English and Arabic Chatbots: A Systematic Literature Review

Abeer S. Alsheddi
Computer Science Department
Imam Muhammad bin Saud Islamic University
King Saud University, Riyadh, Saudi Arabia

Lubna S. Alhenaki
Department of Computer Science, College of Computer and Information Sciences, Majmaah University
Al-Majmaah, 11952, Saudi Arabia

***Abstract*—In recent years, the availability of chatbot applications has increased substantially with the advancement of artificial intelligence techniques, and research efforts have been active in the English language, which presents state-of-the-art solutions. However, despite the popularity of the Arabic language, its research community is still in an immature stage. Therefore, the main objective of this systematic literature review is studying state-of-the-art research – for both the English and Arabic languages – to answer the proposed research questions regarding the development approaches, application domains, evaluation metrics, and development challenges of chatbot applications. The findings show that researchers have devoted more attention to the education domain using retrieval-based approaches while the generation-based approach has grown in popularity recently for providing new responses tasks. Whereas the hybrid approach for ranking multi-possible responses of combining both previous approaches shows a performance improvement. Besides, most metrics used to evaluate chatbot performance are human-based, followed by bilingual evaluation understudy and accuracy metrics. However, defining a common framework for evaluating chatbots remains a challenge. Finally, the open problems and future directions are highlighted to help in developing chatbots with minimal human interference to simulate natural conversations.**

***Keywords*—Chatbots; Arabic language; development approaches; domain applications; evaluation metrics**

I. Introduction

A chatbot is an example of a computer application based on artificial intelligence (AI) that aims to simulate human behavior by conducting a conversation with users using natural language data. The most well-known social applications, such as Telegram and Facebook Messenger, are supported by chatbots. Several organizational benefits of using chatbots include 24-hour availability, endless patience, increased sales, and reduced operational costs [1]. These benefits have led to an increasing demand for the development of chatbots using AI techniques. However, developing effective chatbots that can respond at the level of an actual human is challenging due to the requirement of understanding user inputs, generating appropriate responses, and perceiving the context of the conversation [2].

Over the past decade, a rapid development of chatbots based on English language has taken place in many application domains. In last two years, several *surveys* have been published about chatbots, mostly focused on the implementation approaches, such as those [3],[4],[5]. However , some of the previous research have been limited to specific domains, those researches reviewing the techniques, characteristics, and approaches used in the development used in the development of an intelligent tutoring chatbot applied to education [6], [7]. Moreover, the research in [8] examined previous articles which showed that the personalized learning framework of chatbots helped students improve in their studies. Although the surveys in [9],[10],[11],[12] were careful investigations, they have different aspects *than* this SLR. For example, the study in [11] used a different database and selection criteria. Also, the studies in [9] and [12] presented different research questions. In addition, the evaluation measures and challenges of implementation are not highlighted in [10].

Although developing a chatbot follows similar approaches regardless of which language is being used, for languages such as Arabic, chatbot implementation is challenged by the language's rich morphology, multiple dialects, and orthographic ambiguity [13][14]. However, according to findings obtained from this SLR, in the past three years, research about Arabic chatbots has substantially increased but still has insufficient resources such as available data sets, pretrained models, and tools [13]. Furthermore, to date, few surveys have been done about Arabic chatbots to identify the techniques, metrics, and data sets used. However, chatbots that existed till 2018 are used to process the data in the survey of [15] and did not address some of the same research questions investigated in this SLR [16]. In addition, the survey in [17] highlights one approach to develop a chatbot instead of covering all of the three approaches that will be covered in this SLR. Also, the study in [18] takes a different perspective on a number of applications involving the chatbot. Finally, the study in [19] investigates the characteristics of Arabic chatbots. Overall, the Arabic chatbot research community is still at an immature stage, so the English studies are included to present state-of-the-art solutions.

The objective of this SLR is to present a general overview of the English and Arabic developed chatbots by deeply investigating the current articles in this field. It addresses the domain applications, and approaches used to develop a chatbot and compare these addressed aspects to both languages, English and Arabic. The challenges and evaluation metrics also will be considered. Furthermore, after discussing the findings, the open research problems and future directions for research will be highlighted. The rest of this paper is organized as follows: Section II briefly overviews the chatbot technology. The methodology that follows in this SLR is provided in

Section III. The discussion of finding, open research problems and future research directions are presented in Section IV. Finally, Section V reports conclusion.

II. Chatbot Approaches and Techniques

A. Chatbot Approaches

The selected articles in this SLR generate their responses using different approaches. These approaches can be categorized based on the response generation into rule-based, corpus-based, and hybrid approaches [20]. Fig. 1 presents the approaches along with their common techniques in the selected articles.

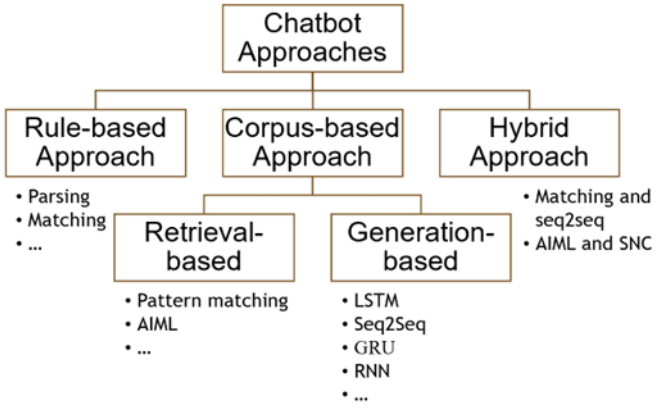

Fig. 1. The General Structure of Chatbot Approaches.

1) Rule-based approach is made up of a set of predefined human-made rules used in the hierarchy to convert user input into an output [20]. The rules break down the input into a sequence of tokens to find a pattern and generate a response rather than generating a new response. Although this approach can be considered easier in terms of implementing, it restricts responses to inputs *within* the predefined rules only and may provide inaccurate responses, leading to an unsatisfactory experience [1].

2) Corpus-based approach uses a knowledge base that contains a statistical language approach to select suitable responses instead of using predefined rules. Most chatbots in this approach produce their responses either two approaches:

- Retrieval-based approach uses information retrieval to get a candidate response from the corpus based on heuristics techniques rather than generating a new response [20]. This technique considers not just input but also the context by identifying keywords to offer the optimal response from a predefined responses (knowledge base) [21].
- Generation-based approach generates the responses based on the dialog context and does not require any predefined response. The chatbot attempts to generate a new response by considering the current and previous user interactions. It may require a large training set, which is potentially tricky to obtain. However, this approach has a high chance of response errors since generating in real-time.

3) Hybrid approach takes advantage of the combined strengths of the generation-based and retrieval-based approaches. Thus, the responses achieve more accurate results by capturing the informative pattern features. The researchers conducted *this* combination by ranking the response from generative and retrieval models or enhancing the informativeness and diversity retrieval responses by feeding them into a generation-based approach [1].

B. Chatbot Techniques

Several techniques are used in this SLR. The following subsections describe these techniques in detail:

1) Parsing. It converts text into meaningful representation string to determine the dependence relationship between its terms or its semantic structure. The parsing technique can be a lexical parsing that converts text into less complicated atomic terms to help extract information and simplify the manipulation. After applying lexical parsing, a syntactical parsing and a semantic parsing can be applied. These two parsing techniques determine a sentence's grammatical structure and extract a specific meaning by converting text to a machine-understandable representation of its meaning [21]. The parsing technique helps chatbots to understand text by identifying the main keywords in it [22]. For example, “set your eyes on my friend” and “could you see my friend” would both generate the same parsed “see my friend”. Moreover, this technique helps to identify the ambiguity in order to ask a user to rephrase his input [23]. For example, two possible ways to interpret the sentence “I saw my friend with my phone”: 1) did my phone help me see my friend; 2) did I see my friend holding my phone.

2) Pattern matching. Chatbots in this technique create a response with patterns where they are made manually, which is a non-trivial process [21]. Although it helps in response time reduction, the responses may be fully predictable and repeated, resulting in dull interactions that lack spontaneity and the human touch [24]. For example, the chatbot recognizes the input “where is the stickers?” and identifies keywords, here is “books”, where each keyword associates with an intent and response, here are “Office supplies” and “Different types of supplies in the office supplies aisle”, respectively.

3) AIML. Chatbots implement a syntax of the pattern matching technique through different technologies like the AIML to retrieve the most suitable response selection [25]. AIML is an open standard language derived from the Extensible Mark-up Language (XML). AIML comprises data objects consisting of two elements which are topics and categories. A topic is an optional top-level element with a set of categories related to that topic, while a category is a rule with a pattern and template matching for input and response, respectively. The objects are sorted in AIML files. Despite its readability, usefulness, and effective use of response time, it must provide a pattern for each conceivable response and update it on a regular basis, which cannot be done automatically [10].

4) RNNs. The RNNs allow the chatbot to handle sequential data and consider the current users' input. Due to the internal short memory, it memorizes the previous users' input. In other words, unlike a traditional Neural *Network*, RNN permits data to remain. The main idea of RNN is saving the output of a particular layer and feeding it as input to the next layer to predict the output [26]. Due to the vanishing or exploding gradient problem [27], the unmodified version of RNN is not appropriate for some applications. LSTM [28] and Gated Recurrent Units (GRU) [29],[30] are different solutions to this problem.

