Diagnosy

Dr. Dina El-sayad
Scientific Computing Department
Faculty of Computers and information Sciences
Ain Shams University
Cairo, Egypt
dina.elsayad@cis.asu.edu.eg

Heba Gamal
Scientific Computing Department
Faculty of Computers and information Sciences
Ain Shams University
Cairo, Egypt
heba_gamal@cis.asu.edu.eg

Malak Ismail
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700850@cis.asu.edu.eg

Manar Alaa
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700853@cis.asu.edu.eg

Fady Makram
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700562@cis.asu.edu.eg

Manar Ibrahim
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700852@cis.asu.edu.eg

Mennattallh Ibrahim
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700856@cis.asu.edu.eg

Manar Mohamed
Sicentific Computing Department
Faculty Of Computers and information Sciences
Ain Shams University
Cairo, Egypt
20201700854@cis.asu.edu.eg

Abstract: The COVID-19 pandemic has strained daily activities and communication between patients and primary care providers, making it challenging to provide adequate primary care services. As a result, healthcare services, particularly remote and automated consultations, have gained increased attention, leading to the rise of medical bots. These bots offer benefits such as 24/7 access to medical counseling, reduced appointment wait times, and cost savings. However, implementing medical bots in Arabic faces challenges due to the language's morphological composition, dialect diversity, and the need for a large corpus. This document introduces the largest Arabic Healthcare Q&A dataset, MAQA, which contains over 430,000 questions across 20 medical specializations. Preprocessing techniques were applied to the dataset, and the data was split into 80% for training and 20% for testing. Three deep learning models, LSTM, Bi-LSTM, and Transformer were used for experimentation and benchmarking. The Transformer model outperformed traditional deep learning models, achieving an average cosine similarity of 83%.

Keywords: Deep learning · LSTM· Arabic corpus. medical bot · Transformers**.** NLP.

I. Introduction

In some regions, access to healthcare facilities is severely limited or entirely absent. This scarcity leaves communities vulnerable, as individuals lack essential medical services and resources to address their health needs. Without nearby healthcare facilities, people may delay seeking treatment for illnesses or injuries, leading to worsened health outcomes or even preventable deaths. The absence of healthcare infrastructure also exacerbates disparities in healthcare access, disproportionately affecting marginalized populations who may already face numerous barriers to receiving adequate medical care.
Misdiagnosis can have serious consequences, especially when it occurs due to a lack of experience or information on the part of the attending physician. In some cases, patients' conditions may be overlooked or incorrectly identified, leading to delays in appropriate treatment or even unnecessary interventions. When doctors lack sufficient expertise or access to comprehensive medical resources, they may struggle to accurately diagnose complex or rare conditions. This not only impacts individual patients but also erodes trust in the healthcare system.

The emergence of the COVID-19 pandemic in 2019 profoundly disrupted human interaction, leading to a significant decrease in face-to-face interactions among individuals. This shift had far-reaching consequences, particularly in healthcare settings, where hospitals became inundated with patients infected with the virus. The fear of contracting COVID-19 deterred some individuals from seeking medical attention, even for unrelated illnesses, resulting in the exacerbation of their health conditions. Tragically, this reluctance to seek timely medical care contributed to the deaths of some individuals who could have otherwise received life-saving treatment.

II. Related Works

There have been numerous works and papers related to the application of diagnosing diseases using machine learning and deep learning, Table [1] shows the results of studies that using ML techniques, Table [2] shows the results of studies that using DL techniques.

