



Arabic Chatbots Challenges and Solutions: A Systematic Literature Review

Soufiyan Ouali¹^{*}, Said El Garouani²

¹ Department of Computer Science, Faculty of Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco

² Department of Computer Science, Faculty of Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco

*Corresponding Author: Soufiyan Ouali

DOI: <https://doi.org/10.52866/ijcsm.2024.05.03.007>

Received February 2024; Accepted April 2024; Available online June 2024

ABSTRACT: Since the beginning of Natural Language Processing, researchers have been highly interested in the idea of equipping machines with the ability to think, understand, and communicate in a human-like manner. However, despite the significant progress in this field, particularly in English, Arabic research remains in its early stages of development. This study presents a Systematic Literature Review on challenges faced in Arabic chatbot development and the proposed solutions. Utilizing the search terms "ARABIC," "CHATBOT," "CHALLENGES," and "SOLUTION," (including synonyms) we systematically surveyed studies published between 2000 and 2023 from Scopus, Science Direct, Web of Science, PubMed, SpringerLink, IEEE Xplore, ACM, Ebescos, and ICI. Moreover, besides Google Scholar and ResearchGate, we employed a manual snowballing technique to discover supplementary relevant research by examining the references of all chosen primary studies. The included studies were assessed for eligibility based on the quality assessment checklist we developed. Out of 3,891 studies, only 64 were deemed eligible. Many challenges were identified, including the scarcity of well-structured datasets. To overcome this (n=35) studies manually collected and preprocessed data. Additionally, Arabic language complexity led (n=53) researchers to adopt pre-scripted rules approaches, followed by generative approaches (n=10), and hybrid approach (n=1). Furthermore, (n=27) studies employed human-based evaluation metrics to assess the chatbot performance. while, (n=11) studies haven't used any metrics. Based on conducted research, a critical research priority is providing Arabic with high-quality resources, such as an Arabic dataset that includes dialectal variations and incorporates empathy, lexicon corpora, and also a word normalization library. These resources will enable the chatbot to interact more naturally and humanely. Additionally, hybrid approaches have shown promising results, particularly in low-resource languages, such as Arabic. Therefore, more focus should be dedicated into implementing hybrid approaches in chatbot development. Furthermore, evaluating the chatbot performance is still an open domain for further research and contribution, highlighting the need for innovative standardized evaluation methods.

Keywords: Arabic Dialogue System, Arabic Conversational Agent, Chatbot, Arabic Natural Language Processing.

1. INTRODUCTION

Chatbot is an intelligent program that attempts to simulate conversations between humankind and machines in a human-like manner [1]. Due to its availability 24/7, quick response, cutting operational costs, and can be implemented in many industries such as education [2, 3], health, and tourism; automating services such as booking flight tickets or hotel reservations all through conversation [9] makes it an appealing solution for organizations [4]. According to an article in Forbes [5] the chatbot market is expected to reach \$102 billion by 2026.

Chatbots have been in use for many years, their development has evolved from utilizing traditional techniques such as rule and retrieval-based models (i.e., predefined rules and scripted patterns) into AI-powered and deep-learning models. For instance, GPT-3 and GPT-4 [6, 7, 107], use much more advanced methods such as sequence to sequence, attention, and transformers model that enable the chatbot to generate responses that mimic the conversations in a human-like manner [8]. Furthermore, the rapid advancement of this field has become a significant concern for researchers. Chatbots have now reached a level where their outputs are often indistinguishable from those generated by humans, posing challenges in discerning between the two [16]. However, this advancement has been primarily in English. To date, the developed Arabic chatbots (AC) are still in the early stages of development [9] [15], the majority are utilizing traditional techniques, as illustrated in Figure. 10 more than 83% of the available AC are based on rule and retrieval approaches. Moreover, the majority are built to execute simple tasks such as Question Answering. This poses the question: Why hasn't the AC research field kept up with the progress? It is therefore imperative to study the state-of-the-art of the developed AC and report the findings to the research community to further the enhancements of this field.

