AraMed: Arabic Medical Question Answering using Pretrained Transformer Language Models

Ashwag Alasmari[1], Sara Alhumoud[2], Waad Alshammari[3]
[1]Department of Computer Science, King Khalid University
[2]Imam Mohammad Ibn Saud Islamic University
[3]King Salman Global Academy for Arabic Language
Saudi Arabia
aasmry@kku.edu.sa, sohumoud@imamu.edu.sa, walshammari@ksaa.gov.sa

Abstract

Medical Question Answering systems have gained significant attention in recent years due to their potential to enhance medical decision-making and improve patient care. However, most of the research in this field has focused on English-language datasets, limiting the generalizability of MQA systems to non-English speaking regions. This study introduces AraMed, a large-scale Arabic Medical Question Answering dataset addressing the limited resources available for Arabic medical question answering. AraMed comprises of 270k question-answer pairs based on health consumer questions submitted to online medical forum. Experiments using various deep learning models showcase the dataset's effectiveness, particularly with AraBERT models achieving highest results, specifically AraBERTv2 obtained an F1 score of 96.73% in the answer selection task. The comparative analysis of different deep learning models provides insights into their strengths and limitations. These findings highlight the potential of AraMed to advance the creation and development of resources specific to Arabic medical question answering research and development.

Keywords: natural Language processing, medical question answering, answer selection, language models

1. Introduction

With the ever-increasing volume of health information available online, finding accurate answers to specific medical requests is becoming more difficult for health consumers (Alasmari & Zhou, 2019, 2021). Medical question and answer (MQA) platforms provide an online space where users can ask direct questions and medical experts can provide answers to these questions. This is the most intuitive ways for people to seek information online, especially when search engines fail to provide relevant and accurate results (Liu et al., 2012). In addition, MQA enables consumers to avoid long wait times when seeking health information, particularly if they require information quickly or need information about managing a health condition at home. Wicks et al. (Wicks et al., 2010) demonstrated that an online medical platform can assist users in managing their own symptoms, dealing with side effects, connecting with other patients, and seeking medication advice.

The explosive demand and growth of users and questions are creating a bottleneck for the limited number of doctors. Therefore, developing ways of automatically answer questions using the information from previously answered questions is important. Particularly for delivering responses that direct consumers to the potentially relevant information about their concern. Ideally, implementing effective solutions can result in the workload of doctors being greatly reduced and the consumer's experience of online medical systems being enhanced. That is, lower overhead cost on the medical institutions, the speed into which the answer is fetched, and harnessing the intelligence of already available MQA forums and datasets. MQA systems are designed to solve the automatic QA issue by the development of several tracks and tasks including question ranking, question similarity, and answers selection. Question ranking is the task of ranking a set of questions according to their relevance to a given question whereas question similarity is concerned with detecting the semantic similarity between two questions (Mishra & Jain, 2016). We focus on the task of answer selection in this work, which is the task of choosing the best answer to a given question from a set of possible answers.

Question Title	حمى ابنتي 38.4 من دون اعراض أخرى. "My daughter has a fever 38.4 and no other symptoms."
Question Description	تعاني ابنتي من حمى 38.4 منذ يومين ولا تعاني من اي اعراض اخرى ماذا افعل هل اخذها للطبيب ام انتظر. "My daughter has a fever 38.4 for two days and she does not have any other symptoms. What should I do? Should I take her to the doctor or wait?"
Question Category	امراض الأطفال "Pediatric"
Relevant Answer (RA)	أنصحك بعرض ابنتك علي طبيب الاطفال للفحص الطبي ومعرفه سبب الحرارة. مادامت الحرارة مستمرة منذ يومين. لإجراء الفحص والتشخيص وكتابه العلاج اللازم. تمنياتي لطفلك بالصحة والسلامة. "I advise you to visit a pediatrician for a medical examination and to find out the cause of the fever. as it continued for two days. This is to conduct the examination, diagnosis and write the necessary treatment. I wish your child health and peace."
RA Doctor specialty	طب أطفال "Pediatrics"
Irrelevant Answer (IA)	اشتباه عرق النساء وهذا عرض لمرض محتاج تعمل اشعه بالرنين المغناطيسي. "It might be sciatica, and this is a symptom of a disease that needs an MRI scan."
IA Doctor specialty	الروماتيزم والمفاصل "Rheumatology and joints"

Figure 1: An example of MQA sample in AraMed. The English translation is not a part of the corpus.

