Arabic Chatbot Evaluation Based on Extractive Question-Answering Transfer Learning and Language Transformers

Tahani N. Alruqi and Salha M. Alzahrani *

Posted Date: 10 July 2023

doi: 10.20944/preprints202307.0609.v1

Keywords: Arabic; chatbot; transfer learning; AraBERT; CAMeLBERT; AarElectra (Generator/Discriminator); AraElectra-SQuAD

Article

Arabic Chatbot Evaluation Based on Extractive Question-Answering Transfer Learning and Language Transformers

Tahani N. Alruqi [1] and Salha M. Alzahrani [2,*]

[1] Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; s44280557@students.tu.edu.sa
[2] Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; s.zahrani@tu.edu.sa
* Correspondence: s.zahrani@tu.edu.sa

Abstract: Chatbots are computer programs that use artificial intelligence to imitate human conversations. Recent advancements in deep learning have shown interest in utilizing language transformers, which do not rely on predefined rules and responses like traditional chatbots. This study provides a comprehensive review of previous research on chatbots that employ deep learning and transfer learning models. Specifically, it examines the current trends in using language transformers with transfer learning techniques to evaluate the ability of Arabic chatbots to understand conversation context and demonstrate natural behavior. The proposed methods explore the use of AraBERT, CAMeLBERT, AraElectra-SQuAD, and AraElectra (Generator/Discriminator) transformers, with different variants of these transformers and semantic embedding models. Two datasets were used for evaluation: one with 398 questions and corresponding documents, and another with 1395 questions and 365,568 documents sourced from Arabic Wikipedia. Extensive experimental works were conducted, evaluating both manually crafted questions and the entire set of questions, using confidence and similarity metrics. The experimental results showed that the AraElectra-SQuAD model achieved an average confidence score of 0.6422 and an average similarity score of 0.9773 on the first dataset, and an average confidence score of 0.6658 and similarity score of 0.9660 on the second dataset. The study concludes that the AraElectra-SQuAD model consistently outperformed other models, displaying remarkable performance, high confidence, and similarity scores, as well as robustness, highlighting its potential for practical applications in natural language processing tasks for Arabic chatbots. The study suggests that the AraElectra-SQuAD model can be further enhanced and applied in various tasks such as chatbots, virtual assistants, and information retrieval systems for Arabic-speaking users. By combining the power of transformer architecture with fine-tuning on SQuAD-like large data, this trend demonstrates its ability to provide accurate and contextually relevant answers to questions in Arabic.

Keywords: Arabic; chatbot; transfer learning; AraBERT; CAMeLBERT; AarElectra (Generator/Discriminator); AraElectra-SQuAD

1. Introduction

Computer programs called "chatbots" mimic human speech. They run based on Artificial Intelligence (AI) technologies and natural language comprehension. Customer service, e-commerce, healthcare, banking, gaming, education, travel, tourism, and other sectors utilize chatbots extensively. English chatbots have been the subject of extensive investigation since 1960th. In 1966, the first chatbot with the name Eliza was created [1]. By returning the user's phrases in the interrogative form, Eliza acted as a psychotherapist which was an inspiration for the creation of later chatbots. Despite its limited communication capabilities, Eliza chatbot employed pattern matching and response selection methods based on template sets; thus, could discuss a restricted range of subjects. Additionally, this early chatbot was unable to maintain lengthy dialogues or gather context from the discourse. These days, a broad range of chatbots have been developed which can be divided based on techniques used into traditional rule-based chatbots and intelligent chatbots. In traditional

 doi:10.20944/preprints202307.0609.v1

rule-based chatbots, a set of rules and template of answers set are used with pattern matching and response selection methods. On the other hand, intelligent or smart chatbots, are developed based on AI, natural language processing (NLP) and understanding (NLU) methods and can respond in a way that is human-like. To respond to queries and access to a sizable body of knowledge, machine learning (ML) and deep learning have been widely used in chatbots research.

Arabic chatbots are uncommon because of the nature and intricacy of the Arabic language. The usage of chatbots in Arabic is a relatively recent trend, but it has a lot of promise. Arabic chatbots have enormous potential and given the explosive rise of digital technology, many companies and organizations are keen to use them, leading to revolutionize how users and consumers speaking Arabic interact with their services. Modern NLP and NLU, which enable the interpretation and creation of computerized methods for natural languages, have enabled the development of Arabic chatbots. As a result, it has become simpler for different sectors, businesses, and companies to create chatbots that can comprehend and answer questions in Arabic. Examples of Arabic chatbots research works include the design and frameworks of building and utilizing them in different applications [2-4]. To build an Arabic chatbot, one needs to comprehend the language, culture, and intricacies of Arabic. In addition to the technical aspects of building a chatbot, the cultural, and social contexts of the regions targeted by the chatbot should be guaranteed to be able engage in meaningful discussions with people. An intelligent chatbot should also be familiarized with the many dialects and nuances in Arabic language and must be able to comprehend the cultural context of interactions. Consequently, rather than relying just on pre-written responses, the chatbot should be able to converse with customers naturally. For instance, there are several dialects and accents in Arabic; therefore, the chatbot must be able to comprehend them while interacting with people. One important characteristic is that the chatbot can understand other languages and translate them into Arabic.

Researchers have devoted significant time and effort to developing frameworks and structures for chatbots in various applications. The incorporation of advanced technologies like natural language processing and question-answer databases enhances the conversational abilities of chatbots. There are different types of question-answering (QA) systems used in building chatbots that have variations in the way they produce answers. One such type is extractive QA, where the system retrieves the answer from a given context and presents it to the user using BERT-like models. Another type is generative QA, which generates a textual answer from scratch using AI-based generation models. Extractive QA models use transfer learning which offers a pre-trained model with a lot of data, and uses self-attention to uncover relationships between words in one sequence that are dependent on one another [5,6]. Few research studies that investigated the use of machine learning, deep learning, and NLP methods in Arabic chatbots [7-10]. To the best of our knowledge, none of these studies have investigated extractive QA and transfer learning methods using language transformers to evaluate Arabic chatbots which can meaningfully understand the conversations context and respond naturally. Therefore, the following objectives are investigated in this study: (i) to study the recent BERT-like language transformers that were pre-trained on large collections and corpora of Arabic language. (ii) to investigate the current datasets available for Arabic chatbots and related fields such as Arabic QA datasets, and finally (iii) to evaluate transfer learning methods for Arabic chatbots using Arabic language transformers namely AraBERT, CAMeLBERT, AraElectra-SQuAD and AarElectra (Generator/Discriminator) with NLP evaluation metrics.

The rest of this study is organized as follows. In section 2, a comprehensive literature review about chatbots is discussed including chatbot system architectures, applications, and methods used in chatbots research namely rule-based, corpus-based, and retrieval-based methods, and extractive versus generative-based AI methods. In section 3, the proposed methodology of this work is explained including the problem formulation and general framework, transformers used for Arabic language, and the methods of transfer learning proposed in this study. The datasets exploration, resources, tools, and evaluation metrics are given in section 4. Experimental results and discussion are presented in section 5. Finally, the concluding remarks and directions to the future works for Arabic chatbots research are given in section 6.

 doi:10.20944/preprints202307.0609.v1

2. Literature review

In this section, we reviewed chatbots literature studies from the data science point-of-view covering the range of the years between 2019-2023. Background knowledge or theoretical concepts were taken from references before this range of years.

2.1. Chatbot system architecture and applications

The general architecture of chatbot input and response generation by Adamopoulou and Moussiades [11] is shown in **Figure 1**. Beginning with a user asks a question via a messaging service like WhatsApp, the User Message Analysis Component (UMAC) analyzes it to determine the entity and intent categorization. Then, the Dialog Management Component (DMC) handles all textual and voice messages that are sent back and forth (questions, requests, and responses). A chatbot must choose its next course of action once it has reached the best interpretation it can. It has several options, including asking for clarification, requesting more background information, remembering what it has learned, and waiting to see what occurs next. Once the request is comprehended, the chatbot seeks data from a database or through API calls from the backend. The knowledge base utilized by rule-based chatbots comprises handwritten replies to user inputs, whereas the generative model uses natural language generation to provide replies. Finally, the chatbot outputs the message to the user to answer what was requested. A variety of factors may be used to classify chatbots. These criteria may include the way in which the chatbots are implemented, the knowledge domain they are used in, how they are used, and the methodologies used to generate responses. Depending on how a chatbot interacts with its input and output, it can be classified as a text or speech-based conversational dialogues**.** While open-domain chatbots may converse about a variety of topics, closed-domain chatbots concentrate on specific knowledge area and may struggle to address unrelated issues. Task-oriented chatbots are created to help users complete specific tasks in a particular domain, whereas non-task-oriented chatbots encourage user participation across several topics and act as informational chatbots. Chatbots are divided depending on their response generation approaches into rule-based, corpus-based, retrieval-based and extractive/generative AI-based chatbots. Rule-based chatbots are designed to respond to specific queries and can effectively produce answers based on established guidelines. However, they are ineligible to handle inputs with spelling and grammar errors, and only handle the most recent message entered. Corpus-based and retrieval-based methods can handle a wider range of questions than rule-based by employing statistical language modeling. Recently, extractive/generative AI-based chatbots are powered by AI designed to communicate with users much like actual people do by using NLP, NLU, ML, deep learning, and more to comprehend the user's intent, with the capability of remembering context and word diction. The process of training AI-based chatbots initiated with datasets collected for the purpose of the chatbot, preprocessing required to clean the data, designing, and training the model which can be either a machine learning or deep learning model, and finally testing and evaluating the model [12].