Author	Year	Approach	Accuracy
Asad Ur Rahman et al [6]	2022	SVM	93.33
Sgar Badlani et al [7]	2021	Random Forest	98.43

Table 1 : Literature studies on traditional papers

Author	Year	Approach	Accuracy	Similarity	BLeU
Mohammed Abdelhay et al [1]	2023	LSTM Bi -LSTM Transformer	-	56 72 80	31 39 58
Abdennour Boulesnane et al [2]	2022	LSTM, GRU BI-LSTM	78 71	-	-
Qiming Bao et al [3]	2020	BERT MaLSTM HBAM	78.2 78.4 81.2	-	-
Prateek Mishra et al [4]	2022	Rasa Architecture	78.7	-	-
Mohamed Boussakssou et al [5]	2022	LSTM GRU	89 85	-	-

Table 2 : Literature studies on deep learning papers

III. MATERIALS

A. *Database used*

MAQA was acquired from a variety of websites such as (altibbi.com, tbeeb.net, cura.healthcare, etc.), it is the largest Arabic Healthcare Q&A dataset. The dataset contains around 430k questions divided into 20 medical specialties. The distribution of questions per category is summarized in Table [3]. All questions are unique. The data is kept in raw format, cleaned but not stemmed, or any other preprocessing has applied after scraping. The questions and answers contain some English symbols and digits, and almost no Arabic diacritics or punctuation. Table [4] shows an example of a question from the "امراض نسائية","امراض الجهاز الهضمي " categories.

The data contains the following columns q_body contains the question content, a_body contains the answer content, q_body_count contains the question content word count, a_body_count Contains the answer content word count, category Contains category name and category_id Contains category number from table categories.

Label	Count
امراض نسائية	103683
امراض المسالك البولية والتناسلية	33847
امراض العضلات والعظام و المفاصل	33050
الامراض الجلدية	29262
الطب العام	26870
امراض باطنية	23722
امراض الجهاز الهضمي	22373
الامراض الجنسية	21773
طب الاسنان	20207
امراض الاطفال	18636
امراض نفسية وعصبية	18295
امراض القلب و الشرايين	15368
جراحة عامة	15185
امراض العيون	14439
انف اذن وحنجرة	13933
الاورام الخبيثة والحميدة	11210
امراض الغدد الصماء	5186
امراض الجهاز التنفسي	4567
جراحة تجميل	1596
امراض الدم	1341

Table 3: Categories and number of its record

Question	Answer	category
اذا كانت مده الدورة عندي 32 يوم فماهي ايام الإباضة عندي لان بدي احمل	18 17 16 15 14	امراض نسائية
السلام عليكم جاي يصير عده تهيج في تحدث عده خربطه في المعدة والقالون العصبي تستخدم علاج مرفق في الفايل ومستقرة على العلاج فقط تحدث مشكله في الاكل ماذا تكل وماذا...	This is not colon treatment it is stomach treatment	امراض الجهاز الهضمي

Table 4: examples of data

B. *Second Datasets*

Disease symptom prediction is an English dataset used to develop a disease prediction or healthcare system. The information includes diseases, symptoms, recommendations to take, and weights. The collection is organized into 41 diseases and 17 symptoms. Initial file for disease prediction classifier, Retrieve a second file with disease descriptions and Third file is for illness prevention. Table [5] show example from this data

IV. METHOD

A. *System Architecture*

Data cleaning and exploration: The processing of Arabic text presents significant challenges, including ambiguity, diglossia, and difficulty in reading and understanding the Arabic script.

Symptom	Disease	Description	precaution
الم المفاصل وجع البطن فقدان الشهية القئ	إلتهاب الكبد أ	التهاب الكبد (أ) هو عدوى الكبد شديدة العدوى الناجمة عن فيروس التهاب الكبد (أ). يعد الفيروس واحدًا من عدة أنواع من فيروسات التهاب الكبد التي تسبب الالتهاب وتؤثر على قدرة الكبد على أداء وظائفه.	استشارة أقرب مستشفى غسل اليدين تجنب الأطعمة الغنية بالتوابل الدهنية

Table 5: Example of second data

Researchers must also address the normalization of inconsistencies in the use of certain letters, dialect words, or diacritical marks. To tackle these challenges, researchers and developers must employ Natural Language Processing (NLP) tools such as tokenization, stemming, and morphology analysis.