5) LSTM. The LSTM is a special kind of RNN [28]. LSTM is designed to handle long-term *dependencies* and solves the vanishing or exploding gradient problem in RNN. Thus, the gates are introduced in LSTM. Gates are the core component in LSTM, which decides which information that will be memorized. Additionally, the gates output the value between zero and one, where zero means do not memorize anything and one means let everything pass to the next state. Moreover, three kinds of gates are available in LSTM: input gates, forget gates, and output gates to control the flow of information. The input gate is responsible for the state update mechanism while the forget gates decide which information should be memorized. Additionally, the output gate determines the output from the hidden layer. The memory cell in LSTM comprises these three gates. Since the LSTM is created as the solution to short-term memory, it is capable of remembering aspects such as gender. Thus, depending on the previously remembered input, the chatbot can use "his/her". There exists a different architecture of LSTM, such as BiLSTM, which considers the input from the opposite direction as well [31]. Besides, the GRU is the main competitor of the LSTM and RNN [29],[30]. Due to its architecture, it is more popular and less complex than LSTM. The input and forget gate are combined to form a single "update gate".

6) Seq2Seq. Seq2Seq structure is the first architecture proposed to solve translation problems: their success bodes well for NLG. The seq2seq is trained end-to-end using different datasets and domains. Furthermore, due to its flexibility, simplicity, and generality, it is widely used to solve different NLP tasks, which makes seq2seq the industry-standard structure [32]. Technically, seq2seq is composed of two RNNs, namely, an Encoder and a Decoder. The Encoder processes the input of the user word-by-word while the Decoder generates the response word-by-word based on previous conversations. For building chatbots, rather than translating from one language to another, the problem was considered as translating the user input to the chatbot response. Additionally, the length of the input and response sequences can differ, which is one of the advantages of seq2seq structure over others.

III. Survey Methodology

The systematic method used in this SLR for reviewing chatbot articles are based on Kitchenham guideline [33]. This guideline consists of three stages which are planning, conducting, and reporting the review.

A. Planning

This subsection presents the research preparation of the SLR, search procedure and finally the inclusion and exclusion criteria.

1) Goals and research questions. The primary purpose of this SLR is focusing on analyzing the state-of-the-art English and Arabic chatbot articles, especially, with respect to their development approaches, application domains, evaluation metric and the main chatbot's development challenges. To achieve these objectives, four research questions are addressed. Table I presents the research questions and motivation.

TABLE I. List of the Research Questions

Number	***Research Question***	***Motivation***
RQ1	What are the main development approaches used for chatbot with regard to the user's generating appropriate response?	Identifying the state-of-the-art approaches and their techniques may be a significance for developers to provide more fit solutions by working and evolving the recent techniques trends
RQ2	What are the several domains used for building the chatbot?	Identifying the several application domains that most common used may help and enable researchers to address the current focus of domain of application as well as boost research in less contributed domains
RQ3	What are the commonly used metrics to evaluate the chatbots' performance?	Identifying the commonly used metric to evaluate the performance may help to improve and standardized the assess of chatbot, besides mitigate the difficult of comparing the performance of different chatbots
RQ4	What are the main challenges facing the implementation of Chatbot?	Open-research problem in development chatbot and the future directions are provided to continue developing the current issues.

2) Databases identification and search procedure. Six digital databases were *selected* which are: IEEE Xplore Digital Library, Springer, ScienceDirect, Google Scholar, Web of Science (ISI), and ACM. The search was done using 15 keywords as mentioned in [15], [9], [11] and presented in Fig. 2. The search procedure uses 15 keywords belonging to the computer science field in the predefined date ranges. There are 14 keywords for English and additional special keyword for Arabic search "Arabchat". For Arabic search, the same 14 keywords are used proceeding by adding "Arabic" term.

3) Inclusion and exclusion criteria. All criteria are applied manually after searching in the six databases. The inclusion criteria are based to the date and type of articles for both English and Arabic research. The selected articles of research work of the journals based on English chatbots started since 2018 till beginning 2022, period at which our actual SLR research study is done. However,due to the lack of Arabic research, the selected Arabic articles from journals and

conferences are starting from 2004 which is the publication year of one of the earliest Arabic articles till beginning 2022 [15]. Moreover, selected range of the earliest approach, rule-based approach, is extended because lately few recent research studies were utilizing this approach. Four inclusion criteria are: 1) Articles published in both English and Arabic Languages; 2) Articles published in specific ranges as mentioned previously; 3) Full text articles; 4) Articles that addressed the proposed research questions.

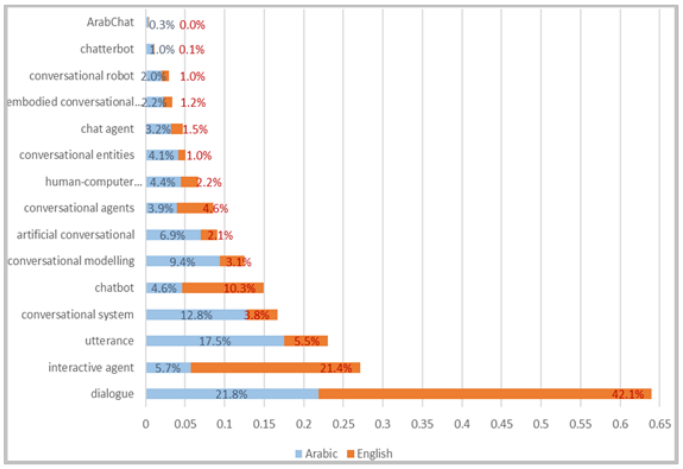

Fig. 2. Total Retrieved Articles for 15 Keywords.

After applying the inclusion criteria, four exclusion criteria are used. The first criterion discards any article that is not related to the chatbot depending on their titles and abstracts. Conspicuously the search in databases sometime returns a huge number of articles that cannot be processed manually. Thus, one ascension in this step is considered which is the returned articles in a database search engine are ordered relevantly to the keywords. Therefore, irrelevant investigated articles are followed by irrelevant articles as well. The second criterion is to remove the duplicate articles that appear in more than one database out of the searched six databases. The third criterion relates to the research questions and involves assessing the candidate articles under the quality assessment as will be discussed in the following subsection. The last criterion keeps journals for both languages and accepts Arabic articles from conferences, while books and theses are filters.

B. Conducting the Review

1) Article selection. The result of 15 keywords searching in all databases returns 59,169 articles. Fig. 2 presents the total number of returned articles for each keyword from all databases in ascending order. The most common keywords is “dialogue” for both English and Arabic chatbot research. Mostly, the greater number of words in each keyword, the smaller number of returned relevant articles, such as the two keywords "dialogue" and "human-computer conversational systems". Moreover, different searching strategies are used in digital databases for retrieving relevant articles. Some of them including in ACM and Google Scholar mostly return a large number of articles exceeding 250K which is caused to restrict the search in them to occur keywords only in the title of articles. Although of that, Google Scholar database still returns the largest number of relevant articles.

2) Data extraction. The result of applying manually exclusion criteria presents in Fig. 3. Firstly, removing irrelevant articles by analyzing their titles and abstracts remained 321 articles that were downloaded. Then 137 duplicated articles were removed. Next, the full texts of the remaining articles were investigated deeply, resulting therefore at filtering 56 articles not addressing the research questions or not passing the quality assessment. The 40 of articles are filtered involving books and theses. As a result, 50 and 38 English and Arabic articles, respectively, are relevant and investigated for the SLR.

3) Quality assessment. The assessment process was conducted in simultaneously with articles extraction. The process was performed for each candidate article individually where the various assessments is discussed until a consensus was achieved. The checklist of ten assessment questions are provided, where a candidate article was selected when at least it gets seven yes answered to the ten questions [34]: 1) Does the article present the objectives of the research clearly? 2) Does the article well-describe the proposed approaches and techniques? 3) Does the article attempt to address an existing issue with chatbot applications? 4) Does the design adhere to well-defined design concepts or principles? 5) Does the article describe the used dataset? 6) Does the article state the results? 7) Does the article state the process of performance evaluation? 8) Dose the article discusses one of domains in the chatbots application? 9) Does the article have a coherent reporting and understandable? 10) Does the article have an appropriate research method appropriate to address the aims of the research?

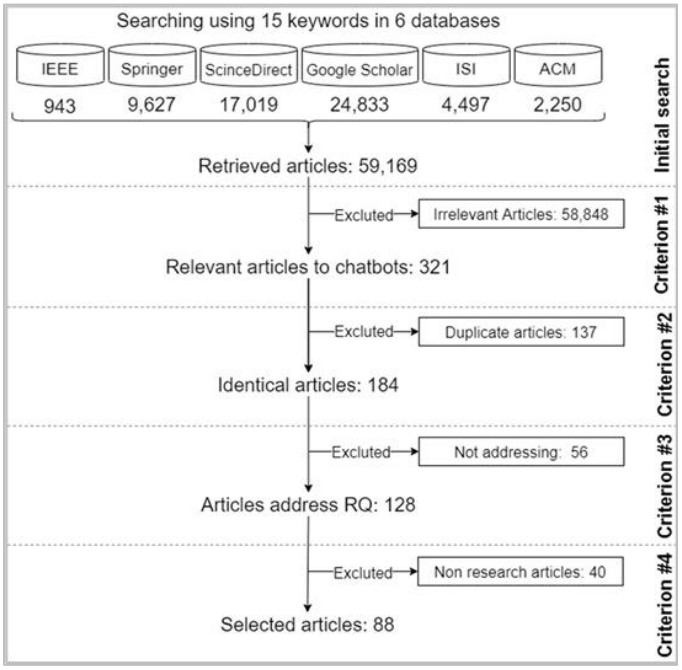

Fig. 3. Data Extraction Step.