Chatbots are used in a wide range of industries and applications, such as customer service, e-commerce, healthcare and medical applications, education, and language learning [10] as summarized in Table 1. In customer service, the chatbot is used to respond to frequently asked inquiries, resolving problems, and delivering information, chatbots are utilized to provide customer care [13,14]. In e-commerce chatbots, product recommendations, responding to inquiries about specific items, and guiding users around a website are used by the chatbots with online shopping [15,16]. Further, healthcare, pharmaceutical, and medical diagnosis chatbots can help with appointment scheduling, giving medical advice and information, and even classifying symptoms [17,18]. Educational chatbots are used to support students, help with registration and enrollment, and give information about courses and activities [19,20]. Finally, for aiding in language learning, chatbots can be used to help users learn a second language, or correct grammar, and aid in writing/speaking improvement [5,21].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

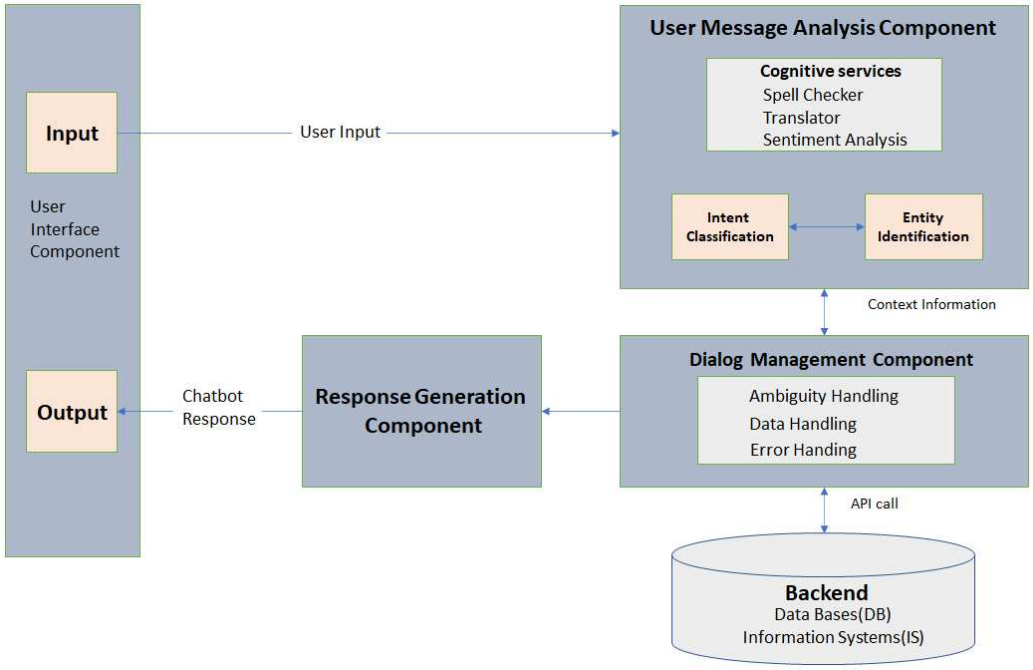


Figure 1. General chatbot input and response generation [11].

Table 1. Summary of research works on chatbots applications including customer service, e-commerce, healthcare and medical diagnosis, education, and language learning.

Industry/Application	Description	Ref.
Customer service	Sentiment and intent analysis and emotion recognition in customer service chatbots	[13,14]
	Goal-oriented conversation management bootstrapping	[14]
e-Commerce and Telecom	Engagement with chatbots versus augmented reality interactive technology in e-commerce	[16]
	Information technology telecom chatbot	[22]
Healthcare and Medical diagnosis	Ask Rosa: digital genetic conversation chatbot about hereditary breast and ovarian cancer	[17]
	AI-Powered health chatbots general architecture	[18]
	Chatbot for disease prediction and treatment recommendation	[23]
	Mental healthcare chatbots	[24,25]
Education	Highly adaptive educational chatbot	[19]
	NEU-chatbot: chatbot for admission of National Economics University	[15]
	Educational and smart chatbots for colleges and universities	[20,26-28]
Language learning	Chatbot assistant for English as a second language learners	[5,21]

2.2. Chatbot methods

2.2.1. Rule-based methods

Traditional and early research works in chatbots have employed rule-based methods [29-31] which utilized a collection of predetermined human-created rules, applied in a hierarchy to transform

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

user input into an output. Instead of creating a brand-new answer, the rules divide the input into a series of tokens to look for patterns and produce a response. Although this strategy is simple and straightforward to develop, it confines responses to inputs that fall within the stated rules only and unpleasant experience with questions/queries that are not in the collection [32].

2.2.2. Corpus-based and retrieval-based methods

Coming after the traditional rule-based chatbots, the corpus-based and retrieval-based methods have been employed widely by utilizing a corpus or knowledge base using a statistical language approaches [32]. The majority of chatbots in this category employed information retrieval to extract a candidate response from the corpus based on heuristics approaches. In these methods, both the input and the context were considered by recognizing keywords to provide the best answer from the corpus/knowledge as opposed to employing predetermined criteria.

2.2.3. Extractive-based and generative-based methods

AI-based solution does not require any prepared replies; instead, it creates the responses depending on the dialog context. By considering both recent and past user interactions, the chatbot tries to come up with a fresh response. It is necessary to gather a sizable training set, which could be challenging. Due to the real-time nature of this method, response failures are quite likely [32]. AI-based chatbots can be classified into extractive and generative QA systems based on the method used to provide the responses. Extractive QA chatbots are designed to answer questions by extracting the relevant information from a given passage of text. They employ algorithms to identify key pieces of text in the passage that are most likely to contain the answer and then extract those pieces of text to provide an answer to the question. Generative QA chatbots, on the other hand, generate responses to questions from scratch using AI generation models trained on large collections. While extractive QA models tend to produce more accurate answers, generative models can generate entirely new responses that may provide a deeper understanding of the given text. Both types of QA models have their own strengths and weaknesses and are used in different applications depending on the desired outcome. The following subsections discuss various methods used in generative and extractive chatbots research including NLP, NLU, ML, recurrent neural networks (RNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), encoder-decoder, sequence-to-sequence (Seq2Seq), and reinforcement and transfer learning.

NLP methods rely on computational linguistics and statistical modeling with textual data. NLU, as a part of NLP, analyzes texts and voices using syntactic and semantic thesauruses and knowledge bases. While semantics relate to the sentence's intended meaning, syntax refers to a sentence's grammatical structure. A general architecture for chatbots, which combined dialogue and communication components with NLU, as well as expert components was developed based on deep learning [8,18]. Their AI-powered chatbots enabled the interaction with users in a more human-like manner while providing accurate answers to their questions related to either open domain [8], or healthcare closed domain [18].

ML methods have been used in several research works for chatbots [7,12,22,23,33-37]. ML chatbots research works were investigated and reviewed from 2020 and before by Suta, Lan, Wu, Mongkolnam and Chan [10]. ML was utilized to understand the relationships and intentions in queries using dimple logistic regression and iterative classifier optimizer which achieved 97.95% accuracy in predicting users' intention higher than other classifiers [33]. Support vector machine (SVM) algorithm to predict the health status of users was integrated with Google API for speech-to-text and text-to-speech conversions [34]. K-nearest neighbor method (KNN) was used to extract symptoms from conversations to provide diagnosis and therapy recommendations [23]. SVM-trained model was examined on women's potential for depression, anxiety, and hypomania. To give users spiritual support and medical guidance, the user's mental health data was gathered and assessed in real time through a chatbot paired with the psychological test scale for additional diagnosis [24]. A chatbot framework using ML was proposed aiming at diagnosing and solving technical issues using extracted data from technical support tickets in HP and Microsoft in Arabic [7].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

RNNs are a subset of artificial neural networks that can handle sequential input by passing information from one step in a sequence to the next via loops [6]. Several research works have used RNN methods for different purposes in chatbots such as classifying the users' intents [38], detecting users' emotions and feelings [13], and handling long conversational dialogues [39]. RNNs were able to comprehend similar-sounding phrase variants and enhanced a conversation intent classification model which obtained 81% accuracy [38]. Another study collected the data based on the discussions and sentiment analysis, and used recent talks as input to RNNs to classify the feelings which obtained 0.76 precision accuracy [13]. Bidirectional RNN and attention model were used to generate responses to long queries (more than 20-40 words) [39].

LSTMs, as kind of RNNs, were created to capture long-term dependencies by retaining and forgetting information over a range of time steps [6]. Different studies have utilized LSTMs in conversational dialogs to provide better understanding of the context of a conversation and more accurate and coherent responses from the chatbot [26,40-43]. A study showed that the performance of a chatbot was improved using an ensemble of LSTM networks, rather than a single LSTM model, trained to learn long-term dependencies and the relationships between events that occur over a prolonged period [40]. The impact of context learning on the overall chatbot performance using LSTM with metaphorical approach was examined [41]. Interactive chatbots based on LSTM and NLP algorithms were developed for teaching, student-serving, and responding to queries posed by pupils [26,42]. Furthermore, a study by Anki, Bustamam, Al-Ash and Sarwinda [43] proposed bidirectional LSTM, known as BiLSTM, to analyze input sequences both forward and backward, enabling to recognize long-term connections between sequence's component parts. The model's performance was assessed using many datasets, and the chatbot achieved an average accuracy of 0.99.

GRUs are kind of RNNs that employ gating mechanisms to regulate the information flow to and from memory cells. With fewer parameters and processes, a GRU unit often performs as well as or better than an LSTM [6]. A study utilized GRUs for chatbots in web interfaces which proved that the performance of GRU was better in question answering than BiLSTM using Facebook bAbi dataset [41].

Encoder-Decoder structures contain the encoder which transforms input data into the internal representation of the network using a fully connected hidden bottleneck layer whereby its activation vector is considered the internal state. On the other hand, the decoder attempts to rebuild the input from the internal data model of the network [6]. Several research works have used encoder-decoder methods and attention mechanisms for chatbots and conversational dialogs [12,44]. Both studies intended to improve responses generation to user queries, the experience and interaction of humans with the chatbots technology.

Seq2Seq models are based on encoder-decoder architectures widely used in NLP tasks such as machine translation, text summarization, QA, text generation, and more, whereby the input is a sequence of data, and the output is another sequence. Seq2Seq models consist of encoder network to convert the input sequence into a fixed-length vector representation and decoder network to decode the representation into the output sequence [6]. A study suggested *midoBot*, a seq2seq-based deep learning Arabic chatbot which textually converse with others on common conversational subjects [45]. Another study developed a chatbot using Seq2seq encoder-decoder architecture trained on brain disorders and mental illnesses data, which was successful to reacts sympathetically with people having mental illnesses [25].

Reinforcement learning models train the machine by using incentives and penalties. The model's objective is to maximize its overall rewards during the training, and it accomplishes this by acting in ways that can alter the environment [6]. The reinforcement learning was utilized in chatbots [46], wherein the reward method enables the chatbot to distinguish between correct and incorrect responses. With the use of deep reinforcement learning algorithms, their chatbot can recognize the tone of the question and respond appropriately. Q-learning, deep Q-neural network (DQN), and distributional reinforcement learning with quantile regression methods were utilized in the suggested system (QR-DQN) and the performance of each method was investigated and evaluated. Ensemble-based deep reinforcement learning for chatbots with sentence clustering and dialogue

 doi:10.20944/preprints202307.0609.v1

clustering was developed and trained on dialogue raw textual data only without any manually-labelled data [47], which concluded that using ensemble chatbot agents were highly correlated with human-rated data.