OSACT 2024 Workshop, pages 50–56

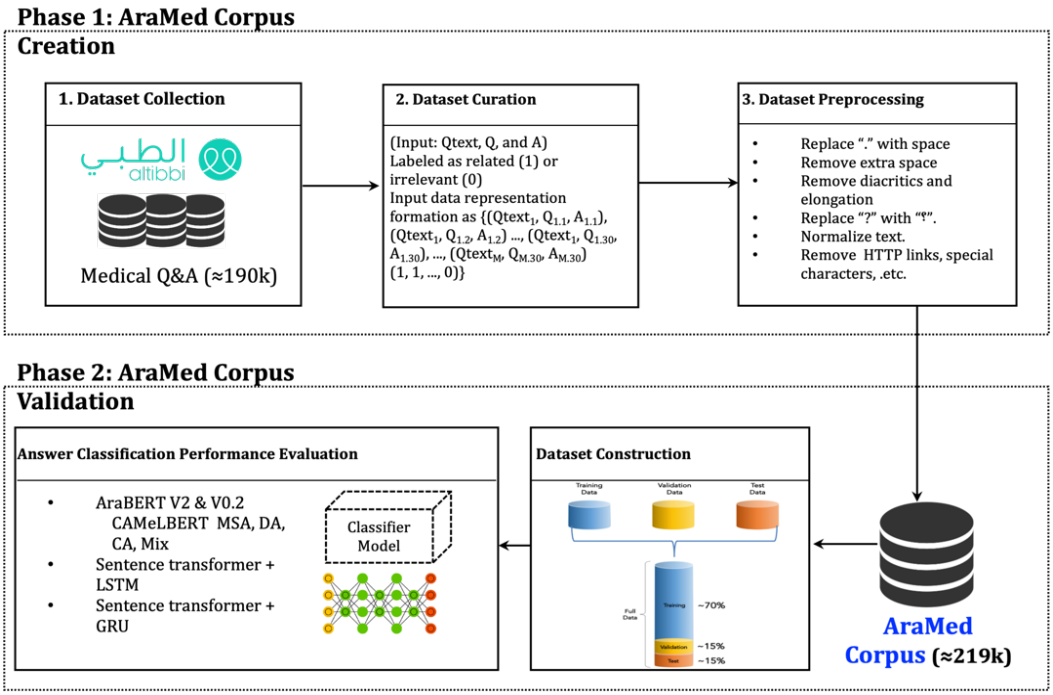

Figure 2: Flow diagram of AraMed dataset creation and validation.

Answer selection has received an increasing attention recently. There are several benefits of answer selection in MQA. Answer selection can improve the accuracy of the answers, reduce the time it takes to answer questions, and increase user satisfaction by providing users with the most relevant answers to their questions. The aim is to identify which of the candidate answers contain the correct answer to a question given a question and a set of candidate answers (Lai et al., 2018). Traditional approaches to answer selection typically rely on recurrent neural networks, including long short-term memory networks and Seq2seq models (Roy et al., 2023). However, the recent development of attention mechanisms, like those found in Transformers, BERT models and large language models has shifted the focus of answer selection research in the question answering research (Zhang et al., 2021; Guo et al., 2022).

The quality and accuracy of MQA systems are dependent mainly on two factors: the size and quality of the corpus and the efficacy of the machine learning model. In English multiple MQA corpora are available that can be used to build more efficient MQA systems. However, the Arabic language currently lacks such comprehensive resources.