C. Reporting the Review

1) Approaches. A detailed comparison of these works is offered in Table II. It is worth noting that no technique is the best selection for all problems. Hence, experimentations in the selected articles are used to compare these techniques. In addition, many chatbots have been developed using state-of-the-art platforms or programming languages for various purposes [5]. Table III presents publicly existing platforms in the selected articles to implement chatbots. The other research works done in [35],[36],[37],[38] develop chatbots using programming languages based on C# in Verbot 5.0, on Snatchbot, Microsoft Visual Studio linked to Google Translate API and the Twilio platform, respectively, as well as some libraries in python are used, such as ChatterBot and chatterbot_corpus.

2) Domains. The functionality of chatbots can be divided into two categories: task-oriented chatbots that interact to fulfill tasks, usually for a specific domain, and non-task-oriented chatbots that engage in open-domain interactions, usually for the sake of amusement. Depending on the selected articles, different domains are divided into two main categories: open domains and closed domains. Table IV analyzes English chatbots followed by Arabic chatbots ordered by their publication years from domain perspective.

3) Evaluation metrics. Human-based and automatic-based evaluation metrics are two main metrics to evaluate chatbots. Human evaluation involves having a group of people communicate with the chatbot and evaluating various aspects using evaluation frameworks or questionnaires. The other categories include different proposed evaluation frameworks, such as sensibleness and specificity average (SSA) and quasi-Turing test method [39][40]. Table V presents a summary and comparison of more than 20 types of evaluation metrics used in chatbot development over the selected articles. While more than 70 articles have detailed the evaluation of their works, about 15 different articles seem not to evaluate their chatbot at all, which were classified under the not available (NA) category. From the table, it can be observed the most common measure used for both languages is human evaluation, then the BLUE and Accuracy for English and Arabic, respectively.

TABLE II. APPROACHES AND TECHNIQUES USED IN IMPLEMENTING SELECTED ARTICLES

Approach	*Techniques/Structures*		*Description*	*Advantages*	*Disadvantages*	*Articles*
Rule-based	Rules and patterns matching		Set of predefined human-made rules	• Simple and lower cost implementation • Quick deploy • No overtime for learning user intent	• Cannot learn on their own. • Fail to response outside its preset understanding	[41],[42],[43]
	Parsing		Converting a text to be less complicated terms	• Providing dependency relationship between words or semantic structure of the text	• Same the rules and patterns matching disadvantages	[44],[45],[46]
Corpus-based: 1.Retrieval-based	Pattern matching		Matching inputs with predefined structures of responses	• Sufficient on simple tasks • Select informative responses from candidate responses • More flexible than the rule-based.	• Providing chatbots without reasoning and creation. • Limited capabilities and repeated responses.	[47],[48],[49],[50],[51], [52],[53],[54],[55],[56],[57], [58],[59],[60],[61],[62],[63], [64],[65],[66],[67]
	AIML		Represents the knowledge as objects which derived from XML	• Advantages of pattern matching • Powerful in designing conversational flow	• Building all the possible patterns manually • Difficult to scaling	[68],[69],[70],[71],[72],[73], [74],[75],[76],[77],[78]
2.Generation-based	LSTM	CNN	CNN usually used for learning features utomatically by utilizing convolution and pooling processes	• Accurate on a larger dataset • Suitable to remember longer sequences	• Uses large number of parameters • Required more memory size • Long execution time • More complex structure	[79],[80]
		CNN and GRU	GRU is a type of RNN technique related with LSTM.			[81]
		RGDDA based on gradient reinforcement learning	Generating responses by utilizing user-specific information			[40]
		Stacked LSTM	Stacked LSTM is comprised of multiple LSTM layers			[82]
		Stacked LSTM and BiLSTM	BiLSTM: Both directions is followed by the input.			[83]
		BiLSTM, HRED	HRED generates context and response			[84]

Approach	*Techniques/Structures*		*Description*	*Advantages*	*Disadvantages*	*Articles*
			embeddingusing three stacked RNNs			
		Attention	Attention is based on RNNs cell and improved technique of seq2seq learning			[85]
	seq2seq	Seq2seq learning based on encoder-decoder architecture	Mapping a sequence of input words to another representation of response sequence.	• Same advantages of tachniques that seq2seq depending on it • Support variable-length size of input and response	• Same disadvantages of tachniques that seq2seq depending on it	[39],[86],[87],[88],[89],[90], [91]
	GRU		In compression to LSTM, GRU required fewer parameters training and it not being required for an additional cell state	• Uses less training parameter • Uses less memory • Take less time in execution • Less complex structure	• Not suitable for large dataset • •Not suitable for long-distance relations	[92],[93]
	RNN	Attention mechanisms besides BiGRU	Attention mechanisms Improved technique of seq2seq learning based on RNNs cell. Resolve the problem of systems' incapability to remember a longer sequence	• Uses less training parameter • Uses less memory • Take less time in execution • Less complex structure	• Suffer from gradient exploding and vanishing problems. • Difficult to process very longer sequences • Not suitable for parallelizing or stacking up	[94]
		Enhancement of RNN-GRU	This technique is improved by adding three additional cells which are Refinement, Adjustment, and output cells			[95]
	Pre-trained model	GPT-2, DIALOGPT, BoB, aubmindlab, CakeChat, asafaya,	To map words to actual number vectors, a language modeling and feature extraction technique was used.			[96],[97],[98],[99], [100]
Hybrid	Combination of generation and retrieval-based approaches		XiaoIce: more popular example from Microsoft	• Easy to select the attributes (relevance) from ranked features list.	• If the hybridization technique is not complementary to each other, the performance quality may decrease.	[101]
			Proposed PS, GP, and PRF			[102]
			Develop matching method based on the seq2seq			[103]
			Develop a model using the Twitter LDA model and attention mechanism			[104]
			Integrate AIML technique with a SNC model			[105]
			Multi-strategy process including LSTM with an attention mechanism beside rule-based technique			[106]

TABLE III. PLATFORMS USED IN IMPLEMENTING SELECTED ARTICLES

Framework	*NLP features*	*API*	*Control conversation*	*Languages*	*Limitation*	*Articles*
Google's DialogFlow	Yes	Yes	Yes	More than 45 languages from Bengali to Vietnamese except Arabic	Limitation of understanding the synonyms and hyponyms besides the documentation isn't very good	[107],[108],[109],[110]
Pandorabots	Yes	Yes	No	Support many languages including Arabic	Limitations to dealing with Arabic spelling mistakes	[76],[77]
IBM Watson Conversation	Yes	Yes	NO	13 languages from Arabic to Spanish	Not support Enhanced intent detection and autocorrection fixes misspellings	[111],[112]
Microsoft Azure	Yes	Yes	Yes	Translating 100 languages from Afrikaans to Yucatec Maya	Replying using translation system which increase the error rate	[60]
Rasa	Yes	Yes	Yes	Can be trained on any languages	Not support Arabic predefined trained entities	[113],[114]
Facebook Bot Engine (Wit.ai)	Yes	Yes	Yes	More than 100 languages from Afrikaans to Zulu	Time consuming when training the chatbot to understand all the different forms of Arabic text	[115],[116],[117]
Chatfuel	NO	Yes	NO	Support many languages including Arabic	Inflexible in terms of conversation flows and multi-languages	[118]
OSCOVA	Yes	Yes	Yes	NA	Not appropriate for complex conversation	[119]
Recast.AI	Yes	Yes	Yes	More than 15 languages from Arabic to Swedish	Poor documentation	[120],[121]

TABLE IV. DOMAINS USED IN SELECTED ARTICLES

Domain	*Languages*	*Articles*
Religion	Classical Arabic	[73],[60],[72]
Education	English	[117],[86],[90],[108],[118],[107],[45],[112],[109],[122]
	Classical and MSA Arabic	[50],[51]
	MSA Arabic	[54],[58],[57],[55],[52],[53],[67],[66],[110],[63],[65]
	Arabic dialects: Saudi Arabic dialect and Jordanian	[77],[78]
Healthcare	English	[41],[44],[47],[70],[49],[71],[119]
	MSA Arabic	[74],[111]
	Arabic Dialects: Egyptian	[100]
Tourism and airline	English	[35],[43]
	MSA Arabic	[59],[75],[46],[113],[115]
Business and customer service	English	[87],[69],[121],[81],[84],[95],[83],[106],[120]
Empathy and personalization	English	[123],[101],[92],[124]
	MSA Arabic	[85],[125]
Open	English	[68],[89],[79],[40],[104],[99],[36],[102],[94],[39],[82],[48],[105],[88],[114],[98],[116]
	MSA Arabic	[64],[42],[61],[56],[80],[93],[38],[62]
	Arabic Dialects: Gulf Arabic and Egyptian	[91],[76]

