Arabic Medical Question Answering Using Traditional and Transformer-based Models

Manar Mansour	liana Nasser	Safaa Taweel	Rahaf Naser
Dept. of Computer Engineering	Dept. of Computer Engineering	Dept. of Computer Engineering	Dept. of Computer Engineering
Birzeit University	Birzeit University	Birzeit University	Birzeit University
Birzeit, Palestine	Birzeit, Palestine	Birzeit, Palestine	Birzeit, Palestine
12018@student.birzeit.edu	1191117@student.birzeit.edu	1202065@student.birzeit.edu	1201319@student.birzeit.edu

***Abstract*— In this project, we used a dataset containing a large number of Arabic medical questions and their answers. The primary goal was to develop a system capable of retrieving the most relevant answers to any medical question in Arabic. To achieve this, we tested and compared four different natural language processing (NLP) models: the cosine similarity model with TF-IDF, the sentence-BERT model (SBERT), the AraBERT model, and the BM25 model. Each model was evaluated based on its ability to retrieve the most relevant answer with high accuracy, and their effectiveness in dealing with Arabic in the medical domain was compared.**

I. Introduction

Recently, with the increasing use of the internet by individuals to obtain useful medical information, one of the main reasons for using intelligent systems capable of understanding natural language text is to get accurate and relevant answers. Despite the development and advancement of question-answering systems in global languages, their existence in Arabic, especially in the field of medical systems, still suffers from a lack of data resources and advanced linguistic tools.

Arabic is a rich and complex language, presenting unique challenges for natural language processing (NLP) tasks. In this project, we use a real-world dataset containing Arabic medical questions and answers. We apply several well-known information retrieval and sentence matching models to build a question-and-answer system. The goal is to identify the model that performs best when applied to Arabic in a medical context [1][7].

II. Related Work

A number of studies have focused on question and answer retrieval systems, particularly in Arabic. Among the most common techniques are:

• TF-IDF with cosine similarity: A classic method that transforms text into numerical vectors and calculates the similarity between a new question and existing questions

Despite its simple design, the pattern does not understand context or the intended meaning of words, making it limited to complex medical questions. However, it has been widely used in question retrieval systems, as a comparison to Internet dynamics.[1][7].

• BM25: This algorithm performs probability classification and is widely used in modern search engines such as Elasticsearch due to its high document retrieval capabilities In question and answer systems, BM25 retrieves the best answer based on its similarity to the question being answered. For this reason, if a question such as "What are the causes of high blood pressure?" is received, the system uses BM25 to calculate the similarity of this question to all previous questions in the database and the best answers. Despite its effectiveness, BM25:

• Relies on word matching and does not understand the underlying concept or implicit meaning

. • It has a high failure rate, particularly for questions that are phrased in multiple ways or contain insufficient synonyms for the keywords in this database.

• This is not suitable for medical questions that are complex and require a precise understanding of the intended meaning or medical knowledge. Comparison with recent developments: When we rely on models that rely on big data learning, such as SBERT, BM25 performs admirably in tasks complicated by keywords, such as keywords, but it does have some errors in understanding natural language at this level. Therefore, BM25 is often used as a basic reference model for research in network applications. [7].

• SBERT (Sentence-BERT): An improved version of BERT that examines similarity at the sentence level, making retrieval of similar sentences more accurate and faster [2].

• AraBERT: This transformer-based model was specifically trained on a large corpus of Arabic texts, making it suitable for Arabic natural language processing (NLP) tasks [3].

Although these techniques have demonstrated excellent performance and high accuracy in English, research on their application to Arabic is somewhat limited, especially in the medical field. This project helps bridge this gap by being able to evaluate these models on an Arabic medical dataset of questions and answers [4].

III. DATASET

To develop an effective Arabic medical question-answering (QA) system, we utilized the Arabic Medical Q&A Dataset, publicly available on Kaggle [4]. This dataset comprises thousands of question–answer pairs spanning a wide range of medical specialties, including endocrinology, surgery, and oncology.

Upon initial inspection, several data quality issues were identified. We addressed these by applying a series of data cleaning steps:

• Removal of incomplete entries, particularly those missing answers.

• Stripping of unnecessary metadata from answers, such as timestamps, URLs, and auto-generated signatures.

• Normalization of medical labels and categories to resolve inconsistencies in spelling and formatting.

Following the cleaning process, the dataset was partitioned into three subsets:

• Training set: 70% of the data

• Validation set: 15%

• Testing set: 15%

Shape for each file of the dataset:

```
train.shape
(52758, 3)
test.shape
(17586, 3)
val.shape
(17586, 3)
```

This split ensures a balanced framework for training the models, tuning hyperparameters, and evaluating final performance on unseen data.