To date, the contribution in the Arabic MQA corpora is limited to two, ARmed (Fehri et al., 2022) and CQA-MD (Adlouni et al., 2019; Balla et al., 2022). This could be due to the lack of resources, the morphological complexity of Arabic, and the multiple forms and dialects that are used for online expression and communication (Boudjellal et al., 2020). In particular, the first contribution (Fehri et al., 2022) created a corpus, ARmed, which included 350 manually annotated questions along with the corresponding responses which were collected from several resources for the purpose of automatically answering medical questions in Arabic. The second contribution is a widely used corpus, CQA-MD, which includes 57,764 question-answer pairs. This dataset was used for the SemEval 2016 and 2017 workshops (Agirre et al., 2016; Nakov et al., 2017), which were collected from three Arabic medical websites including WebTeb, Al-Tibbi, and Islamweb. Several studies built MQA systems utilizing this corpus (Adlouni et al., 2019; Balla et al., 2022). However, CQA-MD is small and unbalanced and contains only 1,943 directly relevant, 24,265 relevant, and 31,556 irrelevant answers. Direct relevant represents the correct answers and is only 4% of the dataset, and it is a low number not sufficient for deep learning purposes. In addition, the average sequence length of combining the new query with the question answer pairs is 1,101 which is considered large as an entry for transformer-based models. Another contribution related to the MQA (Faris et al., 2022) introduced an annotated corpus of 75,000 medical questions only, each of which is assigned one of 15 medical topics; the goal for this was to build an automated question classification. Clearly, the need for an Arabic MQA corpus of a decent size to enable creation of advanced MQA systems is evident.

To address this gap, this paper presents AraMed, a large-scale and a balanced MQA corpus containing 270k pairs of question and answer, along with the question categories, age and gender of the questioner, the specialty of the physician who answered the question, and the date. AraMed is constructed from Altibbi platform data, a popular medical website in the MENA region. This work has two major contributions:

- The collection and curation of an annotated dataset of 270k Arabic question-answer pairs based on health consumer questions

submitted to the Altibbi platform. The code and datasets will be made available upon request.

- Performing preliminary experimentation on the dataset using state-of-the-art pretrained models that are trained specifically for answer selection tasks to benchmark the dataset for future work.

	Train	Test
Questions	109,834	27,459
Answers	219,668	54,918
Categories	85	78
Relevant Answers	109,834	27,459
Irrelevant Answers	109,834	27,459

Table 1: Description of the AraMed corpus.

2. Dataset

In this section, we describe the four main aspects related to the creation of AraMed corpus: data collection, data curation, and data preprocessing. Those aspects are to be described below. Figure 2 illustrates the architecture of the AraMed corpus creation and validation process.

2.1 Data Collection

All data was obtained from the medical platform, Altibbi.com. This platform aims to provide users with reliable, up to date, simplified medical information in Arabic. The website includes thousands of medical articles, a medical glossary, a section of questions and answers (Q&A), the most recent medical news, and telehealth services and consultations. For this work, we collected the most recent Q&A, ranging from 2020 to 2021, and the most useful questions by the vote of users on the website. The resulting dataset includes 219,668 unique Q&A pairs.

2.2 Data Curation

Our corpus involved several preprocessing steps, beginning with the removal of questions with no answers and the removal of duplicate answers for the same question. Additionally, for the learning process to be effective, the data needed to include both questions with correct answers and the same questions with irrelevant answers. As the dataset does not contain irrelevant answers, a rule-based automated annotation approach was applied (Thuwaini & Alhumoud, 2022). This approach is summarized as follows:

For each question that have *n* correct answers, *n* candidate irrelevant answer is appended. To nominate the candidate irrelevant answer, a sliding window is moved over the answers by *m*, where *m=i+10* and *i* denominates the current row. The candidate m^{th} answer is checked against 2 conditions. The first is the category; if both the candidate m^{th} answer category and the current i^{th} answer category are the same, then the irrelevant answer is skipped, and *m* is incremented by 5. This is to confirm that the two answers, the current and the candidate, are not the same or similar. The second condition is that the m^{th} answer has more than 7 words. This is to ensure that the answer has a substantial content that would aid the learning process, as answers with a smaller number of words are not indicative or informative. If the candidate answer satisfies both conditions then it is used as the current question's irrelevant answer, its related data is appended to the question and *i* is incremented by 1 to proceed to the next question.