Data cleansing involves eliminating duplicate and missing data, manually excluding ambiguous sentences, relabeling mislabeled data, and applying data normalization techniques to standardize the text. It is essential to understand the contents of the dataset before selecting and working with it, as it can help detect trends, patterns, outliers, and relationships between variables.

To extract interpretable features from the data, exploratory data analysis (EDA) was used. This involved counting phrases linked to symptoms in each class, visualizing the most used terms, and calculating sentence lengths to estimate the maximum length of the input sequence for each model. This approach helps researchers detect trends, patterns, and relationships between variables in the Arabic language.

Data Augmentation:
Data Augmentation is a strategy to prevent overfitting through regularization, enabled through an intuitive interface. It helps researchers understand the type of priors or additional data needed to improve a system. For example, a question answering dataset might fail with symmetric consistency on comparison questions. The list of augmentations describes the mechanisms available to inject these priors into datasets.

Symbolic augmentation:
Symbolic augmentations are a type of data augmentation that uses auxiliary neural networks or other statistical models to generate data, unlike neural augmentations which use symbolic rules. Symbolic augmentations offer interpretability for human designers and work better with short transformations like replacing words or phrases. However, they are limited in applying global transformations, such as augmenting entire sentences or paragraphs, for information-heavy applications requiring longer inputs.

Rule-based augmentation:
Rule-based Augmentations use if-else programs and symbolic templates to create augmented examples. Easy Data Augmentation offers four augmentations, making it easy to use off-the-shelf. Many of these augmentations are still in the research phase, waiting for large-scale testing and adoption. Examples include random swapping, random deletion, random insertion, and random synonym replacement.

Back-translation augmentation:
Back-translation involves translating text from one language to another and then back to the original language. This process has gained significant interest in machine translation, leading to the creation of large, labeled datasets of parallel sentences. Other text datasets, such as translations between programming languages or writing styles, can also be used for this purpose.

Style augmentation:
The study presents a new augmentation strategy using Deep Networks to enhance data for training other Deep Nets. It explores Neural Style Transfer, a technique that uses artistic style transfers to enhance data. This strategy aims to prevent overfitting to high-frequency features or blurring out language form, focusing on meaning. In the text data domain, it could be used to transfer writing styles between authors for applications like abstractive summarization or context for extractive question answering.

Generative data augmentation:
a promising concept in Deep Learning, generates photorealistic facial images or indistinguishable text passages. It's useful for Transfer Learning but its potential applications are more significant in artistic and representation learning and Data Augmentation.
Tokenization: This model uses word tokenization for chatbot execution, breaking down user queries into a list of words. This process is part of data processing or clean-up to eliminate irregularities or noise.

Lemmatization:
is a simplified version of stemming that normalizes tokenized data, enhancing learning by the model. It is performed before training and execution to ensure similar form of matched words or tokens, enhancing the learning process.

Bag of Words (Bow):
The text data is converted into vector values for training a model. The model matches words related to Diabetes after tokenization and lemmatization. A hard-coded function creates a bag of words for each class, which are then converted to NumPy vectors before training the model.

Pretrained models for preprocessing:
Pretrained models, like BERT, GPT, and RoBERTa, are used to preprocess text data, enhancing accuracy and efficiency. These models, trained on diverse datasets, perform sophisticated tasks like tokenization and context-aware word embeddings with reduced computational effort. This ensures consistency, reliability, and improved performance in natural language processing tasks.

Classification System:
Medical specialty identification is crucial in healthcare, enabling early diagnosis of various diseases. Machine learning techniques, including pattern recognition and fast training algorithms, can help doctors understand patterns and distinguish distinct diseases. This system categorizes unstructured text data into medical specialties, automating administrative processes and identifying potential areas of specialty, allowing medical specialists to manage patients' treatment pathways.