Although there are several reviews on AC available in the literature, their objectives are distinct from the focus of our study. No study to date has neither investigated in-depth the challenges encountered in the development of AC nor stated and investigated the solutions proposed by researchers. In particular, some reviews investigate only the implementation approaches used to build a chatbot, but others investigate only the Arabic language processing challenges. While others examined only a few researches. For instance. In [10], *S. AlHumoud et al*, presented a survey on 12 studies only, also the study by *Abdulkader et al* [11], presented different approaches used to build a chatbot, but they didn't include all the studies. Also, *A. Ahmed* [12], conducted a scoping review of AC, however, they did not cover some significant databases such as Web of Science, Scopus, ERIC, DOAJ, and JSTOR. Furthermore, the study's scope presented by *A. Fuad et al* [15] is limited to articles published between 2018 and 2021, thereby excluding numerous important and informative articles. Moreover, *E. AlHagbani et al* [13], and *Almurayh* [14], conducted surveys that explored the challenges and barriers that hinder the implementation of AC without reviewing the published articles, only basic recommendations for implementing an AC were highlighted in [14]. Additionally, there is only one systematic literature review by *Abeer et al* [9], this study comprehensively presents and compares existing English and Arabic chatbots, providing detailed statistics and discussing their methods and evaluation techniques, in addition, the study highlights the challenges facing chatbots, particularly AC. Despite that, the information provided is limited to a superficial level, and the study doesn't investigate in-depth the challenges encountered in Arabic chatbot developments. Our study, in contrast, distinguishes itself by covering all the existing articles and also analyzing each article in-depth, assessing the problems addressed, challenges faced, solutions proposed, evaluation metrics employed and openly discussing the implementation approaches used.

Considering the above factors and the research gap, the objective of this paper is to conduct a comprehensive, rigorous systematic literature review (SLR) of all the relevant studies that investigate the development of AC. Our methodology involves employing the PRISMA approach to identify, categorize, and present our findings on various aspects of the AC field. Additionally, we aim to provide an in-depth analysis review by comparing and contrasting the overall characteristics of the studies, such as the techniques used, the domain of study, datasets, the challenges faced, and solutions proposed by each article. Finally, Open discussion recommending solutions for Arabic Chatbot challenges.

To organize this SLR, our study contributes to this domain by seeking to answer the following research questions:

- **(RQ1)** What is the current state of the Arabic chatbot research field?
- **(RQ2)** What challenges are encountered in the development of Arabic chatbots?
- **(RQ3)** What solutions have been proposed to address these challenges?
- **(RQ4)** Are the proposed solutions effectively addressing the gap?
- **(RQ5)** What are the recommendations for enhancing the development of this field?

The remainder of this paper is structured as follows: *section 2* presents the background information on chatbots with an overview of Arabic language features and challenges. *section 3* details the methodology of this SLR and the different phases involved. *section 4* presents the findings of the study. *section 5* presents a discussion of the results. In *section 6*, the conclusion, limitations, and potential areas for further research are presented.

2. BACKGROUND INFORMATION

This section has two main objectives. Firstly, it presents an overview of chatbots by providing a brief historical account of their evolution, it describes the general architecture of chatbots, it differentiates between various types and classifications, and discusses the approaches employed for deploying them. Secondly, it provides a thorough in-depth overview and analysis of the Arabic language features and challenges that obstruct deploying an intelligence AC.

2.1 OVERVIEW OF CHATBOT

2.1.1 DEFINITION, TERMINOLOGY AND HISTORY

Chatbot, conversational agent, Virtual agent, Virtual assistant, Bot, or Dialogue system all refer to the same technology [24] which can be identified as a program that has the ability to interact and exchange in a human-like manner whether answering a question, executing a task or just chatting. All these tasks can be done by understanding an input and providing an output, whether it is textual, voice or humanoid mode. According to *Crockett et al* [17], Chatbot is defined as having “the ability to reason and pursue a course of action based on its interactions with humans and other agents”.

The earliest invention of this technologies can be traced back to the 1950s [23]. One of the scientists who laid the cornerstone for the beginning of this field is *John McCarthy*, who is considered the father of Artificial intelligence, decided with the help of other scientists to organize a group called “Dartmouth Workshop” to clarify and develop ideas about thinking machines which marked the founding event of both AI and NLP [18][19]. They discussed the possibility of creating machines that could use language effectively leading to the emergence of NLP as a subfield of AI. This contribution was followed by the work introduced by *Georgetown University and IBM Company* in 1954 [20], the

work of *Chomsky* in 1957 who introduced the syntactic structures [19], the work of *Daniel Bobrow* in 1964 who developed the STUDENT program [21], the work of *Joseph Weizenbaum* in 1965 who built ELIZA [22], and the evolving of this technology continues until now [23, 24].