A unique ID is added for every question, relevant answer, and irrelevant answer. After curation and annotation process, the columns added to the dataset are question ID, relevant answer ID, irrelevant answer, and irrelevant answer ID, and answer date of irrelevant answer.

2.3 Data Preprocessing

The following preprocessing steps are applied to the data:

- Replace "." with space, since it has been used as delimited between tokens, such as, نعم...طبيعي.
- Remove extra spaces.
- Remove diacritics and elongation "ـ" using Pyarabic, an Arabic plugin tool for Python.
- Remove HTTP links, special characters, English alphabet, English numbers, Arabic numbers, and extra spaces using regular expressions, a built-in Python package.
- Normalize text, that is replace the letters ٱ, ٳ, أ, آ with ا
- Replace English question marks "?" with Arabic question marks "؟" to unify the letters.

3. Experimental design

We evaluated the performance of the answer selection classification model using several variants of transformer-based models on various data sets compiled from our developed corpus. In particular, we experimented with multiple variants of transformer-based models and compared their performances on the test set. All models were trained on the same training set and hyperparameters were optimized using the same validation set.

3.1 Transformer-based Models

Transformers achieved state-of-the-art performance in multiple NLP tasks. A transformer is a deep learning language model that is trained on a huge corpus to solve sequence to sequence tasks while easily handling long-range dependencies and predict the probability of the next token given the previous one (Wolf et al., 2020).

In our experiments, we used several Bidirectional Encoder Representations from Transformers (BERT) variants (Devlin et al., 2019). BERT is a transformer-based language model that represents the embedding based on the context of the sentence. There are various Arabic pretrained language models available, including AraBERT (Othman et al., 2020) and CAMeLBERT (Inoue et al., 2021), all of them are

based on transformers produced recently by the Arabic NLP community. The pretrained AraBERT and CAMeLBERT language models are fine-tuned with the target dataset using TensorFlow Estimators[1] and transformers for sequence classification[2], respectively. The maximum sequence length selected was 256 tokens in each QA pair, since the average length of questions and answers is 23.9 and 34.2 respectively. From previous experience using a larger sequence (Thuwaini & Alhumoud, 2022), length (512) has no effect on the accuracy and doubles the execution time.The models are as follows:

- AraBERT version 2 (AraBERT v2) is a version of the AraBERT (Othman et al., 2020), which uses pre-segmentation based on Farasa segmentation (Abdelali et al., 2016) that segment words into stems, prefixes, and suffixes.
- AraBERT version 0.2 (AraBERTv0.2) is a version of the AraBERT(Othman et al., 2020), which uses BERT-compatible tokenization, and the text is preprocessed without the use of Farasa segmentation.
- CAMeLBERT-MSA (Inoue et al., 2021) which pretrained with modern standard Arabic
- CAMeLBERT-DA (Inoue et al., 2021) which pretrained with dialectal Arabic
- CAMeLBERT-CA (Inoue et al., 2021) which pretrained with classical Arabic
- CAMeLBERT-Mix (Inoue et al., 2021) which is pretrained with a mix of Modern Standard Arabic (MSA), dialectal, and classical Arabic.