TABLE V. METRICS USED IN SELECTED ARTICLES

Categorization	*Metrics*	*Articles*
Automatic based Metric	F1-Score	[89],[114],[106],[80],[62],[100]
	Precision	[105],[106],[80],[62]
	Recall	[105],[106],[80],[62]
	Accuracy	[92],[114],[126],[105],[124],[80],[93],[62],[100],[76],[56],[74]
	PPL	[89],[102],[94],[92],[98],[124],[126],[39],[101],[85]
	BLEU	[79],[83],[94],[81],[92],[95],[84],[126],[127],[86],[101],[49],[40],[40],[99],[85],[91]
	ROUGE	[79],[83],[84],[49]
	MAP, P@1 and MRR	[103],[104]
	SkipThoughts cosine similarity, embedding average cosine similarity, vector extrema cosine similarity, BOW and greedy matching scores	[79],[102],[94],[81],[84],[127],[88]
	Other	[87],[95],[84],[39],[116],[40],[99],[109],[43], [48], [79], [102], [127]
NA		[70],[41],[44], [107],[47], [68], [69], [108], [121],[38], [61], [67], [67], [73], [108], [113]
Human based Metric	H: User Satisfaction	[35], [45], [82], [84], [89], [118], [128],[36], [49], [98], [101], [117],[112], [119], [120], [122],[60], [77], [85], [90],[63], [78], [91], [110], [125],[46], [57], [59], [111], [115],[54], [55], [65], [66], [75],[42], [50]–[53], [58], [64], [72]

IV. Discussion

This SLR reviews 88 articles to address the four research questions. This section discusses findings, highlights challenges and open problems with providing future research directions that we expect would help in developing chatbots.

A. Finding

1) Approaches. Initially, no approach or technique is the best for all domain applications. Selecting one approach or another deeply depends on several considerations. With recent improvement in computational recourse, most articles use generation-based approach, which is increasingly important in the development of chatbots. The encoder-decoder architecture is used as the main learning method in chatbots. However, a large number of datasets and high amount of computation time is required. When it comes to selecting the appropriate response from the structured data to respond to the user's input, the retrieval-based approach is the best. There are different improvements for retrieved information in this approach as shown in [47],[65],[66],[69]. However, purely retrieval-based approaches do not perform reasoning, and therefore, they are only suitable for mirroring current knowledge. More recently, significant findings based on performance has been offered by the emergence of the various hybrid approaches. This improvement may be due to the advantages of combining previously mentioned approaches; retrieval-based and generation-based. Moreover, the rule-based approach is straightforward, easy to implement, understand and fast but is too fixed to predefined rules in the database. Thus, extraneous inputs cannot be answered. Thus, it is limited used in developing chatbots in comparison to other approaches.

Furthermore, the generation-based approach in English-language chatbots has more attention. However, overall, there is currently limited research on Arabic-language chatbots. Retrieval-based approach represents the vast majority of the selected approaches in Arabic chatbot research articles, whereas generation-based approach has started to attract attention in 2018 [80]. The most commonly used technique for Arabic-language chatbots is the pattern matching followed by AIML in retrieval-based approach. The LSTM technique is the most employed in developing Arabic chatbots followed by seq2seq and GRU in the generation-based approach.

Besides the three main approaches, some selected articles use publicly available platforms to create and launch their chatbots as presented in Table III. These platforms eliminate the required experience in coding to develop the chatbots from scratch. The platforms simplify development and standardize some implementation processes. Although some of these platforms have a well-described approach, such as Pandora's [129] that uses the retrieval-based approach, most hide the details of their systems. In addition, A closer inspection of the table shows that Google's DialogFlow, Facebook's Wit.ai, IBM Watson Conversation, and Microsoft Azure are cloud-based platforms that support different programming languages besides the natural languages. However, they differ significantly in other aspects [130].

2) Domains. The selected articles show most efforts was devoted to the education domain. A possible explanation might be because of several possible potential areas of education where chatbots can be utilized. According to Fig. 4, there is a significant difference in the published research between Arabic and English chatbots, especially in the business domain. Simultaneously, some of the Arabic chatbots are developed as religious chatbots, whereas English chatbots focus on other domains. Overall, the remaining domains are addressed almost equally in both languages. Furthermore, the research to date has tended to focus on MSA rather than Arabic dialects. This finding may be explained by the fact that MSA has a formal and clear format in written and expression, which helps with analyzing. However, due to the rise of social media platforms, few recently published articles focused on Arabic dialect chatbots, such as those in the Gulf Arabic dialect, Saudi dialect, Egyptian dialect and Jordanian dialect.

In addition, a relationship is observed between the domains and the approaches. Fig. 5 displays the distribution of the investigated approaches in 88 articles across seven domains. Obviously, for the education and health domains, the majority of chatbots seem to be developing using the retrieval-based approach. The possible explanation of this bias is that the chatbots in these domains are always prepared to fit a specific knowledge base. In contrast, most of the chatbots in business, emotions, and open domains are built using the generation-based approach. Thus, chatbots should be able to produce a new natural conversation with more appropriate responses.

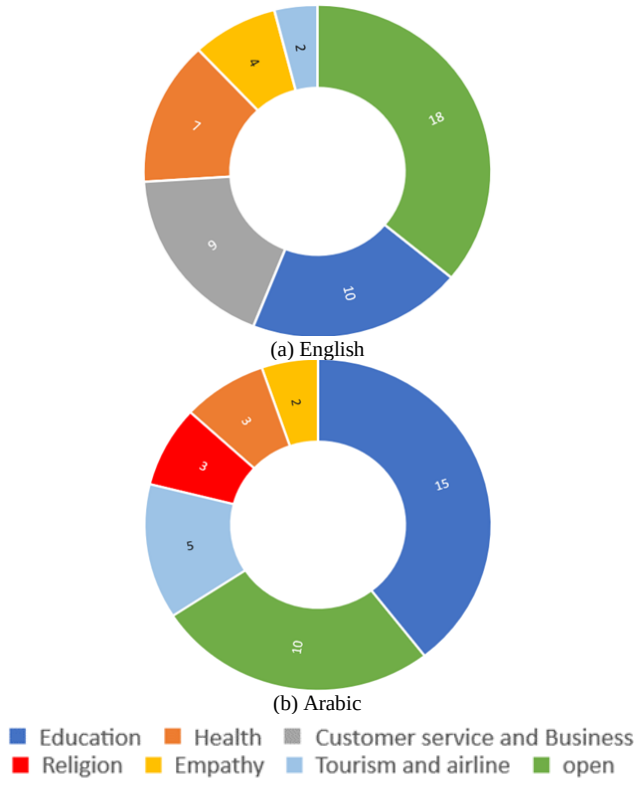

Fig. 4. Finding in Domains with Languages Perspective.

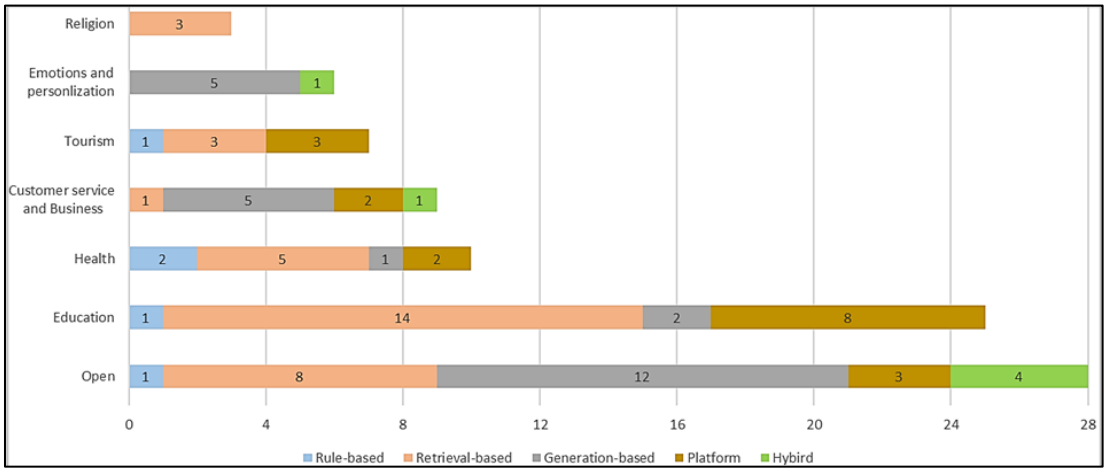

Fig. 5. Finding in Domains with Approaches Perspective.

3) Evaluation metrics. The vast majority of chatbot research uses the human-based metric, followed by BLUE and then accuracy, as presented in Fig. 6, where the most used types of the ROUGE metric are ROUGE-1 (unigram) and ROUGE-2 (bi-gram). Overall, no specific evaluation metric is more represented in articles due to the lack of gold standards of the evaluation. The evaluation of empathy, user satisfaction, and fluency are examples of the needed intervention of human evaluation. In terms of the time and resources, automated evaluation measures are more efficient than human ones. However, they appear to be incapable of accurately assessing the quality and efficacy of the entire conversation even they are easier to use and do not require manual work by human judges.