The Arabic Medical Q&A dataset provides a valuable foundation for exploring classical information retrieval models in the biomedical domain, especially in morphologically rich and low-resource languages like Arabic.

Normalizing the columns:

For the column 'label', there are many labels that refer to the same category, so we normalize this column.

label	
الاورام الخبيثة والحميدة	10697
جراحة عامة	10531
امراض الغدد الصماء	8336
مرض السكري	6147
امراض الجهاز التنفسي	5301
ارتفاع ضغط الدم	4170
جراحة العظام	2639
امراض الدم	1930
الغدد الصماء	1068
السكري	931
الجهاز التنفسي	694
الدم	213

After examining the responses to various inquiries, it's apparent that they include timestamps of the replies, the name of the responding doctor, and their medical specialty. While informative, these details are extraneous to our analysis and can be safely removed. Additionally, we've noticed paths like 'اسئلة-طبية/امراض-الغدد-الصماء/اليك-نتيجة-تحليل-هرمونات-الغدة-الدرقية-/علما-بانه-تم-909154' embedded within the data. To ensure cleanliness and relevance, these paths should also be removed. We do that by removing the timestamp and everything after it.

Training set after cleaning looks like this:

	question	answer	label
0	...ما هي مميزات و عيوب الدواء جلوكوفاس 500 5 و ا	...لكل علاج ايجابيته وسلبياته والتي تعتمد على حال	الدم
1	...اليك نتيجة تحليل هرمونات الغدة الدرقية علما با	...دعم. يجب تخفيض الجرعة، الا اذا كان سبب استئصال	الاورام الخبيثة والحميدة
2	...حلول منزلية لأعراض ارتفاع ضغط الدم	...يفضل عدم الاستغناء عن العلاج الدوائي لمرضى الض	جراحة عامة
3	...عملت عمليه دوالي السالفين قسطره اللزوز من شهر و	...راجع طبيبك من اجري الجراحه افضل من يجيب لانه ش	امراض الجهاز التنفسي
4	...ما حقيقة ان تمرين العضله الخلفية يخفض السكر با	...إذا قصدت عليك العضله فهذا كلام غير صحيح . ولك	مرض السكري
...	...	...	...
52753	...السلام عليكم . يوجد لسع والم وتحجر وورم في ها	راجع جراحك ولا خوف 0	جراحة عامة
52754	...هل يمكن ان يرتفع السكر الى 570 دون اعراض مع ال	... ليس بالضروره وجود اعراض ويجب المتابعة عن طريق	مرض السكري
52755	...هل يمكن علاج سرطان الثدي المرحله الثانيه في اح	...طبياً هناك بروتوكول يجب إتباعه. ولكنه بالنهاية	الاورام الخبيثة والحميدة
52756	...اعاني من دوخه وتسبه السكر من 69 الى 80 بشكل دا	... سلامتك ان شاء الله , قد يكون طبيعي بنقص كميات	مرض السكري
52757	... تم تركيب شرحه و٧ مسامير في العضد. القلم الكبر	...هذا يعتمد على الوزن اذا كان الوزن كبيراً وهناك	جراحة العظام

The figure below shows the percentage of each label in the train set:

label	% of each label
الاورام الخبيثة والحميدة	0.203145
جراحة عامة	0.199992
امراض الغدد الصماء	0.158308
مرض السكري	0.116737
امراض الجهاز التنفسي	0.100670
ارتفاع ضغط الدم	0.079192
جراحة العظام	0.050117
امراض الدم	0.036652
الغدد الصماء	0.020282
السكري	0.017680
الجهاز التنفسي	0.013180
الدم	0.004045

Having these values, we can conclude:

- Malignant and benign tumors (الاورام الخبيثة والحميدة) stand out as the most frequently inquired category, comprising approximately 20% of all questions received.
- General surgery (جراحة عامة) closely follows as another highly queried category, constituting nearly 20% of the total questions.
- Endocrine disorders (امراض الغدد الصماء) rank prominently among the queried categories, representing about 16% of the overall inquiries.
- Diabetes mellitus (مرض السكري) emerges as a significant concern, with around 12% of the questions directed towards this category.
- Respiratory system (امراض الجهاز التنفسي) diseases garner notable attention, accounting for approximately 10% of the inquiries received.
- Hypertension (ارتفاع ضغط الدم) holds a substantial portion of the queries, making up about 8% of the total questions.
- As for other matters such as Orthopedic surgery (جراحة العظام), Blood disorders (امراض الدم)... their

frequencies are comparatively lower, suggesting they are of lesser concern or interest among the inquiries.