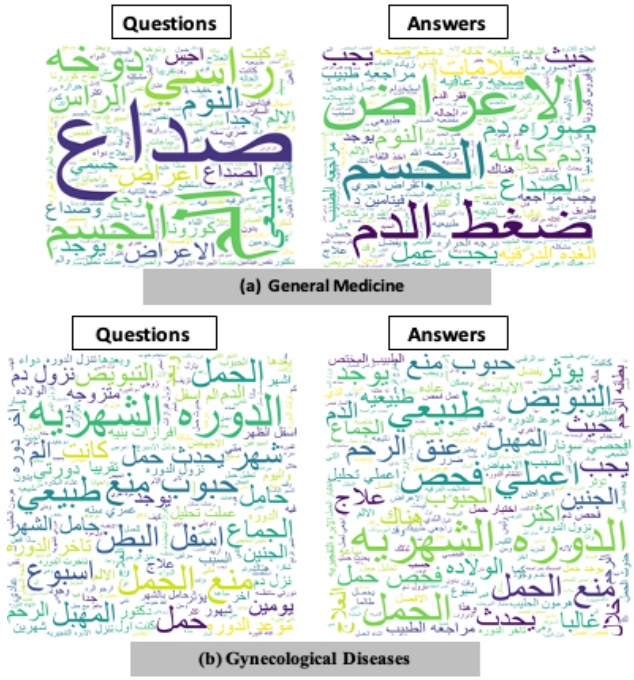

Figure 3: Frequently discussed medical topics in the top two categories in AraMed

3.2 Hybrid Sentence Embedding Models

To tackle the challenge of contextually representing the data, we developed a hybrid model. The first step involves using AraBERTv0.2, in combination with Sentence-BERT (Reimers & Gurevych, 2019), to encode the text data. Unlike BERT, which can capture the contextual relationships between words and phrases, Sentence-BERT is designed to represent entire sentences. By using these two models together, the input text can be encoded into a fixed-dimensional vector representation with 768 dimensions that capture the semantic similarity. Next, the sentence's contextual embedding is fed into a Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) model. A dense layer with a sigmoid activation function is added to the model to predict the similarity.

4. Results

In this section, we present the characteristics of AraMed where we analyze various aspects of the annotated corpus, including the category distribution, and the top medical topics discussed in the corpus. Additionally, we present the performance of various deep learning models on the AraMed dataset for the answer selection task.

4.1 Descriptive Results

The AraMed corpus contains more than 270k questions and answers pairs about numerous medical topics, with an average of 23.9 tokens per medical question and an average of 34.2 tokens per answer. The dataset statistics are in Table 1 and an example from the AraMed corpus is shown in Figure 1.

We listed the categories where the questions in our corpus posted, and total number of questions for each category. For a detailed breakdown of descriptive statistics by category, please refer to Appendix Table 3. The three topics with the highest number of questions are gynecological diseases, sexual health, and general medicine, with more than 10,000 questions per topic.

In this collection, we used word clouds to visualize the highest frequency words in the top two categories to gain insights into the most frequent unigrams and bigrams used. Figure 3 shows word clouds of the most frequent words associated with the questions and answers for top two categories, which are general medicine and gynecological diseases. In general, the diagram shows that the most frequently occurring terms in questions from the general medicine category are صداع "*headache*", راس *"head"*, "*drowsy*", الأعراض *"symptoms"*, ألم الجسم *"body aches"*, and النوم *"sleep"*. Word clouds representing most frequent answers for questions in general medicine are as follows: الأعراض *"symptoms"*, ضغط الدم *"blood pressure"*, سلامات *"feel better"*, مراجعة طبيب *"doctor's visit"*, صورة دم

[1] https://github.com/aub-mind/arabert/

[2] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

"blood sample". For questions related to gynecological diseases, the most frequent words are دورة شهرية *"menstruation"*, الحمل *"pregnancy"*, حبوب منع حمل *"birth control pill"*, منع الحمل *"contraception"*, and اسفل البطن *"lower abdomen"*. Meanwhile, common words in the answers are تبويض *"ovulation"*, طبيعي *"normal"*, اعملي فحص *"do a check"*.