Assist System:
This study explores the use of AI in healthcare, specifically chatbots, to improve scientific practice and develop cost-effective resources. AI advancements have enhanced natural language processing capabilities, allowing chatbots to automate conversations and respond promptly to user inquiries. Innovations like user simulation plans are being proposed to meet new modeling requirements and re-engineer chatbots, particularly in mobile and mental health care.

B. *System implementation*

Chatbot

The chatbot development process is divided into several stages, focusing on preprocessing, classification, and generation of the text data. Each stage involves specific techniques to ensure the quality and functionality of the chatbot.

Preprocessing

a) Data Balancing:
Translation-based Augmentation, Word Swap, Scraping Technique

b) Classification Preprocessing:
Drop unnecessary columns (e.g., a_body, a_bodycount), Determine the maximum length of the questions, Apply label encoding to categorize the data, Tokenize the question body, Pad the question body to ensure uniform length.

c) Generation Preprocessing:
Calculate the average length of the questions and use it as the total length, **Duplicate Removal and Missing Value Handling:** Remove duplicate entries and handle missing values appropriately, **Advertisement Removal:** Eliminate advertisements or promotional content to focus on the main text, **Diacritics Removal:** Remove diacritical marks to simplify text representation, **Repeating Character Removal:** Remove unnecessary repeating characters for text normalization, **Standardizing Arabic Letters:** Ensure uniformity in the representation of Arabic letters, **Replace Numbers with Words:** Convert numerical digits to text, **English Text Detection and Handling:** Detect and handle English text within the Arabic data to maintain coherence, **Data Filtration:** Set maximum length limits for questions and answers, retaining only those within the limit, **Answer Trimming:** Limit the number of words in answers by trimming and updating accordingly, **Word Counting:** Create a word count dictionary for both questions and answers, **Vocabulary Creation:** Define a threshold for word frequency to include in the vocabulary, adding special tokens as needed**, Encoding and Padding:** Convert cleaned questions to their corresponding indices from the vocabulary, pad sequences, and truncate those exceeding the maximum length.

Models

First classification model:

A. *Machine learning models*

The Random Forest model:
is an ensemble learning method used for classification and regression. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Multi (MultinomialNB):
is a probabilistic learning algorithm based on applying Bayes' theorem with the strong independence assumption between the features. It is particularly well-suited for classification with discrete features, such as word counts or term frequencies in text classification.

SVM:
is a powerful supervised learning algorithm used for both classification and regression tasks. SVM works by finding the hyperplane that best separates the data into different classes, with the goal of maximizing the margin between the classes.

B. *Deep learning models*

LSTM model:

LSTM networks are a type of recurrent neural network (RNN) capable of learning long-term dependencies, making them ideal for sequence prediction problems, such as text classification. **LSTM Layer** Captures sequential dependencies in the data.

Transformer Model with GRU for Text Classification:

text classification model using a Transformer architecture enhanced with a GRU (Gated Recurrent Unit) layer. Transformers are well-suited for capturing global dependencies in sequences, while GRU helps in capturing temporal dependencies. The model also utilizes early stopping to prevent overfitting.

Fined tuning for aubmindlab/bert-base-arabertv2 model:

utilizes a pre-trained BERT model for sequence classification. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art model for natural language understanding tasks. Here, we use the AutoModelForSequenceClassification and AutoTokenizer from the transformers library by Hugging Face to fine-tune a pre-trained BERT model on a specific text classification task.

Second generation model:

Seq2Seq models with Long-Range dependencies:

LSTM (Long Short-Term Memory): networks are a type of recurrent neural network (RNN) designed to process sequential data and capture long-term dependencies, overcoming the limitations of traditional RNNs. The key components of LSTMs are memory cells that store information over long periods, updated and modified as the model processes input sequences. LSTMs use gating mechanisms forget gate, input gate, and output gate to control the flow of information through these memory cells. Training an LSTM model involves gradient descent and backpropagation to adjust weights and biases, minimizing prediction errors. Once trained, LSTMs can be used for text generation by iteratively predicting the next word or character based on learned context, producing coherent and contextually relevant sequences. LSTMs excel in capturing long-term dependencies, handling variable-length input sequences, and generating text of arbitrary lengths, making them highly effective for tasks like text generation.