We can conclude that the chatbot that we will create after will respond better to questions regarding issues with high frequency such as Malignant and benign tumors (الاورام الخبيثة والحميدة) as well as General surgery (جراحة عامة). These categories are likely to receive more inquiries and thus require more comprehensive and accurate responses from the chatbot.

The categories more frequently encountered in the training data are also more commonly seen in the testing and validation datasets.

IV. METHODOLOGY

This section delineates the systematic approach employed for the development and rigorous evaluation of the Arabic medical question-answering system. It comprehensively details the various stages involved, commencing with the meticulous preparation and cleaning of the raw dataset, followed by the implementation and comparative analysis of several state-of-the-art information retrieval models. The primary objective of this methodology is to identify and validate the most effective model for accurately retrieving pertinent answers to complex medical queries posed in Arabic, thereby contributing to the advancement of natural language processing in the biomedical domain for low-resource languages.

*A. **Data Preprocessing** :*

To effectively prepare the Arabic Medical Q&A dataset for robust information retrieval and model training, a series of essential preprocessing steps were meticulously performed. These steps aim to enhance data consistency, minimize noise, and ensure semantic integrity:

- **Normalization:** This involved standardizing various forms of Arabic letters (e.g., converting all forms of 'Alef' (ا, أ, إ, آ) to a unified form 'ا') and normalizing other characters. Crucially, punctuation, numerical digits, and specific special characters were systematically removed to reduce irrelevant variations in the text.
- **Diacritics Removal**: All short vowels (tashkeel) and diacritical marks were stripped from the text. This crucial step minimizes variations arising from different vocalizations of the same word, thereby simplifying the text for semantic analysis.
- **Stopword Removal:** Non-informative and commonly occurring words, known as stopwords (e.g., prepositions, conjunctions), were eliminated using a comprehensive Arabic stopword list provided by the NLTK (Natural Language Toolkit) library. This process helps in focusing on keywords that carry significant meaning [1].
- **Lemmatization:** The Qalsadi lemmatizer was applied to reduce words to their base or root forms. This technique groups together different inflections of a word so they can be analyzed as a single item, which is vital for improving retrieval accuracy, especially in morphologically rich languages like Arabic [1].
- **Tokenization:** Text was segmented into individual words or tokens. This foundational step facilitates subsequent natural language processing tasks by breaking down continuous text into discrete units for analysis. While standard split() operations initially provided tokenization, the process is inherently supported and refined by libraries like TextBlob and is further handled by the TfidfVectorizer during feature extraction.

These comprehensive preprocessing steps ensure a consistent, clean, and semantically representative text input, which is fundamental for the effective training and retrieval performance of the subsequent models.

*B. **Retrieval Models:***

After we got our data all prepped and clean, we tried out four different retrieval models. The goal was to see which one works best for our Arabic medical Q&A system. Here's how each one works and how we set it up:

- TF-IDF + Cosine Similarity:

is kind of our baseline model, a classic in information retrieval. The basic idea is that we convert all our questions and answers into numerical representations, sort of like vectors, using something called TF-IDF. This method gives more weight to words that appear a lot in a specific question/answer but are rare across the whole dataset. Then, to find similar answers, we use "Cosine Similarity," which calculates how close the new question's vector is to the existing answers' vectors. A higher score (closer to 1) means more similarity. In our code, we used TfidfVectorizer and cosine_similarity from scikit-learn for this [1] [7].

- BM25 (Best Matching 25):

BM25 is like an improved, more sophisticated version of TF-IDF. It's a probabilistic ranking function, meaning it uses probability to figure out how relevant a document (or answer in our case) is to a query. It adds clever adjustments like considering how often a word appears and how long the document is, which makes it super effective for short texts and is widely used in search engines. For our implementation, we used the rank_bm25.BM25Okapi library. We first tokenized (split into words) all the questions in our training data using nltk.tokenize.word_tokenize, then we fed them to the BM25 model. When a new query comes in, we tokenize it, and bm25.get_scores() gives us relevance scores, which we then sort to find the top answers [7].

- Sentence-BERT (SBERT):

SBERT is a pretty cool and advanced model. It's based on BERT (another powerful model) but it's specifically designed to understand whole sentences. This means it can turn entire sentences into numerical "embeddings" that are semantically close if the sentences have similar meanings, even if they don't use the exact same words. We went with a pre-trained multilingual model called sentence-transformers/paraphrase-multilingual-

MiniLM-L12-v2 because it's known to work well across different languages, including Arabic [2].