2.1.2 CATEGORIES OF CHATBOTS

Chatbots can be classified based on multiple variables and aspects as represented in Figure 1. As for *Shawar et al* [25], they categorized Chatbots depending on the approaches deployed that are whether manually written rules or automatically learning patterns. According to *Hussain et al* [28] Chatbot can be classified into two categories: task-oriented or non-task-oriented, while *K. Moore et al*, [26] and *K. Ramesh et al* [27] found that chatbots based on their response generation method as rule-based, retrieval-based, and generative chatbots. Another study of *A. Fuad et al* [29] stated that a Chatbot can be a Question-Answer agent, task-oriented, or simply a chatbot. However, other studies argued that a Chatbot can be also task-oriented [30,31] or a Question-answering agent [40]. In our viewpoint, the primary characteristic that categorizes chatbots is the training data or the Knowledge domain (i.e., open domain or closed domain). Based on that, we can build either a Rule/retrieval-based chatbot or generative chatbot that can work as an informative agent by answering users’ questions, a task-oriented agent that can execute many tasks such as a flight ticket booking or hotel reservation, and also it can be for the purpose of chatting only (e.g., psychotherapist chatbots ELIZA).

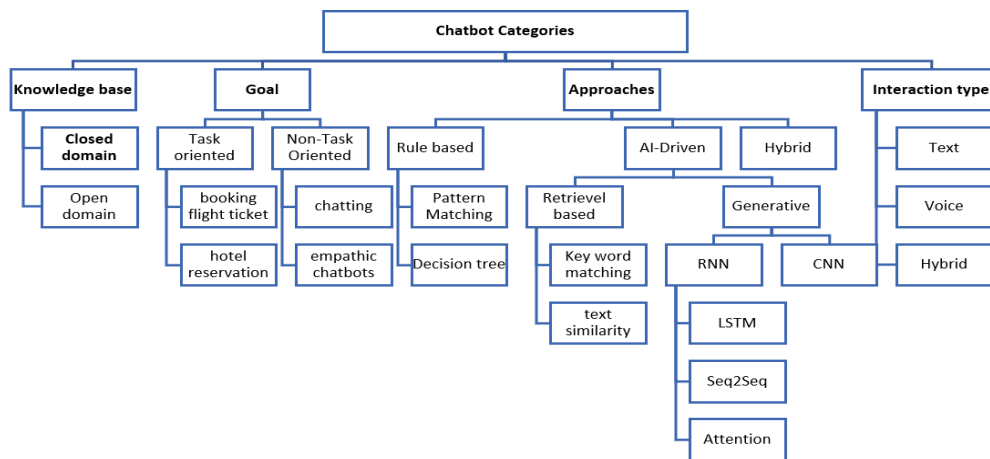


FIGURE 1. Chatbots Classification

2.1.3 CHATBOT APPROACHES

Different approaches were used to develop chatbots in the selected articles as illustrated in Table 3. These approaches can be categorized into three main categories: Pre-Scripted Rules, Generative and Hybrid approaches.

2.1.3.1 PRE-SCRIPTED RULE

Pre-scripted rules are generally built with the aim of matching the user’s utterances to predefined rules and patterns to generate responses or process the user’s query to select responses from a predefined set or database of responses. These types of approaches can be referred to also as Rule-based (e.g., ELIZA [22] and ALICE [33]) or Retrieval-based (e.g., Alexa, Siri, and Google Assistant). Chatbots that adopt these approaches generally don’t generate new responses, and they are usually used for tasks as informative agents via answering users’ questions or task-oriented agent. For instance, booking a hotel room or booking a flight ticket, etc. [34, 35]. Additionally, most of the answers generated by these chatbots are based on the current question and no responses are stored for further utilization [24], they rarely use memory, but only for storing basic information like names or ages. Different algorithms and techniques are used in building a Pre-scripted Rules-based chatbot, such as Pattern Matching, chatscript, word similarity and keyword matching. Furthermore, it is worth noting that in retrieval-based chatbots NLP techniques and Neural Networks (NN) can be integrated to enhance their performance and capabilities. Additional details can be found in [23, 24, 36, 37, 143].

2.1.3.2 GENERATIVE APPROACHES

Generative approaches are designed to use advanced AI and ML techniques to generate responses to user queries via extracting the input from the user’s utterance, processing it using NLP techniques [38], and deploying Natural language understanding (NLU) approaches [29] to understand and extract the context of the user’s query. Furthermore, they employ highly advanced algorithms to achieve a high level of natural language generation (NLG) [40]. These algorithms can generate responses that go beyond the training data and simulate human-like conversations (e.g., ChatGPT). Generative chatbots take into account the entire dialog context rather than just the current turn and generally do not rely on predefined responses for every potential user input, typically they necessitate an extensive training

dataset [23]. The different algorithms used to develop a generative chatbot are NLP, NLU, NLG, NN, Artificial Neural Networks (ANNs), Recurrent Neural Networks (RNN), Bert models, LSTM, Sequence-to-Sequence (Seq2Seq), Bidirectional Recurrent Neural Networks (BRNN). More detailed information about these algorithms can be found in [23, 37, 41]