4.2 Experimental Results

The results presented in Table 2 indicate that the AraBERT models emerged as the top performers. The AraBERTv2 achieved the highest results, with F1 score 96.73%, demonstrating its effectiveness in capturing the contextual semantic of the Arabic language. AraBERTv0.2 closely followed, showing that both versions of AraBERT are highly capable in this task. The CAMeLBERT models, trained on different Arabic text forms, showed good performance, with CAMeLBERT-MSA leading among them.

On the other hand, hybrid models integrating Sentence Transformers with GRU and LSTM showed limitations, indicating the complexity of mixing sentence-level embeddings with sequence models. This finding suggests there is potential for advancing sentence embedding techniques to achieve more precise contextual and semantic representations.

Model	**Accuracy (%)**	**F1 (%)**
AraBERTv2	96.71	96.73
AraBERTv0.2	96.54	96.56
CAMeLBERT-MSA	95.78	95.74
CAMeLBERT-DA	93.72	93.62
CAMeLBERT-CA	94.38	94.31
CAMeLBERT-Mix	95.19	95.15
Sentence transformer AraBERT + GRU	86.29	89.49
Sentence transformer AraBERT + LSTM	86.81	86.57

Table 2: Finetuning and Hybrid models performance

5. Conclusion and Future Work

This paper presents large-scale Arabic Medical Question Answering Corpus (AraMed). AraMed addresses the need for a large-scale resource to study Arabic medical question answering system, answer selection and related tasks. Our evaluation demonstrates the value of AraMed. First, it serves as a benchmark for further research on answer selection and related tasks, achieving solid baseline performance with different variants of pretrained transformer language models. Moving forward, to explore future directions, tracks including advanced models, multimodal incorporation, and dialect-specific model adaptation, holds immense potential for building more efficient, informative, and user-centric medical question answering systems. Also, the QA model could be enhanced by sampling irrelevant answers that are semantically similar. That is done by not skipping similar categories when selecting irrelevant answers. By addressing these future directions, AraMed can become a crucial resource for advancing research in Arabic medical question answering and foster further exploration in related NLP tasks.

6. Ethical Consideration

We ensured the dataset protects user privacy. We anonymized question URLs and user IDs. We went a step further by removing any additional details that could potentially identify individuals, such as email addresses, physical locations, names, or website links. This comprehensive anonymization process allows researchers to safely use the data without the risk of uncovering personal information. AraMed is intended only for research purposes. Following a review process to understand the requester's intent and ensure responsible use, we plan to share the corpus upon research requests to facilitate further advancement in medical question answering research in Arabic.

7. References

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 11–16. https://doi.org/10.18653/v1/N16-3003

Adlouni, Y. El, Rodríguez, H., Meknassi, M., El Alaoui, S. O., & En-nahnahi, N. (2019). A multi-approach to community question answering. *Expert Systems with Applications*, *137*, 432–442. https://doi.org/10.1016/j.eswa.2019.07.024

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., & Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 497–511. https://doi.org/10.18653/v1/S16-1081

Alasmari, A., & Zhou, L. (2019). How multimorbid health information consumers interact in an online community Q&A platform. *International Journal of Medical Informatics*, *131*, 103958. https://doi.org/10.1016/J.IJMEDINF.2019.103958

Alasmari, A., & Zhou, L. (2021). Share to Seek: The Effects of Disease Complexity on Health Information—Seeking Behavior. *J Med Internet Res*, *23*(3), e21642. https://doi.org/10.2196/21642

Balla, H. A. M. N., Llorens Salvador, M., & Delany, S. J. (2022). Arabic Medical Community Question Answering Using ON-LSTM and CNN. *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, 298–307. https://doi.org/10.1145/3529836.3529913

Boudjellal, N., Zhang, H., Khan, A., Ahmad, A.,

Naseem, R., & Dai, L. (2020). A Silver Standard Biomedical Corpus for Arabic Language. *Complexity*, *2020*, 8896659. https://doi.org/10.1155/2020/8896659

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, *abs/1810.0*.

