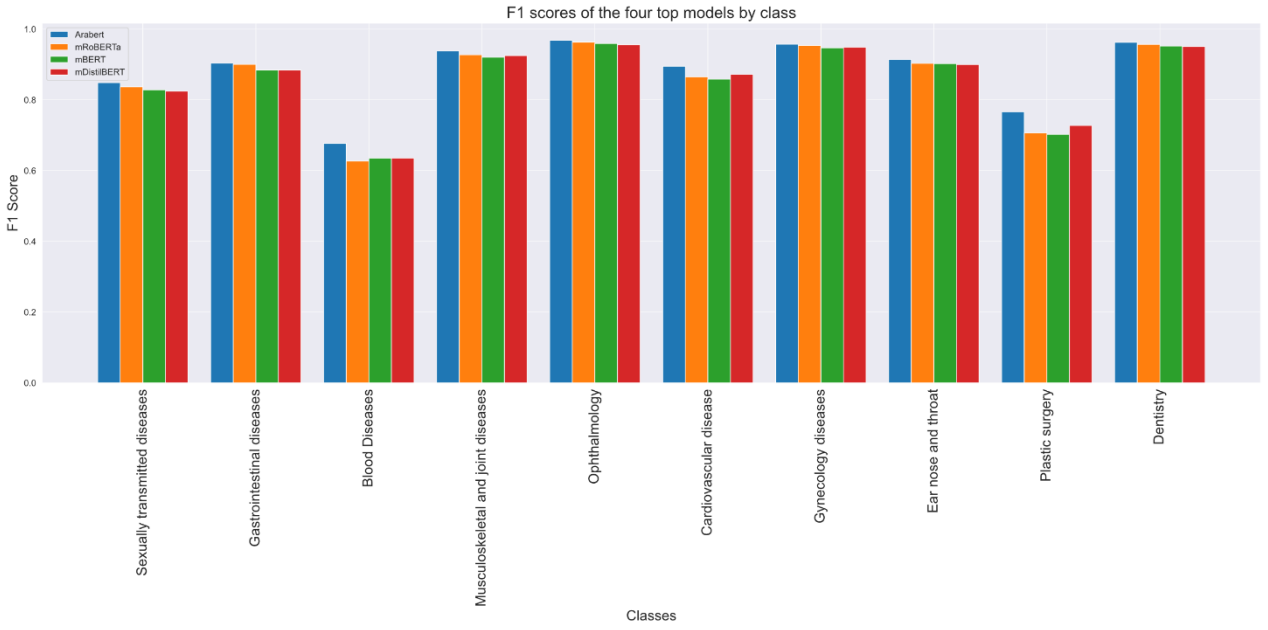

Figure 5. F1-scores of the top four models per class

The precision of pain location is a key to classifying questions, developing methods to accurately identify and interpret key location-indicating words such as "و" (and) or "من" (from) to refine classification outcomes will be an interesting perspective to work on in the coming works. A slight modification of a sentence by adding the pain location can have a significant impact on the assigned class, as shown in Table 10.

Table 10. The impact of pain location on Arabic biomedical sentence classification

Classes translation	Classes	Sentences translation	Sentences
Respiratory diseases	امراض الجهاز التنفسى	I feel a pain in my chest area	اشعر بالم فى منطقه الصدر
Cardiovascular disease	امراض القلب والشرايين	I feel pain in the chest area and near the heart	اشعر بالم فى منطقه الصدر وقرب القلب
General surgery	جراحة عامة	I want to do a nail transplant	اريد اجراء عمليه لنزع الظفر
Ophthalmology	امراض العيون	I want to have a nail removed from my eye	اريد اجراء عمليه لنزع الظفر من عينى