2.1.3.3 HYBRID APPROACHES

Hybrid chatbots integrate features from both pre-scripted rules and Generative approaches. By employing a merge of pre-established responses and generative models, they provide users with a chatbot that is comprehensive and capable in various aspects and also adaptable to various contexts. The objective of hybrid chatbots is to capitalize on the advantages offered by both approaches while minimizing their respective limitations, it helps maintain the conversation context which is not available in the rule-based model and also can answer various questions that have the same meanings which is not available in the machine learning based model [43]. Moreover, one of its main advantages is it enables the chatbot to use AI and ML models however the training data is not sufficient [42].

2.1.4 CHATBOT ARCHITECT

The general architecture of a Chatbot can be categorized into three main components: *User interface*, the platform or system through which users engage with the chatbot, which can take various forms such as a mobile app, website or even a robotic interface. It can be either text-based, voice-based or hybrid. However, the underlying concept remains the same (the user enters a query or command, and the chatbot processes it to produce an output or response). *Knowledge base* serves as Chatbot's central repository of information, it contains the data, patterns and knowledge that the chatbot needs to learn or extract answers from. *Chatbot engine*, houses the model and approaches utilized for generating responses to the users' utterances. Depending on the chosen approach the engine applies relevant techniques to understand and respond to user input effectively.

2.2 OVERVIEW OF THE ARABIC LANGUAGE FEATURES AND CHALLENGES

Despite the vast number of Arabic language speakers, exceeding 400 million, and the increase in the number of Arabic internet users is notable, reaching more than 260 million, integrating Arabic in much-advanced algorithms and models is still in its infancy stages due to many challenges. In the following section, we present the various features that characterize the Arabic language.

1. Arabic language belongs to a family of Semitic languages, it can be represented in three types: Classical Arabic (CA), the language of the Quran and classical Islamic literature i.e., Hadith and Seera, sayings and actions of Prophet Muhammad, it is commonly used in classical poetry, literary, and historical contexts. The Modern Standard Arabic (MSA), is understood by all Arabic speakers and it is used usually in formal contexts such as academic research and written communication. Additionally, there is Colloquial Arabic or dialectal Arabic, a group of regional spoken varieties of the Arabic language, it depends on the region in which it is used, each region has a distinct dialect and they can differ greatly in terms of vocabulary, pronunciation, and syntax, as shown in Supplementary Figure 3. These three distinct linguistic varieties of the Arabic language are different mainly in style and vocabulary. For instance, the sentence "The mighty solid stone" which is written in CA as "كجلمود صخر" would be written in MSA as "الحجر العظيم الصلب" and in dialect (Moroccan) "حجرة قاسحة". Therefore, to build an efficient chatbot these varieties should be taken into account, the chatbot should have the ability to distinguish between them and generate responses according to the user's type of Arabic, which can be very expensive in terms of equipment, resources and time to train the model. [45].
2. Arabic is a language that employs diacritics, small markings or accents that are added to letters to significantly alter the meaning, context, pronunciation and grammatical clarification of a word, any change in diacritics will subsequently change the whole meaning of the word Supplementary Figure 1 (a) illustrates that one word "سجل"–registered– can have at least 16 different nouns and verbs based on diacritics iteration [44, 53]. Moreover, MSA generally does not use diacritics in its writing. Therefore, extracting the word context and sense chatbot will require a high degree of homograph resolution and word sense disambiguation [49].
3. Arabic consists of 28 letters that differ morphologically and phonologically. Some letters share the same character as represented in supplementary Figure 1 (b), for example, some characters can have one or two dots below them and some others can have one to three above them, with each iteration we can have a different letter and sound [52]. Moreover, Arabic letters take different shapes based on their position in the word as illustrated in the supplementary Figure 1 (b). Thus, the chatbot specifically generative ones should be able to recognize the different shapes of each letter which makes the training data not only large but also complex.
4. Arabic language has flexible word order, in addition to the regular order of the sentence (verb, subject and object) (VSO) a sentence can be also presented in the form of (SVO) or (OVS), for example, the sentence "Ahmed eats an apple" can be written like "احمد اكل التفاحة" (SVO) or "اكل احمد التفاحة" (VSO)-eat Ahmed an apple- or "التفاحة اكلها احمد" (OVS)-the apple was eaten by Ahmed- [44, 53, 54]. In addition to the free order Arabic is a pro-drop language, the pronoun can be absent in Arabic sentences which may cause many problems in any syntactic parser or analyzer that aims to know whether the sentence includes the pronoun or no [49]. Moreover, Arabic is considered a clitic

language with complex and rich grammatical structure i.e., Arabic sentences can be structured with no verb and also full sentences can be presented by a single word, for example, the sentence “سيكتبونها”-they will write it- even though it's a single word it has (SVO) [54]. Nevertheless, Arabic words don't use capitalization which makes extracting named entities even difficult. Consequently, applying techniques such as named entity recognition (NER) and Part-of-speech tagging (POS) becomes highly challenging or nearly impassable [55].