Transfer learning leverages the knowledge learned by a pre-trained network on a large dataset to a new, related problem [6]. In order to counteract the negative consequences of the limited availability of data for a chatbot in a specific domain, several research works have been developed to modify a model, originally developed for one task, and apply it to a related task, specifically chatbots [5,14,48,49]. This was achieved by enhancing a pre-existing language model and fine-tuning it on a specific set of conversational data, like medical consultations or customer service encounters. In this regard, chatbots employ the broad linguistic comprehension skills learned from the pre-trained model while simultaneously learning the specific terms and language used in a closed domain. The creation of chatbots using transfer learning-based strategies involved fine-tuning of pre-trained transformer models such as BERT or GPT-2 and pre-trained embeddings like GloVe and ELMo together with the neural network architecture using available conversational data [5]. A method of utilizing transfer learning improved the chatbot performance by 20% for open domains and more than doubled the improvement for closed domains [14], wherein the most favorable outcomes arose when the transfer learning technique was merged with complementary processing techniques such as warm-starting. Another study investigated the use of transfer learning to enhance the ranking of responses in extractive-based QA chatbots [49].

2.3. Discussions and gap

The trend of using both extractive-based and generative-based AI methods for chatbots is increasing. Many methods of NLP, and NLU [8,18], ML [17,20-22,24,50], deep learning including RNN, LSTM, GRU, Encoder-Decoder, and Seq2Seq [5,12,15,25,26,38-45,47], and reinforcement learning [46,47] have been applied successfully in chatbots and conversational dialogs research.

Table 2 summarizes several research works that have been done using these models, datasets used, and their advantages and disadvantages from our perspective. The table also shows a comparison between the results of accuracies obtained by different models implemented in these studies. One of the latest directions in chatbots research is the use of transfer learning by some research works [5,14,48,49] whereby the knowledge leveraged from a pretrained model helps the chatbot to understand and comprehend the context while interacting with people. As there that investigated the use of ML, NLP and deep learning methods in Arabic chatbots [7-9,27,28,51], we believe that modern NLP which enables the interpretation and creation of computerized methods for natural languages, with the recent trends of transfer learning have the potential to the development of Arabic chatbots. Our literature review corroborates that the use of transfer learning with recent language transformers will open directions for a more specialized extractive and generative QA in Arabic chatbots. Therefore, this research aims to bridge this gab towards Arabic chatbots that understand the conversation context and behave naturally as humans do, by conducting a thorough implementation of BERT-like Arabic transformers, comparison, and evaluation using transfer learning on existing QA datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

Table 2. Comparison of research works in chatbots and conversational dialogs including NLP, and NLU, machine learning and deep learning including RNN, LSTM, GRU, Encoder-Decoder, Seq2Seq, with the datasets, accuracy results, pros and cons used in each study.

Method	Description	dataset	Accuracy	Pros.	Cons.	Ref.
NLP NLU	AI-powered healthcare chatbots Arabic NLU chatbot framework	-	-	utilize NLU, NLG, deep learning	inaccurate data decrease accuracy	[8,18]
ML	Acceptance of Chatbot based on Emotional Intelligence through Machine Learning Algorithm	international students with experience in using chatbot	97%	TAM and EI theory to predict users' intentions	data limited to international students, making it difficult to interpret	[33]
	An Improved Chatbot for Medical Assistance using Machine Learning	various sources: medical journals, online forums, and websites	93 %	streamline medical processes and save time	SVM's accuracy may not be perfect	[34]
	Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning	comprised of patient data, medical history, and symptoms	-	alternative to hospital visits-based diagnosis	not as accurate as traditional hospital visits	[23]
	Supervised Machine Learning Chatbots for Perinatal Mental Healthcare	pregnant women, newborns, and their families	-	reduce barriers and help clinicians make accurate diagnoses	not accurately detect subtle changes in mental health	[24]
	A Novel Framework for Arabic Dialect Chatbot Using Machine Learning	extracted IT problems/ solutions from multiple domains		accuracy, response time	no explanation of how ML was employed	[7]
RNN	Intents Categorization for Chatbot Development Using Recurrent Neural Network (RNN) Learning	university guest book available from its website	81%	understands variations in sentence expression	requires big data, difficult or expensive to implement	[38]
	Conversations Sentiment and Intent Categorization Using Context RNN for Emotion Recognition	conversations inside a movie	79%	successful to recognizing emotions in text-based dialogs	only uses a single dataset for testing the algorithm	[13]
	Deep learning with Bidirectional RNN and attention model	Reddit dataset	-	perform English to English translation	No accuracy measured	[39]
LSTM	LSTM Based Ensemble Network to Enhance the Learning of Long-term Dependencies in Chatbot	Cornell Movie Dialog Corpus	71.59%	retain contextual meaning of conversations	-	[40]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

	A Metaphorical Study of Variants of Recurrent Neural Network Models for Context Learning Chatbot	Facebook bAbi dataset	96%	help to create chatbots for web applications	only tests RNN models on a single dataset	[41]
	Natural language processing and deep learning chatbot using long-short term memory algorithm	conversations with users and assessments	-	understand questions and provide detailed answers	does not address accuracy and reliability	[42]
	AI based Chatbots using Deep Neural Networks in Education	set of answer and question pairs	-	provide accurate and useful responses to student queries	incorrect / difficulty handling complex queries	[26]
	AI Chatbot Using Deep Recurrent Neural Networks Based on BiLSTM Model	Cornell Movie Dialog Corpus	99%	outperform other chatbots in accuracy and response time	only compares with a few other systems.	[43]
GRU	A Metaphorical Study of Variants of Recurrent Neural Network Models for A Context Learning Chatbot	Facebook bAbi dataset	72%	-	-	[41]
Encoder-Decoder	AI Chatbot Based on Encoder-Decoder Architectures with Attention	Cornell movie subtitle corpus	-	improve the experience and interaction	lack of review of similar methods	[12]
	Behavioural Chatbot Using Encoder-Decoder Architecture	-	-	increase replicability	focus on chatbot to mimics fictional character, limiting its generalizability	[44]
Seq2seq	Chatbot in Arabic language using seq to seq model.	~81,659 pairs of conversations	-	use common conversational topics	no detailed description of the dataset, making it difficult to replicate	[45]
	Mental Healthcare Chatbot Using Seq2Seq Learning and BiLSTM	The Mental Health FAQ	-	assist mental healthcare	-	[25]
Transfer Learning	Goal-Oriented Chatbot Dialog Management Bootstrapping with Transfer Learning	-	-	overcome low in-domain data availability	focuses on technical aspects not chatbot performance	[14]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

	The Design and Implementation of English Language Transfer Learning Agent Apps	English Language Robot	-	integrate recognition service from Google and GPT-2	no comparison with existing chatbots for language learning	[5]
	Building Chatbot Using Transfer learning: End to end implementation and evaluation	-		show fine tuning and optimizing	no comparison evaluation	[48]
	Reranking of Responses Using Transfer Learning for a Retrieval-Based Chatbot	WOCHAT dataset Ubuntu dialogue dataset		highest ratings from the human subjects	-	[49]
Reinforcement Learning	Evaluating the Performance of Various Deep Reinforcement Learning Algorithms for a Conversational Chatbot	Cornell Movie-dialogs corpus and CoQA (A Conversational QA Challenge)	-	comprehensive review of reinforcement learning	difficult to compare to other approaches	[46]
	Ensemble-based deep reinforcement learning for chatbots	Chitchat data		training ensemble of agents improved chatbot performance	Require more training time	[47]

3. Methodology

3.1. General framework

Arabic chatbots, in general, can be created using a union of ML, NLP, NLU and/or transfer learning techniques which can operate as shown in Figure 2. There are two sides in the chatbot operation which are the user question (i.e., query) referred to as *Q*, and the bot answer (i.e., response) referred to as *R*. A given textual data referred to as context, *C*, is required to train the chatbot. Arabic NLP is required to split the query into tokens (sentences, phrases, words, etc.), and Arabic NLU is also required to exploit the meanings, nuances, and synonyms used in these tokens. ML is part of this framework as we need to train the model with a sizeable dataset of questions and their contexts (*Q*, *C*, *R*). Deep learning can also be utilized to train the chatbot models with a set of questions and their contexts (*Q*, *C*) and learns automatically to generate responses and interact naturally. To elaborate, the context "اسمي تهاني وأسكن في الطائف " is given to the model with queries to be trained answering them. Then, if the user gives a query or question to the chatbot, " اين تعيشين؟"، where we use a different Arabic word to mean the same action of "living", the chatbot should be able to answer the question, "في الطائف". Instead of training the chatbot model from scratch, pre-trained language models which are so-called transformers can be utilized with the transfer learning approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

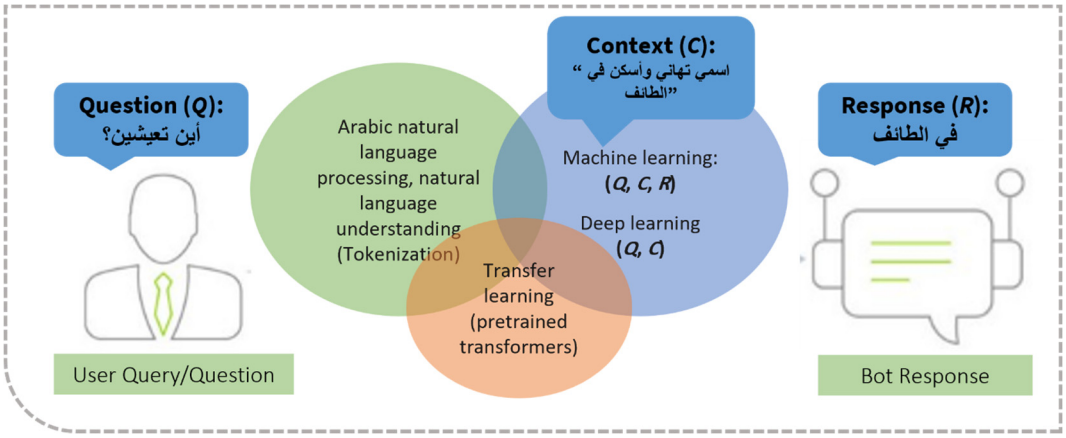


Figure 2. Operational framework of Arabic chatbot using transfer learning.

Question-answering (QA) is the process of using a natural language processing model to automatically answer questions posed in natural language. Extractive QA involves interpreting the question, searching through a large corpus of information, and identifying the most relevant information to extract the correct answer. The goal of our method is to develop an Arabic chatbot based on extractive QA that can accurately and efficiently provide answers to questions posed in Arabic, similar to how a human expert might provide answers. The methodology proposed for this research study is shown in Figure 3. As can be seen, the datasets for Arabic QA that are best related data to this problem were utilized including hundreds of questions, as will be explained shortly. In order to train models for extractive Arabic QA, we implement transfer learning and fine-tuning on four sets of transformers namely AraBERT, CAMeLBERT AraElectra-SQuAD and AraElectra (Generator/Discriminator) transformers with different variations and semantic embedding models, which achieve state-of-the-art results for Arabic NLP problems. When implementing transfer learning, one dense layer and a softmax layer were added to fine-tune the AraBERT and CAMeLBERT pre-trained models because they were pre-trained as general text prediction models, while only a softmax layer was added at the top of AraElectra-SQuAD and AraElectra (Generator/Discriminator) transformers as these were pre-trained for text discrimination and QA data. Finally, our fine-tuned models were then used to predict unanswerable questions and evaluate the performance using confidence and similarity metrics used commonly in NLP research.