Bi-LSTM (A Bidirectional LSTM): extends the standard LSTM by processing input sequences in both forward and backward directions. It consists of two LSTM layers: one processes the sequence from left to right, and the other from right to left. This bidirectional processing allows the model to capture both past and future dependencies for each token. The forward LSTM layer processes the input in a standard sequential manner, while the backward LSTM layer processes it in reverse order. As both layers process the sequence, they maintain hidden states that encode the learned information. After processing, the hidden states from both directions are concatenated, providing a comprehensive representation of the input sequence. The Bi-LSTM model is trained using backpropagation and gradient descent, adjusting weights and biases to minimize prediction errors. For text generation, the model takes an initial input and iteratively predicts the next word or character based on the combined context from both directions. The advantages of Bi-LSTM include capturing bidirectional dependencies and enhanced context understanding, leading to richer text generation. Bi-LSTM models are effective in various natural language processing tasks, such as text generation, sentiment analysis, and named entity recognition, offering improved context understanding compared to unidirectional LSTMs. However, their use should consider computational complexity and potential overfitting.

In recent advancements, the performance of LSTM and Bi-LSTM models has been significantly enhanced by integrating them with attention mechanisms this integration allowed us to approach better result

Transformers: introduced in "Attention Is All You Need [8]" by Vaswani et al., have revolutionized NLP tasks like text generation by effectively capturing long-range dependencies and processing sequential data. During training, transformers adjust weights and biases through techniques like stochastic gradient descent, using a multi-layer structure comprising self-attention mechanisms and feed-forward neural networks. The self-attention layer transforms input embeddings into query, key, and value vectors, computes attention scores via dot products, and generates attention weights through a SoftMax function. This yields a weighted sum representing attended information, which is combined with the original embeddings via residual connections. The feed-forward layer then applies linear transformations, activation functions, and another residual connection, followed by layer normalization. This process, repeated across layers, allows transformers to capture both local and global dependencies. Transformers' advantages include capturing long-range dependencies, parallel processing for faster training and inference, scalability for large datasets, reduced context fragmentation, and interpretability through attention mechanisms. These features make transformers computationally efficient and adaptable across various tasks and domains.

disease-symptom-prediction

Preprocessing

Translate data from English to Arabic by using google translate, Shuffle data, replace _ with space, fill nan by 0, Replace symptom with its weight

Models

We use machine learning model such as (**random forest**), Second **Decision tree** classifier: is a simple yet powerful algorithm used for classification tasks. It works by recursively splitting the data into subsets based on the most significant feature at each node. This implementation demonstrates how to use a Decision Tree for text classification by first converting text data into numerical features and then training a Decision Tree on those features

voting models contain SVM, Random Forest and decision tree

V. Model Evaluation

The deep learning model uses trained word vectors from the pre-trained CBOW model called Aravec, trained on 132,750,000 Arabic documents with 3,300,000,000 words. A word embedding matrix is generated from this model, and the word embedding sentences of the corpus are fed to the network as input features. The model is tested on a held-out test set and evaluated using metrics such as cosine similarity and BLEU score as following:

A. Cosin Similarity

To evaluate our generated answer against the actual answer, we start by getting the embedding vector for each word in the sentence, then get the average for all words' vectors as in equation A1, where A is the sentence vector, Vi is the word vector and N the words count in the sentence. Then, we calculate the vectors product in equation A2, where A and B are two nonzero vectors can be derived by using the Euclidean dot product formula. After that we calculate the cosine similarity between both average vectors as in equation A3, where Ai and Bi are its components of vectors A and B, respectively. Finally, we calculate the Cosine Distance as in equation A4. The greater Cosine Distance, the greater the model efficiency and accuracy (Hendy et al. 2023) [9]