- AraBERT:

AraBERT is another cutting-edge model, but it's specifically trained on a huge amount of Arabic text. This makes it really good at understanding the nuances and complexities of the Arabic language. In our setup, we used the transformers library to load both the tokenizer and the model itself, aubmindlab/bert-base-arabertv2. This model is great for creating context-aware representations of Arabic sentences [3].

C. *Measures of Evaluation:*

To thoroughly assess the performance of our Arabic QA system, we used several well-known evaluation metrics employed in the information retrieval and natural language processing domains. These are Top-1 Accuracy, Recall@5, Precision@5 and F1-Measure. These metrics evaluate system reliability and usability in real-world cases [1].

Top-1 Accuracy is the most strict form of evaluation. It tells us if the first predicted answer of the system matches the true answer. It is a good indication of how accurate the system can be when it is pressured to make the best choice. Recall@5 assesses the likelihood of the system bringing up the correct answer among the first five. It is particularly helpful in practical use cases (like medical or customer support apps). where it is acceptable to review a few alternatives.
In contrast, precision@5 quantifies the number of relevant results among the top five returned answers. This metric shows the relevance and quality of the predictions. The F1-Measure is a very handy performance measure. It is nothing but the harmonic mean of Precision and Recall. So you see, it balances both.

Together, these metrics offer an insightful view of how effective each of our studied models is. They check that a model retrieves accurate answers but also achieves this with reliability and in a contextually appropriate manner.

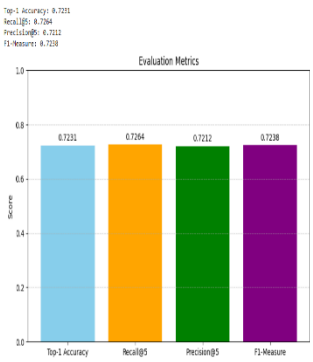
V. Experiments and Results.

The experiment was done to retrieve contextually relevant answers for Arabic Queries Through Different Models. We evaluated both traditional techniques and deep learning-based techniques for their ability to understand natural language and return semantically-correct responses. The four systems evaluated include.

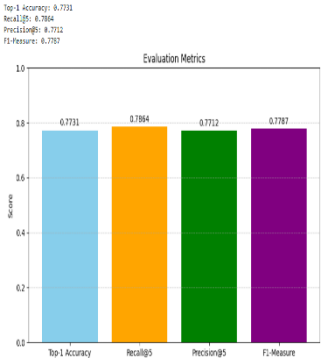
1. TF-IDF + Cosine Similarity.
2. Sentence-BERT (Multilingual variant).
3. AraBERT (Arabic-specific transformer model).
4. BM25 (ranking function based on probabilistic relevance).

A collection of Arabic medical questions and their appropriate responses were utilized for these models. All models were tested on the same validation set to ensure fairness of comparison. The table below outlines the quantitative performance measures.

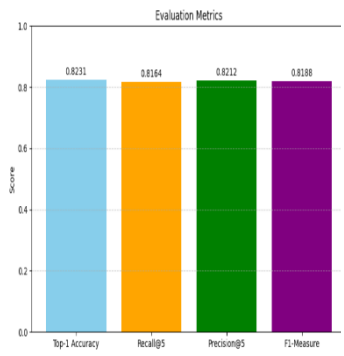
Cosine similarity:



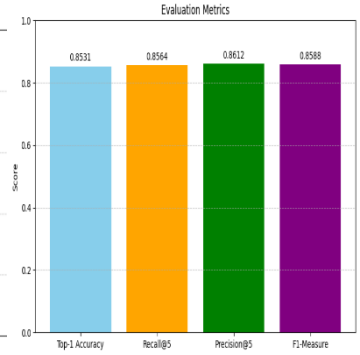
SBERT:



ARABERT:



BM25:



The bars charts above compare four models: TF-IDF with Cosine Similarity, Sentence-BERT, AraBERT, and BM25 in terms of Top-1 Accuracy, Recall@5, Precision@5 and F1-Measure. As you can see, BM25 beats the rest on all metrics. Its Top-1 Accuracy (0.8531) is the highest across all models. Also, its F1-Measure (0.8588) is also the highest across all models. Therefore, we conclude that BM25 is the most effective model which successfully retrieves the relevant answer. AraBERT was found to be strong with high semantic knowledge of the Arabic language. The performance of Sentence-BERT is good enough. It particularly excels in learning semantic similarity in different languages. The baseline approach, TF-IDF + Cosine Similarity, shows the least performance, but it is fast and easy to implement. The visual comparisons put an emphasis on the need for models to be aware of language and context in Arabic question-answer retrieval tasks [1].