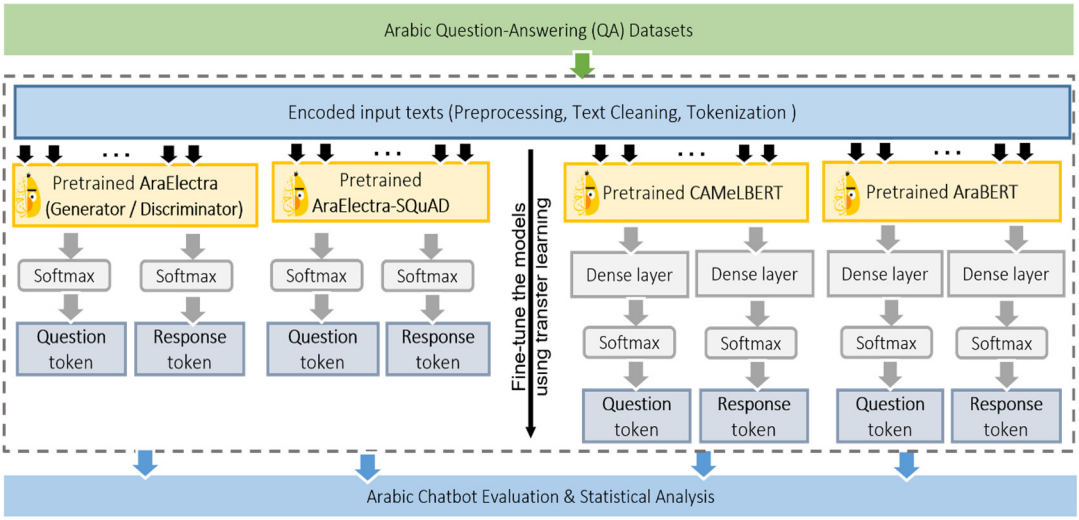


Figure 3. General framework of the methodology used in this study.

 doi:10.20944/preprints202307.0609.v1

3.2. Details of extractive QA and transfer learning method for Arabic chatbot

Extractive QA models involve creating a system that can automatically answer questions based on a given text corpus. To build our extractive QA models, transfer learning was applied with the aim to evaluate and improve the performance of conversational dialogues in Arabic chatbots. In the context of conversational dialogues, transfer learning involves training a model on a large dataset of conversational data to learn general patterns of language use and then fine-tuning the model on a smaller, task-specific dataset to improve performance on that task, similar to the work by [52].

The following are the steps involved in the detailed process of extractive QA used in this study:

- **Dataset preprocessing:** We utilized large datasets of questions and their corresponding answers, together with a large corpus collection of textual documents which contains the contexts these questions and answers were taken from. In this step, we implemented various pre-reprocessing steps to remove any unwanted elements such as special characters, stop words or noisy words. Further, we cleaned the corpus by removing any irrelevant or misleading information.
- **Initialization**: In this study, we used several pre-trained transformers. The final fully connected layer(s) of the pre-trained network was removed and replaced with new layer(s) that represent the questions/queries and responses/answers. This process saved a lot of time and computational resources compared to training a network from scratch, as the network can start from a good initial state based on its prior experience. Several parameters were initialized such as the patch size, the number of epochs, the learning rate, wherein we used initialization settings like the state-of-the-art studies in QA tasks in English.
- **Fine-tuning:** Several BERT-like transformers in Arabic were fine-tuned using large datasets of annotated QA pairs for the task of extractive QA. This step was crucial to achieve the aims of our study whereby the goal of the model was to read a passage of text and extract a concise answer to a given question from the passage. To elaborate, we first provided them with a dataset of questions and their corresponding answers, as well as the passage from which the answer was extracted. The model was then trained to predict the correct answer given a passage and a question. During the fine-tuning process, the transformers' last (i.e., added) layers were trained using a task-specific loss function that aims to optimize the model to generate the correct answer for a given question. The model was trained to select the answer by identifying the start and end positions of the answer in the passage. The fine-tuning process involved adjusting the weights of the pre-trained transformers using backpropagation to optimize the model's output. The loss function was minimized in several epochs to improve the model's accuracy in predicting the correct answer to a given question.
- **Testing and evaluation:** Once the fine-tuning was complete, our models were used for extractive QA in Arabic chatbot. When a user asks a question, the chatbot can feed the question into the model, which will then provide an answer based on corpus collection with a confidence score. Hence, our proposed models were tested on different datasets and real-world scenarios to check their robustness and accuracy. In order to compute the confidence and similarity scores, several semantic embedding models were used. A semantic embedding model is an NLP method that allows words or phrases to be represented as vectors of numbers in a multi-dimensional space. The idea behind this model is that words that are similar in meaning will be located close to each other in this space, while words that are dissimilar will

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

be located far apart. In this study, variants of distilbert and bert-based models for Arabic were employed to predict the answers or responses based on their surrounding context.

The BERT (Bidirectional Encoder Representations from Transformers) is known for its ability to capture a deeper understanding of the context of language by training a model to predict missing words in each sentence. This allows BERT to learn contextual relationships between words and provide more accurate and relevant results for NLP tasks. In this study, several BERT-like transformers developed for Arabic language which can be divided into two sets: the first set contains several variants from AraBERT and CAMeLBERT transformers pre-trained for general text predictions tasks, and the second set contains variants from AraElectra-SQuAD and AraElectra (Generator/Discriminator) transformers pre-trained for text discrimination and QA tasks. The AraBERT is a BERT model for Arabic which has achieved state-of-the-art performance on a range of NLP tasks and has become an important tool for many researchers and practitioners in the field. In this study, we used the base models and the large models, which also differ in the pre-segmentation techniques used. The CAMeLBERT are pre-trained BERT models for Arabic, including Modern Standard Arabic (MSA), dialectal Arabic (DA), and classical Arabic (CA), in addition to a model pre-trained on a mix of the three. These models can provide high-quality contextualized word representations for Arabic text.

On the other hand, models which are variants of the Electra model, which stands for Efficiently Learning an Encoder that Classifies Token Replacements Accurately, were pre-trained to perform well on the Stanford Question Answering Dataset (SQuAD). The AraElectra-SQuAD transformer is a language model for Arabic language comprehension pre-trained to read passages and generate the correct answers to given questions. The AraElectra-SQuAD transformer was fine-tuned specifically on Arabic SQuAD data to excel at QA tasks in Arabic. Another variant of Electra based models are AraElectra-base-generator and AraElectra-base- discriminator developed by the team at AUB Mind Lab, which differs in their underlying architecture and purpose. The generator is a generative model trained on large amounts of Arabic text data and can generate coherent and contextually relevant text based on a given prompt. It is designed to generate new text that is similar in style and content to the data it was trained on. The discriminative model is a classifier trained to classify Arabic texts into different categories based on the examples it was trained on. It is designed to make predictions or decisions based on input features. Table 3 summarizes variants of the transformers used in this study in terms of their size, task, description, and pre-training datasets.

Table 3. Summary of recent transformers pretrained for text generation and question answering tasks for Arabic language.

Transformer name (*based on Huggingface*)	Size	Task	Description	Pre-training datasets
aubmindlab/bert-base-arabertv02 **aubmindlab/bert-base-arabertv2** **aubmindlab/bert-base-arabertv01** **aubmindlab/bert-base-arabert** [1]	base	Text Generation	AraBERT is a pretrained Arabic language model with pre-segmented text, trained and evaluated similar to the original BERT in English.	OSCAR, Arabic Wikipedia, Arabic Books collected from various sources, Arabic News Articles and Arabic text collected from social media platforms, such as Twitter and online forums.
aubmindlab/bert-large-arabertv2 **aubmindlab/bert-large-arabertv02**	large	Text Generation		
aubmindlab/araelectra-base-generator [2]	base	Text prediction, QA	The generator model generates new text based on learned patterns from training data, achieved	OSCAR unshuffled and filtered, Arabic Wikipedia dump from 2020/09/01, the 1.5B words Arabic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

			state-of-the-art performance on Arabic QA datasets.	Corpus, the OSIAN Corpus, and Assafir news articles
aubmindlab/araelectra-base-discriminator [3]	base	Text prediction, QA	The discriminator model classifies or makes predictions based on input features	
CAMeL-Lab/bert-base-arabic-camelbert-mix [4] **CAMeL-Lab/bert-base-arabic-camelbert-ca** **CAMeL-Lab/bert-base-arabic-camelbert-da** **CAMeL-Lab/bert-base-arabic-camelbert-msa**	base	Text Generation	Pre-trained BERT models for Arabic texts with different dialects and structures, formal and informal Arabic.	MSA : Arabic Gigaword, Abu El-Khair Corpus, OSIAN corpus, Arabic Wikipedia, Arabic OSCAR DA : A collection of dialectal data CA : OpenITI (Version 2020.1.2)
ZeyadAhmed/AraElectra-Arabic-SQuADv2-QA [5]	base	QA	AraElectra-based model fine-tuned on question-answer pairs to predict unanswerable questions.	Arabic-SQuADv2.0 dataset

[1]. https://huggingface.co/aubmindlab/bert-base-arabert, [2]. https://huggingface.co/aubmindlab/araelectra-base-generator, [3]. https://huggingface.co/aubmindlab/araelectra-base-discriminator, [4]. https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix, [5]. https://huggingface.co/ZeyadAhmed/AraElectra-Arabic-SQuADv2-QA.

4. Experimental Setup

4.1. Datasets

In order to evaluate the proposed methods, we need to fine-tune the transformers on a corpus collection of Arabic texts, and then to evaluate their confidence in answering questions using a set of Arabic questions. Thus, the datasets utilized in this study contain twofold: a corpus of texts and a set of questions. One dataset was created by Aljawarneh [53] which contains 398 Arabic questions generated using augmentation techniques. The questions are reported in modern standard Arabic (MSA) but has no corpus collection of texts. Therefore, we collected a corpus of 398 documents from the web about the topics found in the questions set. We referred to this dataset as MSA-QA. Another dataset we used is the Arabic language comprehension dataset (ARCD) created by Mozannar, *et al.* [54]. This dataset which we referred to as ARCD-QA, comprises 1,395 distinct questions created by crowd-workers from articles on Arabic Wikipedia. Each question is accompanied by the corresponding article title, the context in which the question was raised, and a set of potential answer choices. The corpus collection we used with this dataset is Arabic Wikipedia dump 2021 which originally contains 600, 000 documents from Wikipedia. However, due to limitations in computing power, 365,568 documents were indexed and used in our study. A third dataset is Arabic AskFM dataset which comprised 98,422 question-answer pairs from AskFM platform posted in dialectal (informal) Arabic (mostly Egyptian dialects), and we referred to this dataset as DA-QA. The focus of the questions in this dataset is Islamic topics, and we used this dataset in some of our initial experiments. To get insights into these datasets we utilized several metrics such as word and character counts, recognition of frequent queries and terms, visual depiction of word frequencies, and word occurrence analysis. Table 4 summarizes the datasets used in this study.