Cosine Similarity Equations

$$A = \sum_{i=1}^{N} \frac{Vi}{N} \tag{A1}$$

$$A.B = \|A\|\ \|B\| \cos\theta \tag{A2}$$

$$\text{cosine similarity} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai^2}\sqrt{\sum_{i=1}^{n} Bi^2}} \tag{A3}$$

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity} \tag{A4}$$

B. BLEU Score

The BLeU score is an algorithm for evaluating the quality of text generated using deep learning algorithms (Papineni et al. 2002) [10]; accuracy is considered the correspondence between a machine's output and that of a human. The base stone of the BLeU score is the familiar precision measure, which is calculated by counting the number of candidate translation words (unigrams) that occur in any reference translation and then divided by the total number of words in the candidate translation as shown in equation A5 (Hendy et al. 2023) [9]. However, as in our bot task, the modified n-gram can be generalized as in equation A6 to the case: one candidate sentence and one reference sentence, where ŷ is candidate sentence and y is one reference sentence. Then, we start with the n-gram count summation as in equation A7 (Hendy et al. 2023) [9]. This count summation cannot be used to compare sentences since it is not normalized. If both the reference and the candidate sentences are long, the count could be huge, even if the candidate is of poor quality (Hendy et al. 2023) [9]. So, we normalize it as in equation A8, and equation A9 shows the final definition of BLEU, where w := (w1, w2, ⋯) is the weighting vector, and Ŝ := (ŷ (1), ⋯, ŷ(M)) is candidate corpus, and S = (S1, ⋯, SM) is reference candidate corpus

BLeU Score Equation

$$p_n(\hat{S};S) := \frac{\sum_{i=1}^{M} \sum_{s \in G_n(\hat{y}^{(i)})} \min(C(s,\hat{y}^{(i)}), \max_{y \in S_i} C(s,y))}{\sum_{i=1}^{M} \sum_{s \in G_n(\hat{y}^{(i)})} C(s,\hat{y}^{(i)})} \quad \text{(A1)}$$

$$\sum_{s \in G_n(\hat{y})} \min(C(s,\hat{y}), C(s,y)) \quad \text{(A2)}$$

$$\sum_{s \in G_n(\hat{y})} C(s,y) = \text{number of n-substrings in } \hat{y} \text{ that appear in } y \quad \text{(A3)}$$

$$p_n(\hat{y};y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s,\hat{y}), C(s,y))}{\sum_{s \in G_n(\hat{y})} C(s,\hat{y})} \quad \text{(4)}$$

$$BLEU_w(\hat{S};S) := BP(\hat{S};S) \cdot \exp\left(\sum_{n=1}^{\infty} w_n \ln p_n(\hat{S};S)\right) \quad \text{(5)}$$

VI. Results

A. Results of Unbalanced Data

For MAQA this unbalanced data we run each class in LSTM, Bi-LSTM and Transformer. We get max average similarity in transformer is 91.8% and average bleu 63.1%.

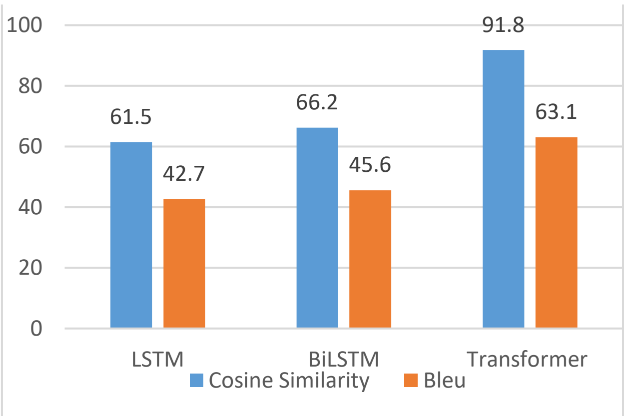

figure 2. results of unbalanced data

In our dataset, the class (أمراض الجهاز التنفسي) demonstrated the highest performance, with 8,764 records and similarity score of 96.65%. This result is achieved using specific model parameters optimized for this class. Conversely, the class (أمراض نسائيه) exhibited the lowest performance, despite having 70,160 records, with a similarity score of 90.2%. This outcome also resulted from applying specific model parameters tailored for this class.