Moreover, several important differences are between the Arabic and English metrics. From Fig. 6, most of the Arabic articles focused on human evaluation whereas human evaluation and the BLEU metric are more popular in English articles. This may be due to several challenges of the Arabic language such as the lack of available data resources.

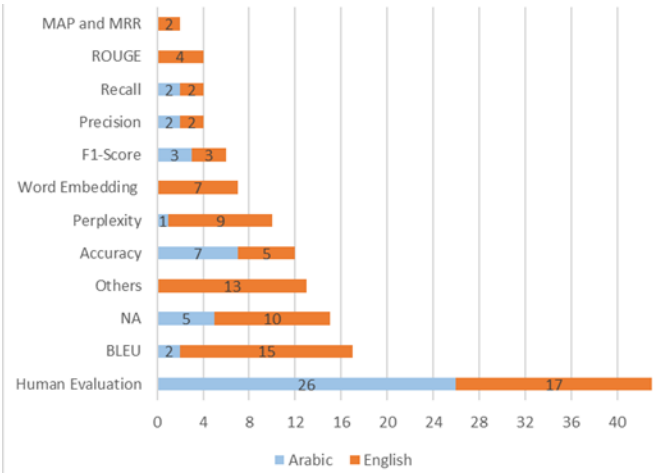

Fig. 6. Finding in Metrics Perspective.

B. Open Research Problems

Several challenges are in chatbots development, but for a comprehensive conversation experience, the development requires a balance between different directions.

1) Ambiguity in conversation. The simple meaning in natural language can be interpreted in more than one way. Therefore, simulating human conversation with a chatbot is quite challenging. Although deep learning has recently taken a lot of researcher attention, there is still a long way to go before chatbots can produce human-like conversation. Moreover, ambiguous word senses are a problem in the building of a semantic conversational AI.

2) Consistency in natural language interpretation. Inaccurate and unsuitable responses are given as a result of inconsistency in interpretation. Microsoft Tay chatbot is a popular example of natural language misinterpretation. It was originally released by Microsoft Corporation via Twitter on March 23, 2016 and was shut down after 16 hours of its launch due to inconsistencies in interpretation [131].

3) Detecting and maintaining conversation. The flow of the conversation is determined by the context of the conversation. Open-domain conversation is another challenge that required determining the topic and keeping track of the context besides detecting when the topic is changed. Despite the spectacular development of chatbots, they are occasionally unable to recognize the intent of users, which makes users frustrated.

4) Data efficiency and time. There is still a problem of little or no datasets being available in certain domains, especially clear in task-oriented systems, where gathering datasets would be costly and time-consuming for new domains. In English, several lexicons are built, such as SentiWordNet and WordNet [132]. By contrast, a lack of Arabic lexicons and resources affects the progress of Arabic chatbots.

5) Evaluation. No universal framework is to evaluate chatbots. There is a lack of unified definitions, metrics, and validated scales in evaluation [133]. The lack of a common

framework makes accurate testing and comparison of different models difficult. Although human evaluation offers qualitative estimation, it is subjective and time-consuming.

6) Privacy and security. Some chatbots are designed to rely on APIs to obtain information, which makes it important to secure the user's data. The kinds of data that the chatbot collects and provides can be used to supplement the support offered by different services, such as in the education domain. This will almost certainly improve the quality and efficiency of the resources available. Moreover, ethical issues of using chatbots are, either from the user side when they abuse the chatbot, or from the chatbot side, when they may save and use the user information for different purposes.

7) Lack of emotions. Some chatbots are developed with predefined conversations that limit their linguistic intelligence and make them mechanical and incapable of communicating in a natural manner. Involving these emotions in chatbots gives them human-like communicative behaviors, recognizing the different meaning possibilities of input and then producing more appropriate responses.

C. Open Research Problems

1) Datasets. Although many word sense disambiguation approaches have been developed [134], [135], they typically increase the computational complexity, which may not be a desirable solution. Moreover, most of the selected articles have built their own dataset, especially for the Arabic research, and others build Arabic corpora [136],[137], [138], whereas others used translation techniques [61],[72]. However, these solutions are limited, and there is still an open-domain problem that needs a lot of attention. In addition, a limitation of Arabic annotated data sets is another problem [136]. Thus, working on providing appropriate data sets and making them available for research can be considered a valuable contribution to chatbot research.

2) Evaluation framework. A new comprehensive framework of evaluation should be provided. Building this framework is not an easy task and may be affected by different factors, such as inputs having multiple semantic meanings and the length of the conversation that may be related to the task itself, for example, the flow of direct questions asked by students differs from entertainment conversations. The framework may also distinguish from different domains, for example, evaluating student understanding differs from completing booking tasks. Thus, the evaluation frameworks must assess providing these tasks to users, and research must define a robust evaluation framework that can mitigate the negative effects of these challenges.

3) Human-like conversation. The current research suffers from limitations in generating a natural conversation, resulting in a noneffective chatbot [10]. A number of factors may affect the behavior of chatbot conversation. First, ambiguous inputs need to be verified and detected to produce appropriate behaviors, such as asking to rephrase the input. Second, bi-linguistic, or multi-linguistic chatbots offer a wider variety of capabilities and provide more user trust. Third, leveraging emotional and contextual cues encourages a user to continue chatting, necessitating further investigation by the researcher into sentiment and emotional analysis. From the Arabic language side, many articles have been conducted on Arabic morphological analysis and generation using a range of methods. This applies in various levels of linguistic complexity, including stemmers [139],[140]. However, Arabic is a derivational and inflectional language that needs a lot of attention and improvement.

4) Extended and different perspectives. Although the selected articles critically survey the state-of-the-art solutions, they focus on specific research. Hence, other perspectives are not addressed in this SLR, and the challenges exposed in this SLR can inspire researchers to focus more on addressing them. Furthermore, due to access constraints, six databases are considered for selecting the articles. Thus, articles in the remaining databases may support or limit the findings in this SLR. Even so, this SLR can be considered as a contribution toward English and Arabic research for the taken criteria.

5) Empirical investigation. This SLR addresses the four research questions without empirical investigation. Involving empirical contributions in future works would give a broader analysis of chatbot development and usage.

V. Conclusion

Chatbot usage has become increasingly prevalent in recent years. One of the key goals of adopting chatbots is minimizing human involvement. The rapid technological advancements of AI aid in achieving this goal and help in the development of more flexible chatbots that are able to produce human-like conversation. This research provides a systematic review of the articles on the evolution of chatbots to investigate the four research questions. A systematic review protocol was used to analyze 50 and 38 articles for English and Arabic works, respectively, and to extract research from six well-known digital databases in computer science. The SLR analyzes the articles in terms of development techniques, domains, evaluation metrics and underlines some challenges and open problems of chatbot development. Furthermore, presenting future directions may assist researchers in identifying crucial aspects that require deep investigation and more development. The findings show that the research domain targeting education receives greater attention from researchers than other domains, and the retrieval-based approach is the most widely utilized approach in this domain. However, this approach is not able to generate a new response that is not predefined in the chatbot's knowledge base. In contrast, the generation-based approach is suitable for tasks that demand providing a new response. Hybrid approaches generally combine between these approaches and are most used for ranking the multiple possible responses when its performance may improve by using one approach rather than another one. However, they still require further developed.

This SLR concludes that current chatbots are still unable to simulate human conversation. Simultaneously, increasing research interest and rapid technological advancements could

evolve chatbot conversation and make chatbots more flexible, fluent, and human-like. Indeed, this SLR provides various recommendations for future articles, which creates a chance for researchers to continue to develop research on chatbots.

ACKNOWLEDGMENT

Lubna Alhenaki would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-258.

REFERENCES

[1] M. McTear, "Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots," Synthesis Lectures on Human Language Technologies, vol. 13, no. 3, pp. 1–251, Oct. 2020, doi: 10.2200/S01060ED1V01Y202010HLT048.

[2] K. Darwish et al., "A Panoramic Survey of Natural Language Processing in the Arab World," arXiv:2011.12631 [cs], Nov. 2020, Accessed: Dec. 15, 2020. [Online]. Available: http://arxiv.org/abs/2011.12631

[3] E. H. Almansor and F. K. Hussain, "Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions," in Complex, Intelligent, and Software Intensive Systems, vol. 993, L. Barolli, F. K. Hussain, and M. Ikeda, Eds. Cham: Springer International Publishing, 2020, pp. 534–543. doi: 10.1007/978-3-030-22354-0_47.

[4] R. Kumar and M. M. Ali, "A Review on Chatbot Design and Implementation Techniques," vol. 07, no. 02, p. 11, 2020.

[5] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in Artificial Intelligence Applications and Innovations, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer International Publishing, 2020, pp. 373–383. doi: 10.1007/978-3-030-49186-4_31.

[6] M. W. Ashfaque, S. Tharewal, S. Iqhbal, and C. N. Kayte, "A Review on Techniques, Characteristics and approaches of an intelligent tutoring Chatbot system," in 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 258–262.

[7] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," Computers and Education: Artificial Intelligence, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.

[8] A. M., K. Ramasamy, S. G., and K. S.R., "A Systematic Survey of Cognitive Chatbots in Personalized Learning Framework," in 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, Mar. 2021, pp. 241–245. doi: 10.1109/WiSPNET51692.2021.9419403.

[9] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," Expert Systems with Applications, vol. 184, p. 115461, Dec. 2021, doi: 10.1016/j.eswa.2021.115461.