Table 4. Summary of the datasets for Arabic question answering (QA) and Arabic chatbots used in this study.

Dataset	Number of Documents	Description

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

	Questions	Answers	Corpus	
MSA-QA	398	398	398	This repository of Arabic Questions Dataset [6] provides an Arabic question for data science and machine learning.
ARCD-QA	1,395	1,395	365,568	The corpus contains a comprehensive Arabic Wikipedia dump 2021 [7], including articles, discussions, and textual information from 2021. The questions were created by crowd-workers in ARCD [8]
DA-QA	98,422	98,422	98,422	Arabic AskFM dataset collection of questions and answers mostly about Islamic topics by various authors in dialectal Arabic (DA) on the AskFM platform.

6. https://github.com/EmranAljawarneh/Arabic-questions-dataset, 7. https://www.kaggle.com/datasets/z3rocool/arabic-wikipedia-dump-2021?datasetId=1179369, 8. https://www.kaggle.com/datasets/thedevastator/unlocking-arabic-language-comprehension-with-the.

4.2. Resources and tools

Resources utilized in this research include two PCs each with Intel(R) Core(TM) i7-10700T CPU with 2.00GHz and 16.0 GB RAM. Models fine-tuning and testing were implemented using TensorFlow 2.8.0, which uses Keras as a high-level API with various complementary libraries such as Ktrain, Scikit-Learn, Matplotlib, and Pandas. The first patch of experiments using 10 questions took around 10-20 minutes each. The second patch of experimental works using ~400 questions took around 2-4 hours each based on the transformers, whilst the third patch using ~1440 questions took around 20 hours each.

4.3. Evaluation

In natural language processing (NLP), confidence metric is a score that measures the level of certainty or probability of a model to accurately predict or classify the correct label or outcome of a given text sample. It is essentially a way to measure the quality of the predicted result. The confidence metric is often expressed as a value between 0 and 1. A higher confidence score indicates that the model is more certain about its prediction, while a lower score indicates greater uncertainty. The confidence metric is used for answer prediction, where the accuracy of the model's prediction is important. There are several mathematical equations used to compute the confidence in QA research, depending on the specific approach and model used. In this research, we used equation called the confidence score:

$$Conf = P(answer \mid question, context) * P(question \mid context) / P(answer)$$

where $P(answer \mid question, context)$ is the probability of the answer given the question and context, $P(question \mid context)$ is the probability of the question given the context, $P(answer)$ is the prior probability of the answer. This equation calculates the probability that a given answer is correct, considering both the likelihood of the answer given the question and context, and the frequency of the question and answer in the corpus collection. Another metric that is often used is the similarity, which evaluates the semantic similarity between two texts of the question and the answer. These metrics are used to assess the relevance and correctness of the answers generated by the chatbot and how well an answer captures the relevant information from the given question, which was calculated in this study as follows:

$$Sim = (question \cdot answer.T)/\big(norm(question) * norm(answer)\big)$$

wherein the vectors representing the *question* and the *answer* to be compared were used, T is the transpose of the *answer* vector, and *norm* is value of matrix norm computed in *NumPy Python*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

5. Experimental Results

5.1. Initial results using a sample of selected questions from MSA-QA and DA-QA datasets

For initial investigation, selected questions were tested from modern standard Arabic and dialectal (i.e., informal) Arabic using the proposed transformers. Table 5 and Table 6 show the experimental results obtained from 10 questions chosen from MSA-QA and DA-QA datasets, respectively. In both tables, we fine-tuned one of the AraBERT transformers (*bert-base-arabert*), and AraElectra-SQuAD transformer using two sizes of semantic embedding models (*bert-base-arabert* and *bert-large-arabert*v2). For each given question shown in the tables, we extracted the answer and then evaluated the similarity and confidence of each model in giving such an answer. The confidences score exceeded 0.8 are presented in bold. While the semantic similarity scores are mostly high, the confidences indicate that the fine-tuned models perform well with some questions and not with the others. To see the big picture of these results, Figure 4 visualizes the confidence scores and Figure 5 visualizes the similarity scores obtained by AraBERT and AraElectra-SQuAD transformers using the selected questions from MSA-QA and DA-QA datasets. As can be seen, the similarities do not indicate differences between the models because there was relevant information for each given question in the texts the models were fine-tuned on. We found a little drop in the similarity scores when we use AraElectra-SQuAD with the *large* semantic embedding model. On the other hand, we found that confidence metrics indicate the differences in the models' performance very well.

Figure 6 shows a better look into the results of confidence scores obtained by AraElectra-SQuAD transformer using the selected questions from MSA-QA indicating the formal context, and the questions from DA-QA indicating the informal context of Arabic. We found that AraElectra-SQuAD transformer performs high with the formal contexts but obtained lower scores with the informal questions. Another notice is that the confidences of the model do not change when changing the semantic embedding models, and the only change happened in the similarity results.

Table 5. Experimental results of 10 selected questions from modern standard Arabic (MSA-QA) dataset.

Question	AraBERT				AraElectra-SQuAD			
	Base		*Large*		*Base*		*Large*	
	Sim.	**Conf.**	**Sim.**	**Conf.**	**Sim.**	**Conf.**	**Sim.**	**Conf.**
1- (' علم البيانات ؟ ')	0.8890	0.3355	0.9284	0.1700	0.9191	0.9993	0.6480	0.9993
2- ('تعريف البيانات العلمية؟')	0.9447	0.1249	0.9486	0.1867	0.9652	0.8601	0.6321	0.8601
3- (' اذكر التباين بين علم البيانات والذكاء الاصطناعي؟ ')	0.9980	1.0	0.9291	0.1234	0.9801	0.5006	0.7498	0.5006
4- (' أذكر لغات جداول الأعمال الأكثر شيوعًا في مجال علم البيانات؟ ')	0.9905	0.5	0.9531	0.1741	0.9778	0.9276	0.8334	0.9276
5- ('يعرف علم البيانات؟')	0.9424	0.25	0.9384	0.2627	0.9343	0.9951	0.6901	0.9951
6- (' اذكر مجالات العمل التي يشارك فيها علم البيانات؟ ')	0.9585	0.1250	0.9895	0.5	0.9747	0.5020	0.8945	0.5020
7- (' أذكرلي لغات البرمجة الأكثر شيوعًا في مجال علم البيانات؟ ')	0.9908	0.1702	0.9887	0.1184	0.9847	0.2424	0.8175	0.2424
8-(' أذكر الاختلافات بين علم البيانات والذكاء الاصطناعي؟ ')	0.9670	1.0	0.9419	0.5476	0.9817	0.4718	0.8070	0.4718
9- (' أذكر الفرق بين علم البيانات والمثقفين الاصطناعي؟ ')	0.9851	0.5102	0.9453	0.2202	0.9831	0.9954	0.8256	0.9954
10- (' اذكر مجالات العمل التي تدخل فيها البيانات علميا؟ ')	0.9470	0.5	0.9677	0.1448	0.9686	0.3397	0.8673	0.3397

Table 6. Experimental results of 10 selected questions from dialectal Arabic (DA-QA) dataset.

Question	AraBERT				AraElectra-SQuAD			
	Base		**Large**		**Base**		**Large**	
	Sim.	**Conf.**	**Sim.**	**Conf.**	**Sim.**	**Conf.**	**Sim.**	**Conf.**
1- ('هوا انا ينفع اقول اني بحب ربنا اووي عشان هوا عسل وبيحبنا؟')	0.9825	0.2331	0.9366	0.1322	0.9729	0.4737	0.8815	0.4737
2- ('بالله عليك اجبني هل يجوز للحائض زيارة المقابر ضروري بالله عليك؟')	0.9801	0.1713	0.9848	0.1724	0.9710	0.8436	0.7874	0.8436

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

3- (‘ ؟التنورة وفوقها بلوزة طويلة مع طرحة تغطي الصدر كده حجاب شرعي’)	0.9902	0.4014	0.9603	0.2438	0.9767	0.9217	0.8508	0.9217
4- (‘؟ هل يجوز أن أقبل يد اخى الأكبر كشكر وعرفان لفضله عليا منذ صغرى’)	0.9809	0.2571	0.9936	0.3674	0.9827	0.4069	0.9191	0.4069
5- (‘؟؟ هل في سنه عن النبي اننا لما نمسح الارض بالمياه نحط عليها ملح’)	0.9858	0.1859	0.975	0.1683	0.9762	0.4853	0.8186	0.4853
6 - (‘ ؟هل ترى تقسيم البدعة الى حسنة وسيئة’)	0.9842	0.3012	0.9371	0.2435	0.9675	0.4547	0.6798	0.4547
7- (‘ ماما بتسأل حضرتك يا شيخ لو هي شارية ليا حاجات للمستقبل أدوات منزلية وغيره هل عليها زكاة أم لا ؟’)	0.986	0.3289	0.9763	0.1594	0.9838	0.9624	0.8619	0.9624
8-(‘ ؟طب يا شيخ بالنسبة لاعياد الميلاد’)	0.9846	0.3816	0.946	0.1539	0.9584	0.3724	0.7328	0.3724
9- (‘ ؟ يا شيخنا هل صلاة المسجد للرجال فرض وتاركه آثم غير مقبول صلاته’)	0.9408	0.2420	0.9839	0.1719	0.9764	0.8567	0.8407	0.8567
10- (‘ ؟ هل يمكن لتوبه صادقه ان تمحو ما قبلها فكأنما ما أذنب المرء قط’)	0.9469	0.4698	0.9851	0.1635	0.9792	0.5950	0.8337	0.5950