For the MAQA run on the entire dataset, we utilized the Transformer model due to its superior performance, achieving an average similarity of 83.8% and BLEU score of 59.58%.

B. Results of Balanced data

We run class by class for each class parameter for his data. The parameters changed based on length of data in file and mean length. Then we run classification to get model file this is suitable file model of generation models. Then we load model and enter question that have predicted category equal to class that model trained on it. This technique we used it to get result in scope of question, we use two generation model in it (fine tune and transformer).

We get in fine tuning model 96.6% similarity and 63% in bleu. We use transformer although and achieve 95.4.% similarity and 61.5% bleu.

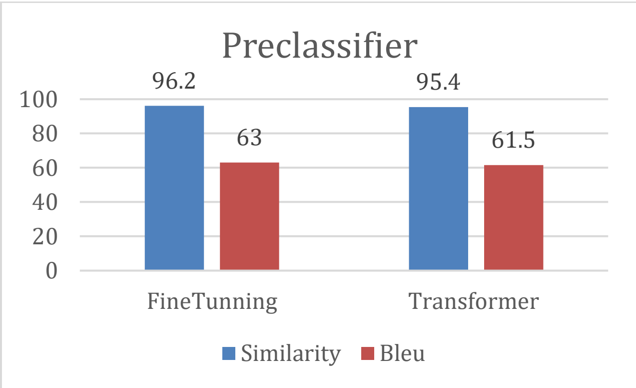

figure 3: Pre-classifier Result

C. Results of all data

For MAQA run in all data we use transformer because it has max average similarity and bleu and achieve similarity 96.2% and bleu 63%.

Data	Models	
Unbalanced Data	**Transformer** Similarity :83% Bleu: 59.9%	
Balanced Data	**Transformer** Similarity :95.7% Bleu: 62%	**Fine Tuning** Similarity :96.2% Bleu: 63%

Table 6: compare between balanced & unbalanced data

D. Diagnosing Expermental results

We split data in 80%train and 20%test. We use machine learning models and train it on embedding data. We use random forest model that give us accuracy 98.75%. We try decision tree and achieve 94.31% accuracy. We try SVM and achieve 80.4%. Final model hard voting in data using (random forest, decision tree, SVM) We achieved 99.95% and this is the best data. Then we take predicted disease and search on description file to get description of this disease. And in finally we search in file of symptom precaution to get some advice to this disease user need to enter at least 4 symptoms to get high accuracy of prediction

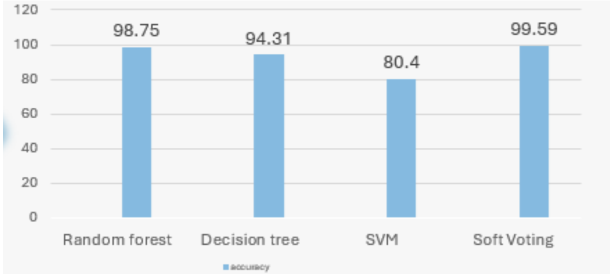

figure 4: Diagnosis Results

VII. Discussion

The discussion section delves into the complexities of the Arabic language and its dialects, emphasizing the challenges they pose for artificial intelligence, particularly in chatbot development. The necessity of chatbots is highlighted, especially in sectors like healthcare where they can provide significant benefits.