5. Arabic has a rich morphology and it is a derivative language, one root can create many words with different contexts and meanings [28, 53]. Additionally, Arabic words are inflected with a variety of characteristics such as gender, number, voice and person [50, 56], for instance, changing the gender of the known is causes changing the verb and other sentence words [46, 47]. Therefore, deploying an effective stemmer is a challenging endeavor, even if the root or stem is extracted successfully there is a high probability of losing the sentence's context and sense which may increase ambiguity. Consequently, NLU may not achieve successful results [57].
6. Arabic chatbot research is known for its limited resources [51]. The Arabic Wordnet (AWN) which is a lexical database for the Arabic language, includes a set of words with their definitions, meanings, synonyms and antonyms; it plays a crucial role in semantic relations, word sense disambiguation and in NLP and AI applications. However, AWN, in its current state, is insufficient for the needs of the Arabic language in comparison with other languages such as English [48,155]. Therefore, the level of ambiguity (Syntactic ambiguity, Semantic ambiguity, Constituent boundary ambiguity, Anaphoric ambiguity) poses a significant challenge for building an efficient Arabic chatbot. Additional details can be found in [49, 58].
7. High probability that the system might not recognize word misspellings in a conversation [58], due to the reasons mentioned above. For instance, in the sentence “أكل التفاح”- “ate the apple”- if the user forgets the letter “أ”-/a/- the sentence will turn into “كل التفاح”, which can have two meanings either “all the apples” or become an imperative sentence “eat the apple”.
8. Political and economic powers have played a significant role in shaping the linguistic panorama of numerous Arabic-speaking nations [60]. As a result, many Arabic speakers are bilingual, thus they often incorporate loanwords from other languages into their speech and also may switch between languages, such as Arabic and another language (e.g., French), within the same conversation or sentence, in Moroccan communication, for instance, speakers frequently merge Arabic and French within a single sentence. (e.g., they may express "I am okay" by saying "انا cava."). Therefore, building an effective chatbot should involve the ability to understand sentences even when they are mixed with different languages.
9. the writing direction which starts from the right to the left can also cause a problem, Due to many models being built to serve left to right languages [59].

3. RESEARCH METHODOLOGY

In order to achieve the objective of this paper, we conduct a SLR because it is considered as an original work, that is implemented using rigorous methodological approaches [61, 62]. This SLR was performed in accordance with the protocol guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [63].

3.1 INFORMATION SOURCES

This study conducted a comprehensive search for scientific journal articles by consulting various databases (i.e., Scopus, Science Direct, Web of Science, PubMed, SpringerLink, IEEE Xplore, ACM, Ebescos, and ICI) which yielded sufficient research material for this investigation. Moreover, other data sources such as Google Scholar, ResearchGate.net, and Doaj.org were also examined. The search criteria encompassed publications in journals, conference articles, and open publications including doctoral theses and dissertations that have the potential to contribute significantly to this study. Furthermore, we employed a manual snowballing technique to discover supplementary relevant research by examining the references of all chosen primary studies. The search process was initiated on November 20th, 2022 and continued until April 1st, 2023.

3.2 SEARCH STRATEGY

3.2.1 SEARCH CONCEPTS AND TERMS

In the process of conceptualizing our search, we initiated by selecting the title of our work as "ARABIC CHATBOT CHALLENGES AND SOLUTION", from which we identified four main concepts: "**Chatbot**", "**Arabic**" language processing, "**Challenges**" encounter building Arabic chatbots, and Proposed "**Solutions**" to overcome these obstacles. To ensure comprehensive coverage of the literature, we conducted a thorough search for all possible synonyms of each concept, thereby minimizing the likelihood of overlooking any relevant articles.

3.2.2 SEARCH PROCESS

In order to enhance the likelihood of identifying studies that are of utmost relevance, we employed logical operators **AND** & **OR** to conjoin the keywords obtained from the Concept and Term procedure. Moreover, the search covered the

title, abstract, and keywords to guarantee the inclusion of pertinent studies. The resulting search query implemented for the present review is as follows (regarding the query structure of each database):

("chatbot" OR "Virtual agent" OR "Conversational agent" OR "Chatterbot" OR "virtual assistant" OR "digital assistant" OR "chat-agent") AND ("Arab" OR "Arabic" OR "Arabic language") AND ("Challenge" OR "Difficulty" OR "Obstacle" OR "problem") AND "Solution").