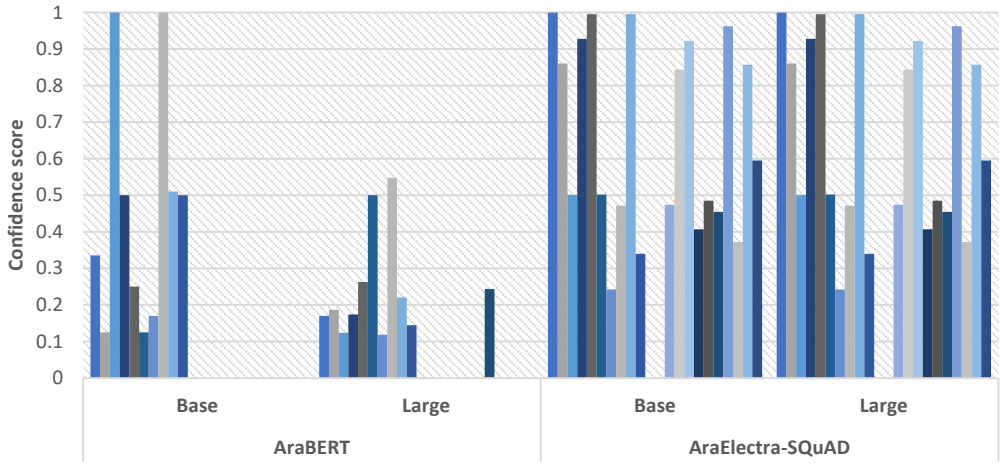
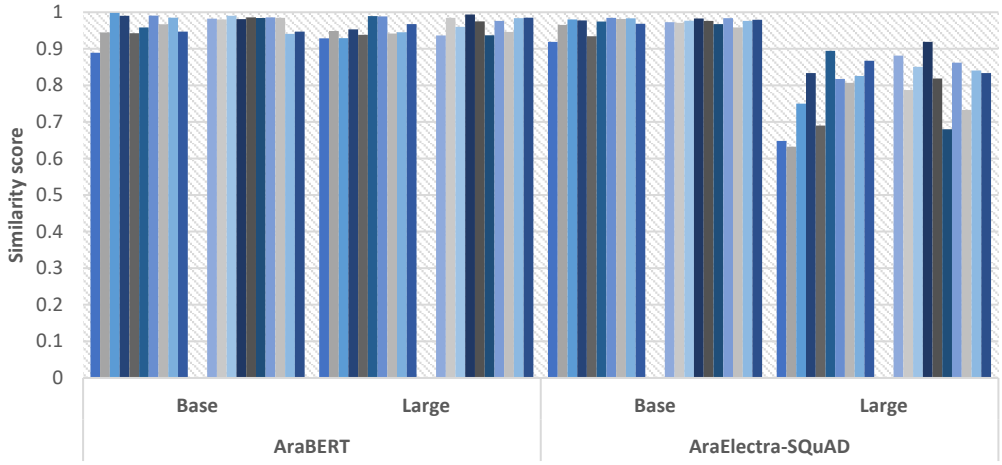


Figure 4. Confidence scores obtained by AraBERT and AraElectra-SQuAD transformers using base and large sizes of embedding models on a sample of selected questions from MSA-QA and DA-QA datasets.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

Figure 5. Similarity scores obtained by AraBERT and AraElectra-SQuAD transformers using base and large sizes of embedding models on a sample of selected questions from MSA-QA and DA-QA datasets.

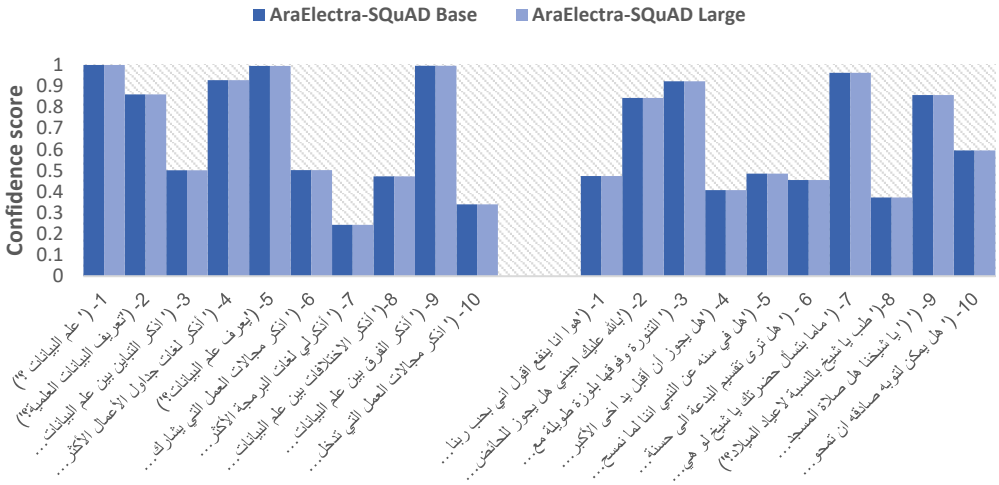


Figure 6. Comparison between confidence scores obtained by formal and informal questions using AraElectra-SQuAD transformer.

5.2. Initial results using a sample of selected questions from MSA-QA and ARCD-QA datasets

To extend our experiments, we aimed in this experimental set to compare MSA-QA and ARCD-QA using a sample of questions generated randomly whereby we investigate the answers by the proposed transformers, and their confidence and similarity scores. In Table 7, we used ten variants of AraBERT-based models with each dataset sample. For MSA-QA, the top-performing model was base-arabic-camelbert-da with an average confidence of 0.4612 and an average similarity of 0.8168. Following closely was the bert-large-arabertv02 transformer resulting in an average confidence of 0.4504and an average similarity of 0.6657. For ARCD-QA, the top-performing model was bert-base-arabert and bert-large-arabertv02 with 0.48 confidence scores. In Table 8, multiple AraElectra-based models were utilized in the experiment, including AraElectra-SQuAD and AarElectra generator and discriminator. We found that AraElectra-Arabic-SQuADv2-QA with the distilbert-base-uncased and bert-base-arabertv2 embedding models outperformed other models with both datasets. Confidences scores obtained in this set of experimental works are visualized in Figure 7 for MSA-QA sample, and Figure 8 for ARCD-QA sample.

Table 7. Experimental results of a sample of generated questions from MSA-QA and ARCD-QA datasets using AraBERT-based transformers.

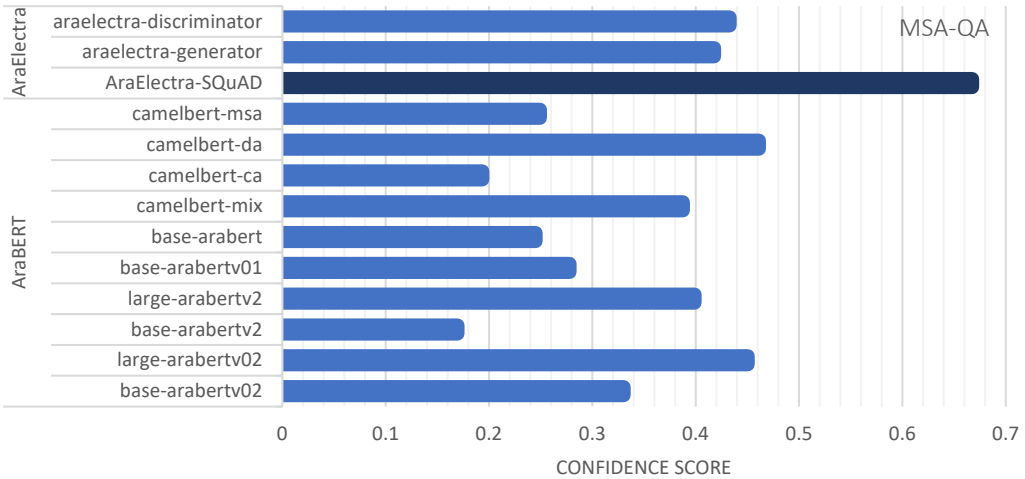
Dataset	Transformer	Semantic Embeddings Model	Avg. Sim.	Avg. Conf.
MSA-QA	bert-base-arabertv02	bert-base-arabertv02	0.8457	0.3304
	bert-large-arabertv02	bert-large-arabertv02	0.6657	0.4504
	bert-base-arabertv2	bert-base-arabertv2	0.8915	0.1695
	bert-large-arabertv2	bert-large-arabertv2	0.7727	0.3989
	bert-base-arabertv01	bert-base-arabertv01	0.8183	0.2779
	bert-base-arabert	bert-base-arabert	0.7776	0.2452
	bert-base-arabic-camelbert-mix	bert-base-arabic-camelbert-mix	0.47667	0.3876
	bert-base-arabic-camelbert-ca	bert-base-arabic-camelbert-ca	0.9625	0.1935
	bert-base-arabic-camelbert-da	bert-base-arabic-camelbert-da	0.8168	0.4612

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

	bert-base-arabic-camelbert-msa	bert-base-arabic-camelbert-msa	0.5394	0.2493
ARCD-QA	bert-base-arabertv02	bert-base-arabertv02	0.7599	0.3320
	bert-large-arabertv02	bert-large-arabertv02	0.6774	0.4816
	bert-base-arabertv2	bert-base-arabertv2	0.6491	0.1822
	bert-large-arabertv2	bert-large-arabertv2	0.6519	0.2913
	bert-base-arabertv01	bert-base-arabertv01	0.8635	0.2271
	bert-base-arabert	bert-base-arabert	0.8507	0.4800
	bert-base-arabic-camelbert-mix	bert-base-arabic-camelbert-mix	0.9122	0.2598
	bert-base-arabic-camelbert-ca	bert-base-arabic-camelbert-ca	0.9352	0.2972
	bert-base-arabic-camelbert-da	bert-base-arabic-camelbert-da	0.8664	0.2937
	bert-base-arabic-camelbert-msa	bert-base-arabic-camelbert-msa	0.7378	0.3381

Table 8. Experimental results of a sample of generated questions from MSA-QA and ARCD-QA datasets using AraElectra-based transformers.

Dataset	Transformer	Semantic Embeddings Model	Avg. Sim.	Avg. Conf.
MSA-QA	AraElectra-Arabic-SQuADv2-QA	bert-base-arabertv2	0.8242	0.6675
	AraElectra-Arabic-SQuADv2-QA	distilbert-base-uncased	0.9786	0.6675
	araelectra-base-generator	bert-base-arabertv2	0.6434	0.4179
	araelectra-base-discriminator	bert-base-arabertv2	0.7652	0.4329
	araelectra-base-generator	distilbert-base-uncased	0.9687	0.3043
	araelectra-base-discriminator	distilbert-base-uncased	0.5688	0.4286
ARCD-QA	AraElectra-Arabic-SQuADv2-QA	bert-base-arabertv2	0.6952	0.6116
	AraElectra-Arabic-SQuADv2-QA	distilbert-base-uncased	0.9806	0.6116
	araelectra-base-generator	bert-base-arabertv2	0.7385	0.1957
	araelectra-base-discriminator	bert-base-arabertv2	0.7388	0.2086
	araelectra-base-generator	distilbert-base-uncased	0.9166	0.3206
	araelectra-base-discriminator	distilbert-base-uncased	0.8962	0.4593



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

Figure 7. Comparison between confidence scores obtained by AraBERT and AraElectra based transformers on a sample of selected questions from MSA-QA dataset.