Our project, "DIGNOSY," aims to enrich the Arabic language's digital presence and improve healthcare delivery through a versatile chatbot capable of communicating in Arabic dialects. The chatbot is built on sequence-to-sequence models, incorporating LSTM, Bi-LSTM, and Transformer architectures.

Experimental results demonstrate the chatbot's efficiency, although it was trained on a limited dataset. Analysis reveals a correlation between the chatbot's effectiveness and the sentence length of the input data, impacting both questions and answers.

In summary, the proposed chatbot shows promise in enhancing healthcare communication in Arabic, with potential for further improvements and applications in other fields .

VIII. Conclusion

The Arabic language and its dialects are very complex. This poses a new challenge for artificial intelligence. Chat bots are more important now than ever because they can save you a lot of different services in many fields. So, creating The Chabot will be versatile in Arabic, a wonderful addition to enriching the Arabic language and its uses. Furthermore, the healthcare field is essential in our lives. Given its importance, it will be useful to use chatbots to support this area and provide health assistance to patients. In this paper, we propose a new Chabot called" DIGNOSY" to help patients and improve healthcare delivery. The proposed Chabot adopts the Arabic dialect to communicate with a larger number of patients. Technically Speaking, the created" DIGNOSY" is based on the sequence-to-sequence model with LSTM, BiLSTM, and Transformer. Furthermore, the experimental results showed good efficiency of the proposed Chabot despite the number of data used in this study (num of data). Moreover, the analysis also showed the extent to which the Chabot's effectiveness was related to the length of the sentences used, whether of questions or answers.

IX. Acknowledgment

This project is the result of the cooperation of multiple individuals, without whom this system would not exist today.

All praise and thanks to ALLAH, who provided me with the ability to complete this work. I hope to accept this work from me.

I am grateful to *my parents* and *my family* who are always providing help and support throughout the whole years of study. I hope I can give that back to them.

I also offer my sincerest gratitude to my supervisors, *Prof. Dr. Dina El-Sayad* and *T.A. Radwa El-Hussieny* who have supported me throughout my thesis with their patience, knowledge, and experience.

Finally, I would like to thank my friends and all the people who gave me support and encouragement.

X. References

[1] Abdelhay, Mohammed, Ammar Mohammed, and Hesham A. Hefny. "Deep learning for Arabic healthcare: MedicalBot." *Social Network Analysis and Mining* 13.1 (2023): 71.

[2] Boulesnane, Abdennour, et al. "Dzchatbot: a medical assistant chatbot in the algerian arabic dialect using seq2seq model." *2022 4th international conference on pattern analysis and intelligent systems (PAIS)*. IEEE, 2022.

[3] Bao, Qiming, Lin Ni, and Jiamou Liu. "HHH: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention." *Proceedings of the Australasian computer science week multiconference*. 2020.

[4] Mishra, Prateek, et al. "Personalized Healthcare Chatbot: Dataset and Prototype System." *International Conference on Computational Intelligence in Communications and Business Analytics*. Cham: Springer International Publishing, 2022.

[5] Chatbot in Arabic language using seq to seq model https://www.researchgate.net/publication/355961405_Chatbot_in_Arabic_language_using_seq_to_seq_model

[6] Health Consultant Bot: Primary Health Care Monitoring Chatbot for Disease Prediction https://journal.50sea.com/index.php/IJIST/article/view/193

[7] Multilingual Healthcare Chatbot Using Machine learning https://www.researchgate.net/publication/352668726_Multilingual_Healthcare_Chatbot_Using_Machine_Learning

[8] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.

[9] Hendy, Amr & Abdelrehim, Mohamed & Sharaf, Amr & Raunak, Vikas & Gabr, Mohamed & Matsushita, Hitokazu & Kim, Young Jin & Afify, Mohamed & Awadalla, Hany. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. 10.48550/arXiv.2302.09210

[10] Papineni K, Roukos S, Ward T, et al (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, https://doi.org/10.3115/10730 83.1073135