With the intention to procure a more comprehensive range of references pertaining to the research topic, certain articles were obtained directly from the query. (“Arabic chatbot” or “Arabic conversation agent”). Additional articles from various sources including Google Scholar, ResearchGate, and Google search engine were included to provide background knowledge for the research, Master and doctoral thesis papers were also searched for potential contributions. Furthermore, we used a manual snowballing method to identify additional relevant studies by exploring references of all selected primary studies. Due to the rarity of Arabic chatbot articles we had to extend our years range covering all articles from 2000 to 2023.

3.3 ELIGIBILITY CRITERIA

In the process of choosing articles for review, we identified key criteria that determine whether the article will be included or excluded. Firstly, all studies related to the development of chatbots that can communicate in Arabic were included whether it is textual, vocal or humanoid. Secondly, because of the rarity of Arabic contributions we had to widen our search range, including peer review articles, conference papers and book chapters, and our year range, we included papers between 2000 and 2023. Thirdly, we eliminated studies lacking abstracts or with unavailable full text. Finally, to eliminate the bias that may result from poor translation only articles published in English language were included.

3.4 SELECTION PROCESS

For the selection process task, we used Zotero [65] as a Citation and research management tool due to its compatibility with many databases and library resources, accessible by internet browser and it can be integrated with Microsoft Word and other word editors [66]. After downloading all the articles, we organized and saved the work then we used Zotero feature to detect and delete all the duplicated articles, additional duplicates were deleted manually through the screening phase. After discussing the eligibility criteria and deciding the main objectives of this step, the screening phase was performed by two authors (SO) and (SEG). The screening process consisted of three phases: title screening, abstract screening, and full-text screening. In order to perform a rigorous SLR we created inclusion criteria for each phase of the screening, as presented in Table 1. After conducting the title and abstract screening, the full texts of all potentially eligible papers were then retrieved and reviewed independently by the two authors to determine if they met the inclusion criteria. Any disagreements regarding inclusion and classification were discussed until a consensus was achieved. The eligible articles were identified using the PRISMA flow diagram presented in Figure. 2.

Table 1. Inclusion Criteria for each phase of the Screening.

Screening by title	Screening by abstract	full-text Screening
Inclusion Criteria:	Inclusion Criteria:	Inclusion Criteria:
<ul style="list-style-type: none"> • Any article that mentions the development of Arabic chatbot • Any title contains one of the concept’s terms in the search strategy • Title highlighting Natural Language Processing with Arabic Language 	<ul style="list-style-type: none"> • Well Structured abstract • describes the chatbot-building process • highlight the challenges encountered and solutions proposed • mention the dataset used • mention evaluation metric performed to assess the chatbot performance. 	<ul style="list-style-type: none"> • Describes the chatbot-building process • Suggest solution to overcome challenges encounter developing Arabic Chatbot

3.5 DATA COLLECTION PROCESS AND DATA ITEMS

The first and second authors independently extracted data from the articles ultimately selected for inclusion in the review. Any disagreements regarding relevant data were discussed until a consensus was achieved. The eligible selected articles were analyzed in-depth one by one, and the relevant information extracted from each article are the following: a). DOI. b). the authors, journal name, the title of the article and the study area. c). the year of

publication. d). the challenges identified in each study. e). the approach or technique used or solution proposed. f). the data set used in the training phase. g). the evaluation metric used and the resulting score.

3.6 QUALITY ASSESSMENT CHECKLIST

To assess the risk of bias in the included studies, we developed a quality assessment checklist consisting of thirteen questions, as presented in Table 2, these questions represent what elements are important the most in our study for the data extraction phase.