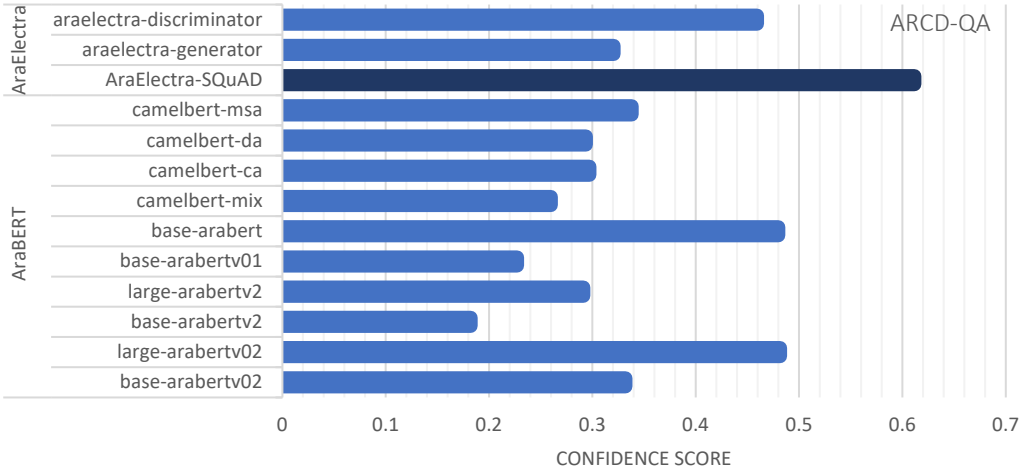


Figure 8. Comparison between confidence scores obtained by AraBERT and AraElectra based transformers on a sample of selected questions from ARCD-QA dataset.

5.3. Experimental results using all questions from MSA-QA and ARCD-QA datasets

In the final set of our experimental works, we used the whole dataset questions and indexed collections specified previously in Table 4. A number of models were selected based on our experiments conducted in the previous parts as described in section 5.1 and 5.2. These models were selected to be evaluated based on their confidence results and robustness in their performance. Table **9** and Figure 9 show our Arabic chatbot evaluation results using 398 questions in MSA-QA with their collected corpus, and 1395 questions ARCD-QA with 365,568 indexed Wikipedia documents. Significant findings include that AraElectra-based fine-tuned models outperformed AraBERT-based fine-tuned models, and AraElectra-SQuAD model achieved the highest performing model with all datasets. Its best confidence and similarity scores were achieved when using *distilbert* semantic embeddings and the results were 0.9660 similarity and 0.6658 confidence. Therefore, AraElectra-SQuAD model can be further enhanced in fine-tuning and in practice in various natural language processing tasks, such as chatbots, virtual assistants, and information retrieval systems, for Arabic-speaking users. By combining the power of the transformer architecture with the fine-tuning on SQuAD, AraElectra-SQuAD was able to provide accurate and contextually relevant answers to questions in Arabic.

Table 9. Arabic chatbot evaluation results using all questions in MSA-QA, and ARCD-QA with Wikipedia dump collection.

	Dataset	Transformer	Semantic Embeddings Model	Avg. Sim.	Avg. Conf.
AraBERT-based	MSA-QA	bert-base-arabertv02	bert-base-arabertv02	0.8256	0.3897
		bert-large-arabertv02	bert-large-arabertv02	0.8365	0.2128
		bert-large-arabertv2	bert-large-arabertv2	0.7673	0.4251
		bert-base-arabic-camelbert-da	bert-base-arabic-camelbert-da	0.9229	0.3634
	ARCD-QA	bert-base-arabertv02	bert-base-arabertv02	0.6986	0.2038
		bert-large-arabertv02	bert-large-arabertv02	0.6241	0.5465
		bert-base-arabert	bert-base-arabert	0.9396	0.2426

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2023 doi:10.20944/preprints202307.0609.v1

		bert-base-arabic-camelbert-msa	bert-base-arabic-camelbert-msa	0.7727	0.1901
AraElectra-based	MSA-QA	AraElectra-Arabic-SQuADv2-QA	bert-base-arabertv2	0.8268	0.6422
		AraElectra-Arabic-SQuADv2-QA	distilbert-base-uncased	0.9773	0.6422
		araelectra-base-generator	bert-base-arabertv2	0.7013	0.3616
		araelectra-base-discriminator	bert-base-arabertv2	0.7218	0.3291
	ARCD-QA	AraElectra-Arabic-SQuADv2-QA	bert-base-arabertv2	0.6852	0.6657
		AraElectra-Arabic-SQuADv2-QA	distilbert-base-uncased	0.9660	0.6658
		araelectra-base-generator	distilbert-base-uncased	0.9036	0.2908
		araelectra-base-discriminator	distilbert-base-uncased	0.8573	0.4147

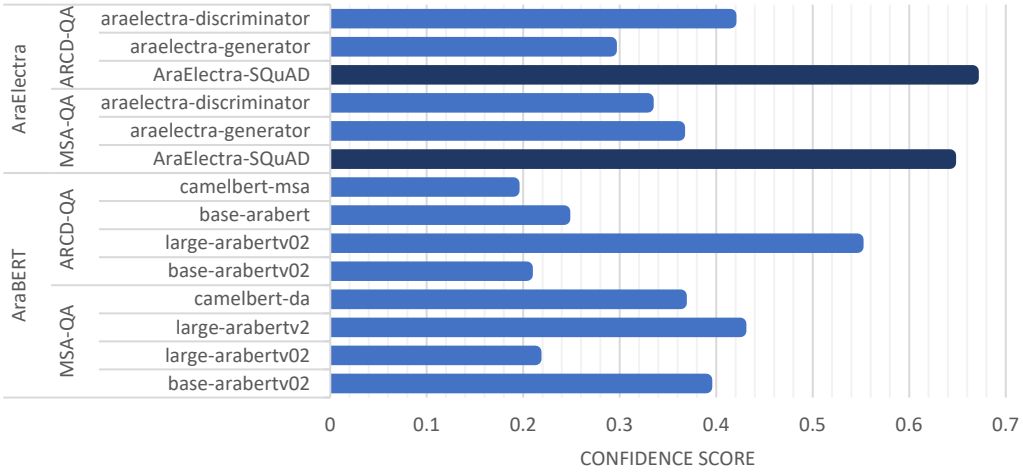


Figure 9. Comparison between confidence scores obtained by AraBERT and AraElectra-SQuAD transformers on a sample of selected questions from ARCD-QA dataset.

6. Conclusion and future works

Chatbots are AI-based programs designed to imitate human dialogues. This paper addresses the problem of Arabic chatbots because they are rare and less known than English chatbots due to the nature and intricacy of the Arabic language. This work shed light on a comprehensive review of previously published studies that applied to chatbots using extractive and generative models, machine learning, deep learning, and the current trends of transfer learning. Using pre-trained models and transfer learning overcome the problems of limited data availability and allow the generalization of Arabic language conversational dialogs to understand conversation context and behave naturally. Different Arabic QA datasets were utilized to investigate the use transfer learning techniques using ten AraBERT-based and CAMeLBERT transformers, as well as six AraElectra-SQuAD and AarElectra generator and discriminator transformers. We evaluated different variants of these transformers and semantic embedding models using 398 questions with corresponding documents, and 1,395 questions and 365,568 documents indexed from Arabic Wikipedia. Through extensive experimentation, we observed that AraElectra-based fine-tuned models yielded promising results with both datasets. The AraElectra-Arabic-SQuADv2-QA model consistently demonstrated the top performance of 0.66 confidence and 0.96 similarity scores. For future work, it would be

beneficial to explore additional datasets to validate the findings and assess the models' performance in diverse contexts and improve the results in terms of confidence. Additionally, evaluating the models' performance on other language-related tasks or expanding the research to include multilingual question answering would be valuable directions for future studies.

Author Contributions: "Conceptualization, T.N. and S.M; methodology, T.N.; software, T.N.; validation, S.M.; formal analysis, S.M.; investigation, T.N.; resources, T.N.; writing—original draft preparation, S.M.; writing—review and editing, S.M.; visualization, S.M.; supervision S.M. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The indexed dataset is available on this link: bit.ly/3pubIc9.

Conflicts of 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.

Acknowledgments: The researchers would like to acknowledge Deanship of Scientific Research, Taif University for funding this work.

References

1. Caldarini, G.; Jaf, S.; McGarry, K.; McGarry, K. A Literature Survey of Recent Advances in Chatbots. **2022**, *2022*, 41-41, doi:10.3390/info13010041.
2. Ali, D.A.; Habash, N. Botta: An Arabic Dialect Chatbot. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, 2016; pp. 208-212.
3. Al-Ghadhban, D.; Al-Twairesh, N. *Nabiha: An Arabic Dialect Chatbot*; 2020.
4. Joukhadar, A.; Saghergy, H.; Kweider, L.; Ghneim, N. Arabic dialogue act recognition for textual chatbot systems. In Proceedings of the Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers, 2019; pp. 43-49.
5. Shi, N.; Zeng, Q.; Lee, R. Language Chatbot-The Design and Implementation of English Language Transfer Learning Agent Apps. In Proceedings of the 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering, AUTEEE 2020, 2020/11//, 2020; pp. 403-407.
6. Vasilev, I.; Slater, D.; Spacagna, G.; Roelants, P.; Zocca, V. *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow, 2nd Edition*; Packt Publishing: 2019.
7. Alhassan, N.A.; Saad Albarrak, A.; Bhatia, S.; Agarwal, P. A Novel Framework for Arabic Dialect Chatbot Using Machine Learning. *Computational Intelligence and Neuroscience* **2022**, *2022*, doi:10.1155/2022/1844051.
8. Alruily, M. ArRASA: Channel Optimization for Deep Learning-Based Arabic NLU Chatbot Framework. *Electronics (Switzerland)* **2022**, *11*, doi:10.3390/electronics11223745.
9. Ghaddar, A.; Wu, Y.; Bagga, S.; Rashid, A.; Bibi, K.; Rezagholizadeh, M.; Xing, C.; Wang, Y.; Xinyu, D.; Wang, Z.; et al. Revisiting Pre-trained Language Models and their Evaluation for Arabic Natural Language Understanding. **2022**, arXiv:2205.10687, doi:10.48550/arXiv.2205.10687.
10. Suta, P.; Lan, X.; Wu, B.; Mongkolnam, P.; Chan, J.H. An overview of machine learning in chatbots. *International Journal of Mechanical Engineering and Robotics Research* **2020**, *9*, 502-510, doi:10.18178/ijmerr.9.4.502-510.
11. Adamopoulou, E.; Moussiades, L. Chatbots: History, technology, and applications. *Machine Learning with Applications* **2020**, *2*, 100006-100006, doi:10.1016/j.mlwa.2020.100006.
12. Ali, A.; Zain Amin, M. Conversational AI Chatbot Based on Encoder-Decoder Architectures with Attention Mechanism Application of Multilayer Perceptron (MLP) for Data Mining in Healthcare Operations View project Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions View project Conversational AI Chatbot Based on Encoder-Decoder Architectures with Attention Mechanism. *Artificial Intelligence Festival* **2019**, *2*, doi:10.13140/RG.2.2.12710.27204.
13. Majid, R.; Santoso, H.A. Conversations Sentiment and Intent Categorization Using Context RNN for Emotion Recognition. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, 2021/3//, 2021; pp. 46-50.
14. Ilievski, V.; Musat, C.; Hossmann, A.; Baeriswyl, M. Goal-Oriented chatbot dialog management bootstrapping with transfer learning. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, 2018; pp. 4115-4121.