[10] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," Machine Learning with Applications, vol. 2, p. 100006, Dec. 2020, doi: 10.1016/j.mlwa.2020.100006.

[11] G. Caldarini, S. Jaf, and K. McGarry, "A Literature Survey of Recent Advances in Chatbots," Information, vol. 13, no. 1, p. 41, Jan. 2022, doi: 10.3390/info13010041.

[12] S. Singh and H. Beniwal, "A survey on near-human conversational agents," Journal of King Saud University - Computer and Information Sciences, p. S1319157821003001, Nov. 2021, doi: 10.1016/j.jksuci.2021.10.013.

[13] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, pp. 1–22, Dec. 2009, doi: 10.1145/1644879.1644881.

[14] M. Hijjawi and Y. Elsheikh, "Arabic Language Challenges in Text Based Conversational Agents Compared to The English Language," IJCSIT, vol. 7, no. 3, pp. 1–13, Jun. 2015, doi: 10.5121/ijcsit.2015.7301.

[15] S. AlHumoud, A. Al, and W. Aldamegh, "Arabic Chatbots: A Survey," ijacsa, vol. 9, no. 8, 2018, doi: 10.14569/IJACSA.2018.090867.

[16] A. A. Elmadany, S. M. Abdou, and M. Gheith, "A Survey of Arabic Dialogues Understanding for Spontaneous Dialogues and Instant Message," IJNLC, vol. 4, no. 2, pp. 75–94, Apr. 2015, doi: 10.5121/ijnlc.2015.4206.

[17] E. S. AL-Hagbani and M. B. Khan, "Support of Existing Chatbot Development Framework for Arabic Language: A Brief Survey," in 5th International Symposium on Data Mining Applications, vol. 753, M. Alenezi and B. Qureshi, Eds. Cham: Springer International Publishing, 2018, pp. 26–35. doi: 10.1007/978-3-319-78753-4_3.

[18] A. Fuad and M. Al-Yahya, "Recent Developments in Arabic Conversational AI: A Literature Review," IEEE Access, vol. 10, pp. 23842–23859, 2022, doi: 10.1109/ACCESS.2022.3155521.

[19] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouani, A. Abd-alrazaq, and M. Househ, "Arabic Chatbot Technologies: A Scoping Review," Computer Methods and Programs in Biomedicine Update, p. 100057, Apr. 2022, doi: 10.1016/j.cmpbup.2022.100057.

[20] D. Jurafsky and J. Martin, Speech and Language Processing. 2022.

[21] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A Survey on Conversational Agents/Chatbots Classification and Design Techniques," in Web, Artificial Intelligence and Network Applications, vol. 927, L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, Eds. Cham: Springer International Publishing, 2019, pp. 946–956. doi: 10.1007/978-3-030-15035-8_93.

[22] R. Agarwal and M. Wadhwa, "Review of State-of-the-Art Design Techniques for Chatbots," SN COMPUT. SCI., vol. 1, no. 5, p. 246, Sep. 2020, doi: 10.1007/s42979-020-00255-3.

[23] V. V., J. B. Cooper, and R. L. J., "Algorithm Inspection for Chatbot Performance Evaluation," Procedia Computer Science, vol. 171, pp. 2267–2274, 2020, doi: 10.1016/j.procs.2020.04.245.

[24] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A Survey of Design Techniques for Conversational Agents," in Information, Communication and Computing Technology, vol. 750, S. Kaushik, D. Gupta, L. Kharb, and D. Chahal, Eds. Singapore: Springer Singapore, 2017, pp. 336–350. doi: 10.1007/978-981-10-6544-6_31.

[25] Q. Motger, X. Franch, and J. Marco, "Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges," arXiv:2106.10901 [cs], Jun. 2021, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2106.10901

[26] I. Sutskever, J. Martens, and G. Hinton, "Generating Text with Recurrent Neural Networks," p. 8.

[27] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," Int. J. Unc. Fuzz. Knowl. Based Syst., vol. 06, no. 02, pp. 107–116, Apr. 1998, doi: 10.1142/S0218488598000094.

[28] "Long Short-Term Memory."

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv:1412.3555 [cs], Dec. 2014, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1412.3555.

[30] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv:1409.1259 [cs, stat], Oct. 2014, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1409.1259.

[31] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," p. 9.

[32] O. Vinyals and Q. Le, "A Neural Conversational Model," arXiv:1506.05869 [cs], Jul. 2015, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1506.05869.

[33] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," vol. 2, Jan. 2007.

[34] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," Journal of Systems and Software, vol. 125, pp. 207–219, Mar. 2017, doi: 10.1016/j.jss.2016.11.027.

[35] V. Kasinathan, M. H. A. Wahab, S. Z. S. Idrus, A. Mustapha, and K. Z. Yuen, "AIRA Chatbot for Travel: Case Study of AirAsia," J. Phys.: Conf. Ser., vol. 1529, no. 2, p. 022101, Apr. 2020, doi: 10.1088/1742-6596/1529/2/022101.

[36] M. Vanjani, Milam Aiken, and M. Park, "Chatbots for Multilingual Conversations," Jul. 2019, doi: 10.5281/ZENODO.3264011.

[37] "snatchbot." https://snatchbot.me/.

[38] Y. M. Mohialden, M. T. Younis, and N. M. Hussien, "A Novel Approach to Arabic Chabot, Utilizing Google Colab and the Internet of Things: A Case Study at a Computer Center," WEB, vol. 18, no. 2, pp. 946–954, Dec. 2021, doi: 10.14704/WEB/V18I2/WEB18365.

[39] D. Adiwardana et al., "Towards a Human-like Open-Domain Chatbot," arXiv:2001.09977 [cs, stat], Feb. 2020, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/2001.09977.

[40] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, and J. Zhu, "Personalized response generation by dual-learning based domain adaptation," Neural Networks, vol. 103, pp. 72–82, 2018.

[41] J. Weizenbaum, "ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine," ACM, vol. 9, no. 1, pp. 36–45, 1966.

[42] M. Makatchev et al., "Dialogue patterns of an arabic robot receptionist," in Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10, Osaka, Japan, 2010, p. 167. doi: 10.1145/1734454.1734526.

[43] B. Liu and C. Mei, "Lifelong Knowledge Learning in Rule-based Dialogue Systems," p. 5.

[44] K. Colby, S. Weber, and F. Hilf, "Artificial Paranoia," Artificial Intelligence, pp. 1–25, 1971.

[45] J. Jia, "CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning," Knowledge-Based Systems, vol. 22, no. 4, pp. 249–255, May 2009, doi: 10.1016/j.knosys.2008.09.001.

[46] A. Moubaiddin, O. Shalbak, B. Hammo, and N. Obeid, "Arabic Dialogue System for Hotel Reservation based on Natural Language Processing Techniques," CyS, vol. 19, no. 1, Mar. 2015, doi: 10.13053/cys-19-1-1962.

[47] Navida Belgaumwala, "Chatbot: A Virtual Medical Assistant," International Journal for Research in Applied Science and Engineering Technology, vol. 7, pp. 1042–1050, Jun. 2019, doi: 10.22214/ijraset.2019.6179.

[48] H. Candra, "Designing a Chatbot Application for Student Information Centers on Telegram Messenger Using Fulltext Search Boolean Mode," p. 10.

[49] N. A. I. Omoregbe, I. O. Ndaman, S. Misra, O. O. Abayomi-Alli, and R. Damaševičius, "Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic," Journal of Healthcare Engineering, vol. 2020, pp. 1–14, Sep. 2020, doi: 10.1155/2020/8839524.

[50] O. G. Alobaidi, K. A. Crockett, J. D. O'Shea, and T. M. Jarad, "Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education," p. 7, 2013.

[51] O. G. Alobaidi, K. Crockett, J. D. O'Shea, and T. M. Jarad, "The Application of Learning Theories into Abdullah: An Intelligent Arabic Conversational Agent Tutor:," in Proceedings of the International Conference on Agents and Artificial Intelligence, Lisbon, Portugal, 2015, pp. 361–369. doi: 10.5220/0005197003610369.

[52] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, A. Latham, and M. Kaleem, "Development of an Arabic Conversational Intelligent Tutoring System for Education of children with ASD," in 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, Jun. 2017, pp. 24–29. doi: 10.1109/CIVEMSA.2017.7995296.

[53] S. Aljameel, J. O'Shea, K. Crockett, A. Latham, and M. Kaleem, "LANA-I: An Arabic Conversational Intelligent Tutoring System for Children with ASD," in Intelligent Computing, vol. 997, K. Arai, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2019, pp. 498–516. doi: 10.1007/978-3-030-22871-2_34.

[54] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "ArabChat: An arabic conversational agent," in 2014 6th International Conference on Computer Science and Information Technology (CSIT), 2014, pp. 227–237.

[55] M. Hijjawi, Z. Bandar, and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," ijacsa, vol. 7, no. 2, 2016, doi: 10.14569/IJACSA.2016.070247.

[56] M. Hijjawi, Z. Bandar, and K. Crockett, "User's utterance classification using machine learning for Arabic Conversational Agents," in 2013 5th International Conference on Computer Science and Information Technology, Amman, Jordan, Mar. 2013, pp. 223–232. doi: 10.1109/CSIT.2013.6588784.