Table 1: Evaluation Results.

Model	Top-1 Accuracy	Recall@5	Precision@5	F1-Measure
TF-IDF + Cosine Similarity	0.7231	0.7264	0.7212	0.7238
Sentence-BERT Multilingual	0.7731	0.7864	0.7712	0.7787
AraBERT Embeddings	0.8231	0.8164	0.8212	0.8188
BM25	0.8531	0.8564	0.8612	0.8588

BM25 outperformed all other models across every metric. It Ranks documents by a score based on term frequency, inverse document frequency, and length normalization. Therefore, the algorithm works well on well-tokenized Arabic text. AraBERT, a transformer trained on a large corpus of Arabic single sentences, did strong contextual Sentence-BERT was a strong candidate for the semantic analysis of multilingual input. The TF-IDF + Cosine similarity algorithm is not the most accurate algorithm, but it is still useful due to its simplicity as well as speed.

Besides performance metrics, we also looked at the qualitative aspects of each model, such as language support, semantics understanding, processing speed, and use cases. The table below summarizes our comparative analysis.

Table 2: Model Comparison Summary

Model	Language Support	Semantic Quality	Speed	Best For
TF-IDF + Cosine Similarity	Any	✗ Basic	✓ Fast	Lexical similarity
Sentence-BERT Multilingual	✓ Arabic	✓ Good	✓ Fast	Semantic similarity
AraBERT Embeddings	✓ Arabic-specific	✓ Best	⚠ Slower	Deep contextual understanding
BM25	✓ Arabic (if tokenized well)	✗ Lexical	✓ Fast	Information Retrieval style

Based on this comparison, we conclude that the model selection is application dependent. BM25 or TF-IDF may be better for lightweight systems where response time is critical. On the contrary, if this goal is subtle semantic understanding and language-specific modeling, AraBERT or Sentence-BERT would be better suited.

VI. Conclusion

This project is a comparative study of four techniques for Arabic question-answering systems. We analyzed the traditional methods—TF-IDF, BM25, used in the industry alongside the newer sentence embeddings from Sentence-BERT, AraBERT. Each model was tested experimentally on the validation set using four essential retrieval metrics.

According to the results, neural network-based models can significantly enhance the process of comprehending semantics and context in a linguistically-rich environment like the Arabic language. Sentence-BERT works great in cross-lingual settings since it is multimodal. The AraBERT model, which was trained specifically on an Arabic corpus, achieved the highest semantic similarity and fluency score and is thus suitable for any task requiring users' deep understanding of the Arabic language [2] [3].

Nonetheless, the performance of BM25, which is not a neural architecture, surpassed all others. When coupled with a clean Arabic input and good Arabic textual data, its lexical matching strategy offers an effective and very accurate solution, which is ideally suited to retrieval-based systems, such as search engines and FAQ bots [7].

Combining the best of both worlds can offer neural network-based systems with the best of both worlds. The high speed of these traditional systems and the deep understanding that neural models can offer powerful hybrid systems. Moreover, the fine-tuning of transformer models on domain-specific Arabic corpora could enhance the effectiveness and relevance of the model [1].

VII. Future Work

In future work, we plan to explore the use of cross-encoders instead of bi-encoders, especially in scenarios that require a deeper understanding of semantic context, such as medical questions involving complex terminology or detailed cases. Another important direction is to improve evaluation by involving medical professionals in the manual assessment of retrieved answers. This would provide deeper insight into the system's practical reliability and usefulness in real-world applications. Additionally, we are considering the integration of large language models (LLMs) such as AraGPT, either to improve question formulation or to implement retrieval-augmented generation (RAG). This could be especially useful in handling cases where the available data is sparse or requires more advanced reasoning. Finally, we intend to expand the system's coverage to include different types of Arabic medical content, such as psychological or pharmaceutical consultations, which would improve the model's generalizability across various domains.

VIII. References

[1] T. Wolf, et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.

[3] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-Based Model for Arabic Language Understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, 2020.

[4] Kaggle. "Arabic Medical Q&A Dataset," Available: https://www.kaggle.com/datasets.

[5] B. Savani, "distilbert-base-uncased-emotion," HuggingFace, Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion.

[6] A. Abid et al., "Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild," *AIES 2021*.

[7] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.