For each question on the assessment checklist, two answers were given ('1' or '0'). Considering Q10, the answer depends on the number of citations; if it is above 5, the answer will be 1; otherwise, it is 0. The quality of each article was determined based on the resulting score percentage obtained, which is calculated as follows: the sum of the answer's score * 100 / the total number of questions (n=13). We decided that the minimum threshold value for articles with good quality is 75%. While the preferred threshold is 85%, due to the scarcity of Arabic research and articles, we had to lower our threshold score. However, it remains acceptable. Therefore, if the score obtained is equal to or above 75%, the article was considered high quality. If the score obtained is between 50% and 75%, the article was considered medium quality. Otherwise, the article was considered low quality. The authors (SO, SEG), independently evaluated the quality of the studies. Any disagreements regarding article score were discussed until a consensus was achieved

Table 2. Quality assessment checklist

#	Question
Q1	is the study relevant to our research?
Q2	Is the study present a new method or solution to develop an Arabic chatbot?
Q3	Is the problem statement clear?
Q4	Is the Experiment setup clearly explained?
Q5	Is the dataset used mentioned in the work?
Q6	Are the evaluation techniques used explained?
Q7	Are the results compared to previous studies, or no?
Q8	Is the conclusion explained clearly and linked to the purpose of the study?
Q9	Is the source of the article credible (published in a ranked venue)?
Q10	Has the study been cited in other publications?
Q11	is the presentation of the work well-formed and clear?
Q12	Are the challenges of Arabic processing been highlighted?
Q13	Is the future work mentioned?

4. RESULTS.

4.1 STUDY SELECTION

During this phase, we utilized the PRISMA framework [131] to apply the inclusion-exclusion criteria and screen the collected articles for eligibility.

To begin, we initiated our data collection process by gathering articles extracted from various databases, by which we identified (n=3,891) studies. Subsequently, we eliminated duplicate articles and merged those with similar content to reduce redundancy, resulting in the removal of (n=804) articles. Following this, we conducted a title and abstract screening to eliminate irrelevant data. Out of the remaining articles, (n=2912) were excluded from consideration, as they did not meet our inclusion criteria. Next, we performed a full-text screening of the remaining articles (n=175) to assess relevance and eligibility. We utilized the quality assessment checklist that we developed (Table 4) to further filter articles, ultimately selecting the most relevant and efficient articles for our systematic review, resulting in the inclusion of (n=64) articles. The screening process steps are presented in Figure. 2.

4.2 STUDY CHARACTERISTICS

4.2.1 SOURCE OF THE ARTICLES

Figure. 3 represents the distribution of studies by source type. Most of the articles included in the study are peer-reviewed articles (56%). Considering the scarcity of Arabic publications, we had to include also conference papers (39%) and book chapters (5%). However, during the quality assessment phase, all the included studies were carefully verified to minimize the risk of bias.

4.2.2 PUBLICATION YEAR

The selected studies were published between 2002 and 2023 as represented in the Figure. 4, with most of them being published between 2018 and 2022. The reason for this can be explained by the existence of several platforms that facilitate the development of chatbots in Arabic such as Pandorabots [132] and Rasa [133].

4.2.3 REGION OF STUDY

Figure. 5 represent the distribution of studies based on their respective regions. These studies were conducted in a total of 12 countries worldwide, most of the studies originated from Saudia Arabia (n=19) followed by Jordan (n=17), the contributions of these two countries constitute (55%) of the studies. The remaining regions, a total of 10, collectively contribute 45% of the studies.

4.2.4 DOMAIN OF STUDY

Figure. 7 shows that most of the studies were closed-domain with a percentage of (72%). Although there is a wide array of open-domain datasets accessible; the chatbot can be trained on various types of text corpora such as movie corpus or a novel or news-paper which are available. However, due to the complexity of deploying advanced models with Arabic content, the integration of open-domain methodologies in Arabic chatbots remains relatively scarce, accounting for only (28%) of the overall coverage. Out of (72%) of the closed domain studies, (41%) of the studies have been developed for educational purposes as illustrated in Figure. 6. A possible explanation might be because of the simplicity of building their own dataset, all they need is the campus and student information which can be gathered through a survey or a questionnaire (Table 3).

4.2.5 CHATBOT LANGUAGE

Figure. 8 represents the distribution of studies based on chatbot languages, as mentioned before Arabic language can be expressed in three distinct ways; CA, MSA and dialectal. The distribution shows that (69%) are based on CA and MSA datasets, (11%) of chatbots are based on dialectal datasets, and (20%) are multilingual supporting Arabic and other languages. The wide adaptation of CA and MSA can be explained by two reasons. Firstly, considering them a universal language, it can be understood by all Arabic speakers. Secondly, as illustrated in Figure. 5 most of the research are from Saudia Arabia where the official language is CA and MSA, also the fact that dialectal datasets are extremely rare.

4.2.6 CHATBOT INTERACTION TYPE

Figure. 9 presents the distribution of studies by chatbot Interaction mode. A chatbot may be text-based, voice-based or multimodal. Most of the studies are developing text-based chatbots (90%), plausibly due to the increased use of messaging technologies. Additionally, the availability of text-based datasets is more abundant compared to speech datasets.