[57] M. Hijjawi, Z. Bandar, and K. Crockett, "A Novel Hybrid Rule Mechanism for the Arabic Conversational Agent ArabChat," p. 10, 2015.

[58] M. Hijjawi, H. Qattous, and O. Alsheiksalem, "Mobile Arabchat: An Arabic Mobile-Based Conversational Agent," ijacsa, vol. 6, no. 10, 2015, doi: 10.14569/IJACSA.2015.061016.

[59] Z. Noori, Z. Bandar, and K. Crockett, "Arabic Goal-oriented Conversational Agent Based on Pattern Matching and Knowledge Trees," p. 7, 2014.

[60] Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, S. M. Yassin, M. Z. Khan, and Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, "SeerahBot: An Arabic Chatbot About Prophet's Biography," ijircst, vol. 9, no. 2, pp. 89–97, Mar. 2021, doi: 10.21276/ijircst.2021.9.2.13.

[61] N. Mavridis, A. AlDhaheri, L. AlDhaheri, M. Khanii, and N. AlDarmaki, "Transforming IbnSina into an advanced multilingual interactive android robot," in 2011 IEEE GCC Conference and Exhibition (GCC), Dubai, United Arab Emirates, Feb. 2011, pp. 120–123. doi: 10.1109/IEEEGCC.2011.5752467.

[62] N. O. Alshammari and F. D. Alharbi, "Combining a Novel Scoring Approach with Arabic Stemming Techniques for Arabic Chatbots Conversation Engine," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 21, no. 4, pp. 1–21, Jul. 2022, doi: 10.1145/3511215.

[63] S. Z. Sweidan, S. S. Abu Laban, N. A. Alnaimat, and K. A. Darabkh, " SIAAA - C: A student interactive assistant android application with chatbot during COVID - 19 pandemic," Comput Appl Eng Educ, vol. 29, no. 6, pp. 1718‐1742, Nov. 2021, doi: 10.1002/cae.22419.

[64] L. D. Riek et al., "Ibn Sina Steps Out: Exploring Arabic Attitudes Toward Humanoid Robots," p. 8, 2010.

[65] S. Z. Sweidan, S. S. Abu Laban, N. A. Alnaimat, and K. A. Darabkh, "SEG-COVID: A Student Electronic Guide within Covid-19 Pandemic," in 2021 9th International Conference on Information and Education Technology (ICIET), Okayama, Japan, Mar. 2021, pp. 139–144. doi: 10.1109/ICIET51873.2021.9419656.

[66] H. ElGibreen, S. Almazyad, S. B. Shuail, M. A. Qahtani, and L. ALhwiseen, "Robot Framework for Anti-Bullying in Saudi Schools," in 2020 Fourth IEEE International Conference on Robotic Computing (IRC), Taichung, Taiwan, Nov. 2020, pp. 166–171. doi: 10.1109/IRC.2020.00033.

[67] Y. Almurtadha, "LABEEB: Intelligent Conversational Agent Approach to Enhance Course Teaching and Allied Learning Outcomes attainment," JACSM, vol. 13, no. 1, pp. 9–12, 2019, doi: 10.4316/JACSM.201901001.

[68] R. Wallace, "Artificial linguistic internet computer entity (alice)," City, 1995.

[69] Department of Computer Applications Cochin University of Science and Technology Cochin, India and Reshmi. S, "EMPOWERING CHATBOTS WITH BUSINESS INTELLIGENCE BY BIG DATA INTEGRATION," ijarcs, vol. 9, no. 1, pp. 627–631, Feb. 2018, doi: 10.26483/ijarcs.v9i1.5398.

[70] S. Roca, J. Sancho, J. García, and Á. Alesanco, "Microservice chatbot architecture for chronic patient support," Journal of Biomedical Informatics, vol. 102, p. 103305, Feb. 2020, doi: 10.1016/j.jbi.2019.103305.

[71] D. Zhang, X. Chen, Y. Zhang, and S. Qin, "Template-based Chatbot for Agriculture Related FAQs," undefined, 2021, Accessed: May 03, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Template-based-Chatbot-for-Agriculture-Related-FAQs-Zhang-Chen/ec0f4128378064b8014edeb6cbb9cdfa5834ac3a.

[72] B. A. Shawar and E. Atwell, "Accessing an Information System by Chatting," in Natural Language Processing and Information Systems, vol. 3136, F. Meziane and E. Métais, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 407–412. doi: 10.1007/978-3-540-27779-8_39.

[73] B. Shawar and E. Atwell, "An Arabic chatbot giving answers from the Qur'an," in Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, vol. 2, pp. 197–202, 2004.

[74] B. Abu Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA," Int. J. Emerg. Technol. Learn., vol. 6, no. 1, pp. 37–43, Mar. 2011, doi: 10.3991/ijet.v6i1.1502.

[75] T. Kadeed, Construction of Arabic Interactive Tool Between Humans and Intelligent Agents. 2014.

[76] D. A. Ali and N. Habash, "Botta: An Arabic Dialect Chatbot," p. 5.

[77] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic Dialect Chatbot," IJACSA, vol. 11, no. 3, 2020, doi: 10.14569/IJACSA.2020.0110357.

[78] N. A. Al-Madi, K. A. Maria, M. A. Al-Madi, M. A. Alia, and E. A. Maria, "An Intelligent Arabic Chatbot System Proposed Framework," in 2021 International Conference on Information Technology (ICIT), Amman, Jordan, Jul. 2021, pp. 592–597. doi: 10.1109/ICIT52682.2021.9491699.

[79] Y. Zhang et al., "Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization," p. 11.

[80] A. M. Bashir, A. Hassan, B. Rosman, D. Duma, and M. Ahmed, "Implementation of A Neural Natural Language Understanding Component for Arabic Dialogue Systems," Procedia Computer Science, vol. 142, pp. 222–229, 2018, doi: 10.1016/j.procs.2018.10.479.

[81] M. Aleedy, H. Shaiba, and M. Bezbradica, "Generating and Analyzing Chatbot Responses using Natural Language Processing," IJACSA, vol. 10, no. 9, 2019, doi: 10.14569/IJACSA.2019.0100910.

[82] O. Octavany and A. Wicaksana, "Cleveree: an artificially intelligent web service for Jacob voice chatbot," TELKOMNIKA, vol. 18, no. 3, p. 1422, Jun. 2020, doi: 10.12928/telkomnika.v18i3.14791.

[83] J. Kapočiūtė-Dzikienė, "A Domain-Specific Generative Chatbot Trained from Little Data," Applied Sciences, vol. 10, no. 7, p. 2221, Mar. 2020, doi: 10.3390/app10072221.

[84] T. Hori, W. Wang, Y. Koji, C. Hori, B. Harsham, and J. R. Hershey, "Adversarial training and decoding strategies for end-to-end neural conversation models," Computer Speech & Language, vol. 54, pp. 122–139, Mar. 2019, doi: 10.1016/j.csl.2018.08.006.

[85] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven Arabic Conversational Chatbot," p. 11.

[86] K. Palasundram, N. Mohd Sharef, N. A. Nasharuddin, K. A. Kasmiran, and A. Azman, "Sequence to Sequence Model Performance for Education Chatbot," Int. J. Emerg. Technol. Learn., vol. 14, no. 24, p. 56, Dec. 2019, doi: 10.3991/ijet.v14i24.12187.

[87] T. Hu et al., "Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media," arXiv:1803.02952 [cs], Mar. 2018, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1803.02952

[88] J. Prassanna, "Towards Building A Neural Conversation Chatbot Through Seq2Seq Model," vol. 9, no. 03, p. 5, 2020.

[89] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?," arXiv:1801.07243 [cs], Sep. 2018, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1801.07243

[90] R. Patel, N. Bhagora, P. Singh, and K. Namdev, "CLOUD BASED STUDENT INFORMATION CHATBOT," vol. 02, no. 04, p. 4.

[91] T. Alshareef and M. A. Siddiqui, "A seq2seq Neural Network based Conversational Agent for Gulf Arabic Dialect," in 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, Nov. 2020, pp. 1–7. doi: 10.1109/ACIT50332.2020.9300059.

[92] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human–machine dialogue modelling with the fusion of word- and sentence-level emotions," Knowledge-Based Systems, vol. 192, p. 105319, Mar. 2020, doi: 10.1016/j.knosys.2019.105319.

[93] M. Boussakssou, H. Ezzikouri, and M. Erritali, "Chatbot in Arabic language using seq to seq model," Multimed Tools Appl, vol. 81, no. 2, pp. 2859–2871, Jan. 2022, doi: 10.1007/s11042-021-11709-y.

[94] S. Kim, O.-W. Kwon, and H. Kim, "Knowledge-Grounded Chatbot Based on Dual Wasserstein Generative Adversarial Networks with Effective Attention Mechanisms," Applied Sciences, vol. 10, no. 9, p. 3335, May 2020, doi: 10.3390/app10093335.

[95] V.-K. Tran and L.-M. Nguyen, "Gating mechanism based Natural Language Generation for spoken dialogue systems," Neurocomputing, vol. 325, pp. 48–58, Jan. 2019, doi: 10.1016/j.neucom.2018.09.069.

[96] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], May 2019, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[97] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," p. 12.