5. Conclusion

Biomedical question classification is an essential component for enhancing various biomedical applications, including question answering, decision support, and retrieval systems. In this paper, we have proposed an Arabic biomedical question classification system. We built vector representations that capture contextual and semantic information within biomedical text and questions, which present combined challenges related to the complex morphology of the Arabic language and the specialized terminology of the biomedical domain. Our representation is able to adapt the rich information encoded in pretrained transformer models (BERT, AraBERT, BioBERT, RoBERTa, and DistilBERT). Using the encoder part of transformer architecture and an attention mechanism, we extracted essential information from the Arabic biomedical text. These representations strengthen our classification system, enabling it to accurately predict the relevant category of Arabic biomedical questions. We carried out several experiments using the biomedical dataset MAQA, which is an Arabic Healthcare Q&A dataset. The obtained results show that AraBERT outperforms other models and reaches 93.35% in terms of F1-score. However, mBERT, XLM-RoBERTa, and multilingual DistilBERT provided equally promising results for Arabic biomedical question classification. Furthermore, we compared the BERT model with its light version, DistilBERT, which achieved a good result from the first training epoch. Due to the sensitive nature of the biomedical field, the speed of disease identification is crucial. Fast models are an option; some of these include DistilBERT as a good example proving that. Future work will focus on combining multiple models using an ensemble technique to further enhance Arabic biomedical question classification. We intend also to integrate this component in our future Arabic biomedical question answering system.

6. Declarations

6.1. Author Contributions

Conceptualization, I.A.T.; methodology, I.A.T.; software, I.A.T.; validation, H.A. and S.O.E.A.; formal analysis, I.A.T.; investigation, I.A.T., H.A., and S.O.E.A.; resources, I.A.T., H.A., and S.O.E.A.; data curation, I.A.T.; writing—original draft preparation, I.A.T.; writing—review and editing, H.A. and S.O.E.A.; visualization, I.A.T.; supervision, H.A. and S.O.E.A.; project administration, H.A. and S.O.E.A.; funding acquisition, I.A.T., H.A., and S.O.E.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available in the article.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Sarrouti, M., & El Alaoui, S. O. (2017). A machine learning-based method for question type classification in biomedical question answering. Methods of Information in Medicine, 56(3), 209–216. doi:10.3414/ME16-01-0116.

[2] Xu, S., Cheng, G., & Kong, F. (2016). Research on question classification for automatic question answering. 2016 international conference on Asian language processing (IALP), 218–221. doi:10.1109/IALP.2016.7875972.

[3] Babu, A., & Boddu, S. B. (2024). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. Exploratory Research in Clinical and Social Pharmacy, 13, 100419. doi:10.1016/j.rcsop.2024.100419.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 4171–4186.

[5] Tama, B. A., & Lim, S. (2020). A comparative performance evaluation of classification algorithms for clinical decision support systems. Mathematics, 8(10), 1–24. doi:10.3390/math8101814.

[6] Hassan, E., Abd El-Hafeez, T., & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. Scientific Reports, 14(1), 01 2024. doi:10.1038/s41598-024-51615-5.

[7] Momtazi, S. (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. Information Processing and Management, 54(3), 380–393. doi:10.1016/j.ipm.2018.01.001.

[8] Hamza, A., En-Nahnahi, N., Zidani, K. A., & El Alaoui Ouatik, S. (2021). An Arabic question classification method based on new taxonomy and continuous distributed representation of words. Journal of King Saud University - Computer and Information Sciences, 33(2), 218–224. doi:10.1016/j.jksuci.2019.01.001.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 1–12.

[10] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for sentence summarization. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 379–389.

[11] Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. Mining Text Data, 9781461432234, 77–128. doi:10.1007/978-1-4614-3223-4_4.

[12] Mikolov, T., Sutskever, I., Chen, K., *Corrado*, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. ArXiv Preprint, ArXiv:1310.4546. doi:10.48550/arXiv.1310.4546.

[13] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1532–1543. doi:10.3115/v1/d14-1162.

[14] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. doi:10.1162/tacl_a_00051.

[15] Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data, 6(1), 52. doi:10.1038/s41597-019-0055-0.

[16] Lahbari, I., & El Alaoui, S. O. (2024). Exploring Sentence Embedding Representation for Arabic Question Answering. International Journal of Computing and Digital Systems, 15(1), 1229–1241. doi:10.12785/ijcds/150187.