Faris, H., Habib, M., Faris, M., Alomari, A., Castillo, P. A., & Alomari, M. (2022). Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: A deep learning approach. *Journal of Ambient Intelligence and Humanized Computing*, *13*(4), 1811–1827. https://doi.org/10.1007/s12652-021-02948-w

Fehri, H., Dardour, S., & Haddar, K. (2022). ARmed question answering system. *Concurrency and Computation: Practice and Experience*, *34*(21), e7054. https://doi.org/10.1002/cpe.7054

Guo, Q., Cao, S., & Yi, Z. (2022). A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, *37*(11), 8548–8564. https://doi.org/10.1002/int.22955

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. *WANLP*.

Lai, T. M., Bui, T., & Li, S. (2018). A Review on Deep Learning Techniques Applied to Answer Selection. *Proceedings of the 27th International Conference on Computational Linguistics*, 2132–2144.

Liu, Q., Agichtein, E., Dror, G., Maarek, Y., & Szpektor, I. (2012). When web search fails, searchers become askers. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12*, 801. https://doi.org/10.1145/2348283.2348390

Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, *28*(3), 345–361. https://doi.org/10.1016/j.jksuci.2014.10.007

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., & Verspoor, K. (2017). SemEval-2017 Task 3: Community Question Answering. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 27–48. https://doi.org/10.18653/v1/S17-2003

Othman, N., Faiz, R., & Smaïli, K. (2020). Improving the Community Question Retrieval Performance Using Attention-Based Siamese LSTM. In E. Métais, F. Meziane, H. Horacek, & P. Cimiano (Eds.), *Natural Language Processing and Information Systems* (pp. 252–263). Springer International Publishing.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv Preprint arXiv:1908.10084*.

Roy, P. K., Saumya, S., Singh, J. P., Banerjee, S., & Gutub, A. (2023). Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, *8*(1), 95–117. https://doi.org/10.1049/cit2.12081

Thuwaini, W. A., & Alhumoud, S. (2022). TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT With BiLSTM. *IEEE Access*, *10*, 91509–91523. https://doi.org/10.1109/ACCESS.2022.3198955

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., & Heywood, J. (2010). Sharing Health Data for Better Outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, *12*(2), e19. https://doi.org/10.2196/jmir.1549

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & others. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Zhang, W., Chen, Z., Dong, C., Wang, W., Zha, H., & Wang, J. (2021). Graph-Based Tri-Attention Network for Answer Ranking in CQA. *ArXiv*, *abs/2103.03583*. https://api.semanticscholar.org/CorpusID:232135063

8. Appendices

8.1 Dataset Description

Each question "Q" has question ID, question category, gender, age, question time, question title, question description, number of answers, relevant answer ID, date of relevant answer, relevant answer, the answer specialty of the physician, irrelevant answer ID, doctor specialty of irrelevant answer, date of irrelevant answer, irrelevant answer.

To offer a deeper understanding of the AraMed corpus, let's delve into the nature of the questions and answers. The AraMed corpus is primarily composed of questions and answers formulated in Modern Standard Arabic (MSA). This reflects the platform's focus on providing reliable and standardized health information. However, the corpus also encompasses a variety of question types. Users engage in both knowledge-seeking queries, such as "What are the symptoms of the common cold?", and those seeking specific consultations, like "I have a persistent cough. What could it be?". This diversity enriches the data by reflecting real-world information needs within Arabic-speaking communities.

Categories	Number of Questions
Gynecological diseases	24399
Sexual health	18569
General Medicine	10656
Musculoskeletal and joint diseases	7898
Skin disease	5574
Gastrointestinal diseases	5567
Urinary and venereal diseases	5063
Sexually transmitted diseases	4980
Pharmacology	4266
Cardiovascular disease	4113
General Surgery	3513
Dental disease	2410
Otolaryngology	3364
Pregnancy and birth	3140
Psychiatric illness	2753
Children's diseases	2713
Internal medicine	2684
Ophthalmology	2277

Table 3: Total number of questions for each category.