[98] H. Song, Y. Wang, K. Zhang, W.-N. Zhang, and T. Liu, "BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data," arXiv:2106.06169 [cs], Jun. 2021, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/2106.06169

[99] Y. Zhang et al., "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation," arXiv:1911.00536 [cs], May 2020, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1911.00536.

[100] T. Wael, A. Hesham, M. Youssef, O. Adel, H. Hesham, and M. S. Darweesh, "Intelligent Arabic-Based Healthcare Assistant," in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, Oct. 2021, pp. 216–221. doi: 10.1109/NILES53778.2021.9600526.

[101] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot," Computational Linguistics, vol. 46, no. 1, pp. 53–93, Mar. 2020, doi: 10.1162/coli_a_00368.

[102] L. Zhang, Y. Yang, J. Zhou, C. Chen, and L. He, "Retrieval-Polished Response Generation for Chatbot," IEEE Access, vol. 8, pp. 123882–123890, 2020, doi: 10.1109/ACCESS.2020.3004152.

[103] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots," arXiv:1805.02333 [cs], May 2018, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1805.02333

[104] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Response Selection with Topic Clues for Retrieval-based Chatbots," arXiv:1605.00090 [cs], Sep. 2016, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1605.00090.

[105] Y. S. Wijaya and F. Zoromi, "Chatbot Designing Information Service for New Student Registration Based on AIML and Machine Learning," vol. 1, no. 1, p. 10, 2020.

[106] M. Nuruzzaman and O. K. Hussain, "IntelliBot: A Dialogue-based chatbot for the insurance industry," Knowledge-Based Systems, vol. 196, p. 105810, May 2020, doi: 10.1016/j.knosys.2020.105810.

[107] O. Zahour, "Towards a Chatbot for educational and vocational guidance in Morocco: Chatbot E-Orientation," IJETER, vol. 9, no. 2, pp. 2479–2487, Apr. 2020, doi: 10.30534/ijatcse/2020/237922020.

[108] S. S. Ranavare and R. S. Kamath, "Artificial Intelligence based Chatbot for Placement Activity at College Using DialogFlow," vol. 68, no. 30, p. 10.

[109] S. Sajjapanroj, P. Longpradit, and K. Polanunt, "A Prototype of Google Dialog Flow for School Teachers' Uses in Conducting Classroom Research," p. 14.

[110] W. El Hefny, Y. Mansy, M. Abdallah, and S. Abdennadher, "Jooka: A Bilingual Chatbot for University Admission," in Trends and Applications in Information Systems and Technologies, vol. 1367, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds. Cham: Springer International Publishing, 2021, pp. 671–681. doi: 10.1007/978-3-030-72660-7_64.

[111] Department of Computer Science, Universita Degli Studi di Trento, Italy, A. Fadhil, A. AbuRa'ed, and Information & Communication Technologies, Universitat Pompeu Fabra Barcelona, Spain, "OlloBot - Towards A Text-Based Arabic Health Conversational Agent: Evaluation and Results," in Proceedings - Natural Language Processing in a Deep Learning World, Oct. 2019, pp. 295–303. doi: 10.26615/978-954-452-056-4_034.

[112] S. Memeti and S. Pllana, "PAPA: A parallel programming assistant powered by IBM Watson cognitive computing technology," Journal of Computational Science, vol. 26, pp. 275–284, May 2018, doi: 10.1016/j.jocs.2018.01.001.

[113] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamdi, "AI Chatbot for Tourist Recommendations: A Case Study in the City of Jeddah, Saudi

Arabia," Int. J. Interact. Mob. Technol., vol. 14, no. 19, p. 18, Nov. 2020, doi: 10.3991/ijim.v14i19.17201.

[114] N. T. M. Trang and M. Shcherbakov, "Enhancing Rasa NLU model for Vietnamese chatbot," vol. 9, p. 6, 2021.

[115] A.-H. Al-Ajmi and N. Al-Twairesh, "Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach," IEEE Access, vol. 9, pp. 7043–7053, 2021, doi: 10.1109/ACCESS.2021.3049732.

[116] "JACOB Voice Chatbot Application using Wit.ai for Providing Information in UMN," IJEAT, vol. 8, no. 6S3, pp. 105–109, Nov. 2019, doi: 10.35940/ijeat.F1017.0986S319.

[117] University of Jeddah, Jeddah, Saudi Arabia and A. A. Qaffas, "Improvement of Chatbots Semantics Using Wit.ai and Word Sequence Kernel: Education Chatbot as a Case Study," IJMECS, vol. 11, no. 3, pp. 16–22, Mar. 2019, doi: 10.5815/ijmecs.2019.03.03.

[118] F. O. Chete and G. O. Daudu, "An Approach towards the Development of a Hybrid Chatbot for Handling Students' Complaints," p. 10.

[119] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test," IEEE Trans. Emerg. Topics Comput., vol. 9, no. 3, pp. 1170–1182, Jul. 2021, doi: 10.1109/TETC.2020.2974478.

[120] V. Oguntosin and A. Olomo, "Development of an E-Commerce Chatbot for a University Shopping Mall," Applied Computational Intelligence and Soft Computing, vol. 2021, pp. 1–14, Mar. 2021, doi: 10.1155/2021/6630326.

[121] K. Khavya, "Banking Bot," International Journal of New Technology and Research, vol. 4, no. 7, p. 263023.

[122] K. Mageira, D. Pittou, A. Papasalouros, K. Kotis, P. Zangogianni, and A. Daradoumis, "Educational AI Chatbots for Content and Language Integrated Learning," Applied Sciences, vol. 12, no. 7, p. 3239, Mar. 2022, doi: 10.3390/app12073239.

[123] W.-N. Zhang, Q. Zhu, Y. Wang, Y. Zhao, and T. Liu, "Neural personalized response generation as domain adaptation," World Wide Web, vol. 22, no. 4, pp. 1427–1446, Jul. 2019, doi: 10.1007/s11280-018-0598-6.

[124] A. I. Niculescu, I. Kukanov, and B. Wadhwa, "DigiMo - towards developing an emotional intelligent chatbot in Singapore," in Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures, Honolulu HI USA, Apr. 2020, pp. 29–32. doi: 10.1145/3391203.3391210.

[125] T. Naous, W. Antoun, R. A. Mahmoud, and H. Hajj, "Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data," arXiv:2103.04353 [cs], Mar. 2021, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2103.04353.

[126] University Politehnica of Bucharest, A. Grosuleac, S. Budulan, University Politehnica of Bucharest, T. Rebedea, and University Politehnica of Bucharest, "Seeking an Empathy-abled Conversational Agent," in RoCHI - International Conference on Human-Computer Interaction, 2020, pp. 103–107. doi: 10.37789/rochi.2020.1.1.16.

[127] J. Kim, S. Oh, O.-W. Kwon, and H. Kim, "Multi-Turn Chatbot Based on Query-Context Attentions and Dual Wasserstein Generative Adversarial Networks," Applied Sciences, vol. 9, no. 18, p. 3908, Sep. 2019, doi: 10.3390/app9183908.

[128] N. Asghar, I. Kobyzev, J. Hoey, P. Poupart, and M. B. Sheikh, "Generating Emotionally Aligned Responses in Dialogues using Affect Control Theory," arXiv:2003.03645 [cs], Apr. 2020, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2003.03645.

[129] "pandorabots." www.pandorabots.com.

[130] M. Canonico and L. D. Russis, "A Comparison and Critique of Natural Language Understanding Tools," CLOUD COMPUTING, p. 7, 2018.

[131] "openai." https://openai.com/api/.

[132] "Learning Multilingual Subjective Language via Cross-lingual Projections," p. 8.

[133] A. B. Kocaballi, L. Laranjo, and E. Coiera, "Understanding and Measuring User Experience in Conversational Interfaces," Interacting with Computers, vol. 31, no. 2, pp. 192–207, Mar. 2019, doi: 10.1093/iwc/iwz015.

[134] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09, Boulder, Colorado, 2009, p. 19. doi: 10.3115/1620754.1620758.

[135] A. Zouaghi, L. Merhbene, and M. Zrigui, "Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation," Artif Intell Rev, vol. 38, no. 4, pp. 257–269, Dec. 2012, doi: 10.1007/s10462-011-9249-3.

[136] C. Lhioui, A. Zouaghi, and M. Zrigui, "A Rule-based Semantic Frame Annotation of Arabic Speech Turns for Automatic Dialogue Analysis," Procedia Computer Science, vol. 117, pp. 46–54, 2017, doi: 10.1016/j.procs.2017.10.093.

[137] C. Lhioui, A. Zouaghi, and M. Zrigui, "The Constitution of an Arabic Touristic Corpus," Procedia Computer Science, vol. 142, pp. 14–25, 2018, doi: 10.1016/j.procs.2018.10.457.

[138] B. A. Shawar and E. S. Atwell, "Using corpora in machine-learning chatbot systems," IJCL, vol. 10, no. 4, pp. 489–516, Nov. 2005, doi: 10.1075/ijcl.10.4.06sha.

[139] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II, Las Vegas, NV, USA, 2005, pp. 152-157 Vol. 1. doi: 10.1109/ITCC.2005.90.

[140] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "An application of pattern matching stemmer in arabic dialogue system," in KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, 2011, pp. 35–43.