[17] Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for Arabic language understanding. arXiv preprint, arXiv:2003.00104. doi:10.48550/arXiv.2003.00104.

[18] Abdelhay, M., & Mohammed, A. (2022). MAQA: Medical Arabic Q & A dataset. Harvard Dataverse, Cambridge, United States.

[19] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (1907). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. doi:10.48550/arXiv.1907.11692.

[20] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. doi:10.1093/bioinformatics/btz682.

[21] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint, arXiv:1910.01108. doi:10.48550/arXiv.1910.01108.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December, 5999–6009.

[23] Mutabazi, E., Ni, J., Tang, G., & Cao, W. (2023). An Improved Model for Medical Forum Question Classification Based on CNN and BiLSTM. Applied Sciences (Switzerland), 13(15), 8623. doi:10.3390/app13158623.

[24] Vihikan, W. O., & Trisna, I. N. P. Indonesian health question multi-class classification based on deep learning. Journal of Information Systems and Informatics, 6(3), 1931–1944.

[25] Mansour, M., Tohamy, M., Ezzat, Z., & Torki, M. (2020). {A}rabic Dialect Identification Using {BERT} Fine-Tuning. Proceedings of the Fifth Arabic Natural Language Processing Workshop, 308–312.

[26] Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., & Dai, L. (2021). ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. Complexity, 2021, 1–6. doi:10.1155/2021/6633213.

[27] Zafar, A., Sahoo, S. K., Varshney, D., Das, A., & Ekbal, A. (2024). KIMedQA: towards building knowledge-enhanced medical QA models. Journal of Intelligent Information Systems, 62(3), 833–858. doi:10.1007/s10844-024-00844-1.

[28] Hammoud, J., Vatian, A., Dobrenko, N., Vedernikov, N., Shalyto, A., & Gusarova, N. (2021). New Arabic Medical Dataset for Diseases Classification. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13113 LNCS, 196–203. doi:10.1007/978-3-030-91608-4_20.

[29] Al-Smadi, B. S. (2024). DeBERTa-BiLSTM: A multi-label classification model of Arabic medical questions using pre-trained models and deep learning. Computers in Biology and Medicine, 170, 107921. doi:10.1016/j.compbiomed.2024.107921.

[30] Yu, H., Liu, C., Zhang, L., Wu, C., Liang, G., Escorcia-Gutierrez, J., & Ghoneim, O. A. (2023). An intent classification method for questions in "Treatise on Febrile diseases" based on TinyBERT-CNN fusion model. Computers in Biology and Medicine, 162, 107075. doi:10.1016/j.compbiomed.2023.107075.

[31] Kofi Akpatsa, S., Lei, H., Li, X., Kofi Setornyo Obeng, V.-H., Mensah Martey, E., Clement Addo, P., & Dodzi Fiawoo, D. (2022). Online News Sentiment Classification Using DistilBERT. Journal of Quantum Computing, 4(1), 1–11. doi:10.32604/jqc.2022.026658.

[32] Aftan, S., & Shah, H. (2023). Using the AraBERT Model for Customer Satisfaction Classification of Telecom Sectors in Saudi Arabia. Brain Sciences, 13(1), 147. doi:10.3390/brainsci13010147.

[33] El-Alami, F. zahra, Ouatik El Alaoui, S., & En Nahnahi, N. (2022). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. Journal of King Saud University - Computer and Information Sciences, 34(10), 8422–8428. doi:10.1016/j.jksuci.2021.02.005.

[34] Houssein, E. H., Mohamed, R. E., Hu, G., & Ali, A. A. (2024). Adapting transformer-based language models for heart disease detection and risk factors extraction. Journal of Big Data, 11(1). doi:10.1186/s40537-024-00903-y.

[35] Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, 11–16. doi:10.18653/v1/n16-3003.

[36] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11856 LNAI, 194–206. doi:10.1007/978-3-030-32381-3_16.

[37] Abdelhay, M., Mohammed, A., & Hefny, H. A. (2023). Deep learning for Arabic healthcare: MedicalBot. Social Network Analysis and Mining, 13(1), 71. doi:10.1007/s13278-023-01077-w.