15. Nguyen, T.T.; Le, A.D.; Hoang, H.T.; Nguyen, T. NEU-chatbot: Chatbot for admission of National Economics University. *Computers and Education: Artificial Intelligence* **2021**, *2*, doi:10.1016/j.caeai.2021.100036.
16. Moriuchi, E.; Landers, V.M.; Colton, D.; Hair, N. Engagement with chatbots versus augmented reality interactive technology in e-commerce. *Journal of Strategic Marketing* **2021**, *29*, 375-389, doi:10.1080/0965254X.2020.1740766.
17. Siglen, E.; Vetti, H.H.; Lunde, A.B.F.; Hatlebrekke, T.A.; Strømsvik, N.; Hamang, A.; Hovland, S.T.; Rettberg, J.W.; Steen, V.M.; Bjorvatn, C. Ask Rosa – The making of a digital genetic conversation tool, a chatbot, about hereditary breast and ovarian cancer. *Patient Education and Counseling* **2022**, *105*, 1488-1494, doi:10.1016/j.pec.2021.09.027.
18. Khadija, A.; Zahra, F.F.; Naceur, A. AI-Powered Health Chatbots: Toward a general architecture. In Proceedings of the Procedia Computer Science, 2021; pp. 355-360.
19. Baha, T.A.I.T.; Hajji, M.E.L.; Es-Saady, Y.; Fadili, H. Towards highly adaptive Edu-Chatbot. In Proceedings of the Procedia Computer Science, 2021; pp. 397-403.
20. K., H.K.; Palakurthi, A.K.; Putnala, V.; K., A.K. Smart College Chatbot using ML and Python. In Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 3-4 July 2020, 2020; pp. 1-5.
21. Vyawahare, S.; Chakradeo, K. Chatbot assistant for english as a second language learners. In Proceedings of the 2020 International Conference on Convergence to Digital World - Quo Vadis, ICCDW 2020, 2020/2//, 2020.
22. Gowda, M.P.C.; Srivastava, A.; Chakraborty, S.; Ghosh, A.; Raj, H. Development of Information Technology Telecom Chatbot: An Artificial Intelligence and Machine Learning Approach. In Proceedings of the Proceedings of 2021 2nd International Conference on Intelligent Engineering and Management, ICIEM 2021, 2021/4//, 2021; pp. 216-221.
23. Mathew, R.B.; Varghese, S.; Joy, S.E.; Alex, S.S. Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 23-25 April 2019, 2019; pp. 851-856.
24. Wang, R.; Wang, J.; Liao, Y.; Wang, J. Supervised machine learning chatbots for perinatal mental healthcare. In Proceedings of the Proceedings - 2020 International Conference on Intelligent Computing and Human-Computer Interaction, ICHCI 2020, 2020/12//, 2020; pp. 378-383.
25. Rakib, A.B.; Rumky, E.A.; Ashraf, A.J.; Hillas, M.M.; Rahman, M.A. Mental Healthcare Chatbot Using Sequence-to-Sequence Learning and BiLSTM. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021; pp. 378-387.
26. Chempavathy, B.; Prabhu, S.N.; Varshitha, D.R.; Vinita; Lokeswari, Y. AI based Chatbots using Deep Neural Networks in Education. In Proceedings of the Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2022, 2022; pp. 124-130.
27. Almurayh, A. The Challenges of Using Arabic Chatbot in Saudi Universities. *IAENG International Journal of Computer Science* **2021**, *48*.
28. Zahour, O.; Benlahmar, E.H.; Eddaoui, A.; Ouchra, H.; Hourrane, O. A system for educational and vocational guidance in Morocco: Chatbot e-orientation. In Proceedings of the Procedia Computer Science, 2020; pp. 554-559.
29. Thorat, S.A.; Jadhav, V. A Review on Implementation Issues of Rule-based Chatbot Systems. *Social Science Research Network* **2020**.
30. Singh, J.; Joesph, M.H.; Jabbar, K.B.A. Rule-based chabot for student enquiries. In Proceedings of the Journal of Physics: Conference Series, 2019/6//, 2019.
31. Maeng, W.; Lee, J. Designing a Chatbot for Survivors of Sexual Violence: Exploratory Study for Hybrid Approach Combining Rule-based Chatbot and ML-based Chatbot. In Proceedings of the 5th Asian CHI Symposium 2021, 2021/5//, 2021; pp. 160-166.
32. Alsheddi, A.S.; Alhenaki, L.S. English and Arabic Chatbots: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications* **2022**, *13*, 662-675, doi:10.14569/IJACSA.2022.0130876.
33. Rokaya, A.; Md Touhidul Islam, S.; Zhang, H.; Sun, L.; Zhu, M.; Zhao, L. Acceptance of Chatbot based on Emotional Intelligence through Machine Learning Algorithm. In Proceedings of the Proceedings - 2022 2nd

International Conference on Frontiers of Electronics, Information and Computation Technologies, ICFEICT 2022, 2022; pp. 610-616.

34. Achuthan, S.; Balaji, S.; Thanush, B.; Reshma, R. An Improved Chatbot for Medical Assistance using Machine Learning. In Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2022 - Proceedings, 2022; pp. 70-75.
35. Goel, R.; Arora, D.K.; Kumar, V.; Mittal, M. A Machine Learning based Medical Chatbot for detecting diseases. In Proceedings of the Proceedings of 2nd International Conference on Innovative Practices in Technology and Management, ICIPTM 2022, 2022; pp. 175-181.
36. Goel, R.; Goswami, R.P.; Totlani, S.; Arora, P.; Bansal, R.; Vij, D. Machine Learning Based Healthcare Chatbot. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, 2022; pp. 188-192.
37. Mahanan, W.; Thanyaphongphat, J.; Sawadsitang, S.; Sangamuang, S. College Agent: The Machine Learning Chatbot for College Tasks. In Proceedings of the 7th International Conference on Digital Arts, Media and Technology, DAMT 2022 and 5th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, NCON 2022, 2022; pp. 329-332.
38. Prasetyo, A.; Santoso, H.A. Intents Categorization for Chatbot Development Using Recurrent Neural Network (RNN) Learning. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, 2021/3//, 2021; pp. 551-556.
39. Dhyani, M.; Kumar, R. An intelligent Chatbot using deep learning with Bidirectional RNN and attention model. In Proceedings of the Materials Today: Proceedings, 2019; pp. 817-824.
40. Patil, S.; Mudaliar, V.M.; Kamat, P.; Gite, S. LSTM based Ensemble Network to enhance the learning of long-term dependencies in chatbot. *Int. J. Simul. Multidisci. Des. Optim.* **2020**, *11*, 25.
41. Pathak, K.; Arya, A. A Metaphorical Study of Variants of Recurrent Neural Network Models for A Context Learning Chatbot. In Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, 2019/11//, 2019; pp. 768-772.
42. Kasthuri, E.; Balaji, S. Natural language processing and deep learning chatbot using long short term memory algorithm. *Materials Today: Proceedings* **2021**, doi:10.1016/j.matpr.2021.04.154.
43. Anki, P.; Bustamam, A.; Al-Ash, H.S.; Sarwinda, D. High Accuracy Conversational AI Chatbot Using Deep Recurrent Neural Networks Based on BiLSTM Model. In Proceedings of the 2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020, 2020/11//, 2020; pp. 382-387.
44. Jalaja, T.; Adilakshmi, D.T.; Sharat Chandra, M.S.; Imran Mirza, M.; Kumar, M. A Behavioral Chatbot Using Encoder-Decoder Architecture : Humanizing conversations. 2022/11//, 2022; pp. 51-54.
45. Boussakssou, M.; Ezzikouri, H.; Erritali, M. Chatbot in Arabic language using seq to seq model. *Multimedia Tools and Applications* **2022**, *81*, 2859-2871, doi:10.1007/s11042-021-11709-y.
46. Rajamalli Keerthana, R.; Fathima, G.; Florence, L. Evaluating the performance of various deep reinforcement learning algorithms for a conversational chatbot. In Proceedings of the 2021 2nd International Conference for Emerging Technology, INCET 2021, 2021/5//, 2021.
47. Cuayáhuitl, H.; Lee, D.; Ryu, S.; Cho, Y.; Choi, S.; Indurthi, S.; Yu, S.; Choi, H.; Hwang, I.; Kim, J. Ensemble-based deep reinforcement learning for chatbots. *Neurocomputing* **2019**, *366*, 118-130, doi:10.1016/j.neucom.2019.08.007.
48. Kulkarni, A.; Shivananda, A.; Kulkarni, A. Building a Chatbot Using Transfer Learning. In *Natural Language Processing Projects : Build Next-Generation NLP Applications Using AI Techniques*; Apress: Berkeley, CA, 2022; pp. 239-255.
49. Aksu, I.T.; Chen, N.F.; D'Haro, L.F.; Banchs, R.E. Reranking of Responses Using Transfer Learning for a Retrieval-Based Chatbot. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Marchi, E., Siniscalchi, S.M., Cumani, S., Salerno, V.M., Li, H., Eds.; Springer Singapore: Singapore, 2021; pp. 239-250.
50. Vijayaraghavan, V.; Cooper, J.B.; Rian Leevinson, R.L. Algorithm Inspection for Chatbot Performance Evaluation. In Proceedings of the Procedia Computer Science, 2020; pp. 2267-2274.
51. Ahmed, A.; Ali, N.; Alzubaidi, M.; Zaghouani, W.; Abd-alrazaq, A.; Househ, M. Arabic chatbot technologies: A scoping review. *Computer Methods and Programs in Biomedicine Update* **2022**, *2*, 100057-100057, doi:10.1016/j.cmpbup.2022.100057.
52. Wolf, T.; Sanh, V.; Chaumond, J.; Delangue, C. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. **2019**.

 doi:10.20944/preprints202307.0609.v1

53. Aljawarneh, E. Arabic Questions Dataset. Available online: https://github.com/EmranAljawarneh/Arabic-questions-dataset (accessed on 9 Feb. 2023).
54. Mozannar, H.; Hajal, K.E.; Maamary, E.; Hajj, H. Neural Arabic Question Answering. **2019**.