4.2.7 APPROACH USED

Figure. 10 presents the distribution of studies by Chatbot approach. The majority of studies (83%) were developing chatbots based on Rule or Retrieval approaches, (16%) were based on Generative approaches and only (1%) were built using hybrid approaches. The bias can be elucidated by the findings presented in Figure. 6 and Figure 7, where a significant majority (72%) of studies focused on closed-domain applications, these types of chatbots are designed to fulfill specific and modest tasks such as hotel reservations or provide simple answers, therefore retrieval-based approaches prove to be the most suitable for fulfilling these types of services.

4.2.8 EVALUATION METRICS

Evaluating the performance of a developed chatbot is essential, and the evaluation metrics play a crucial role in this process. However, evaluating chatbots, particularly those designed for the Arabic language is still a challenging task. Figure. 11 presents the evaluation metrics used in the selected articles which can be classified into two categories, human-based (manually) and automatic-based metrics. 42% of the studies used manual approaches through conducting conversation logs or surveys about user feedback. 41% of studies were using automatic metrics (30% were using classical automatic metrics such as accuracy, precision and recall, and 11 % were using advanced automatic metrics such as blue score and Perplexity). 17% of studies didn't use any metric However, Chatbot evaluation metrics are still critical and require more research (more detailed information can be found in section 6). The characteristics of the included studies are detailed in Table 3 and Table 4.

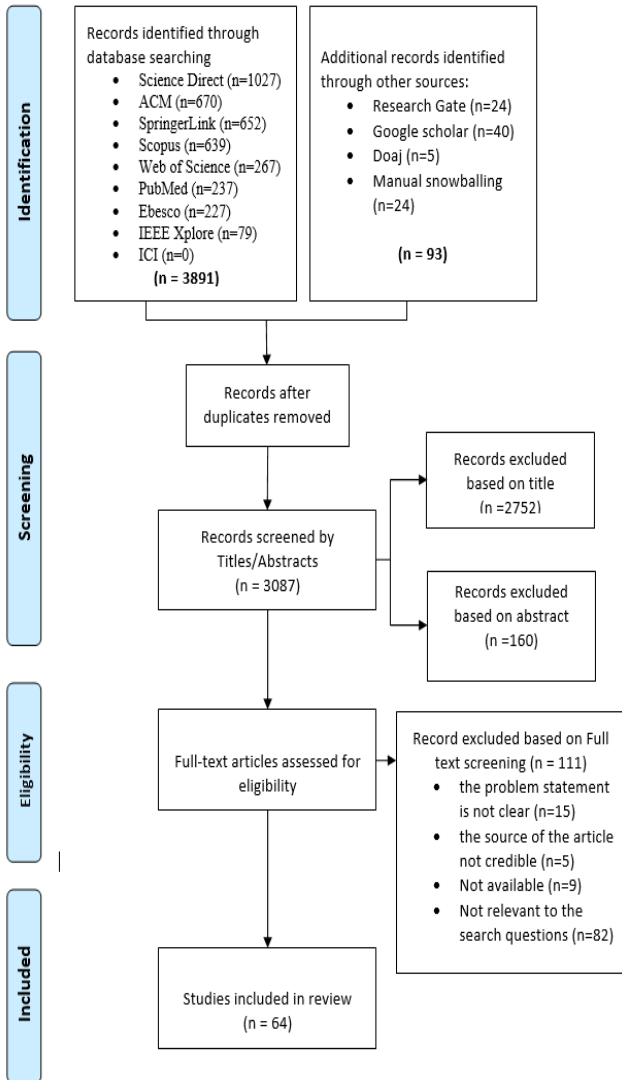


FIGURE 2. Prisma Flowchart.

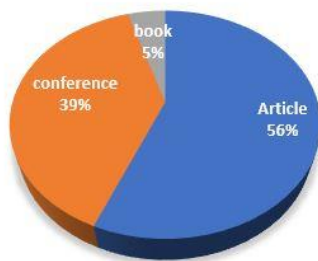


FIGURE 3. Distribution by article's source.

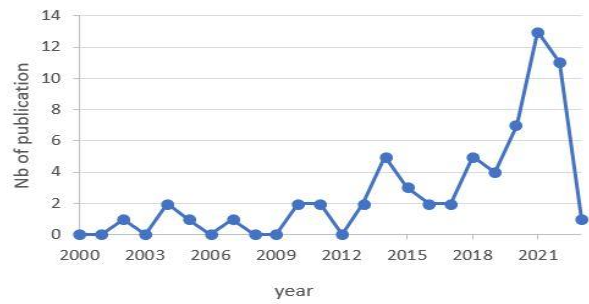
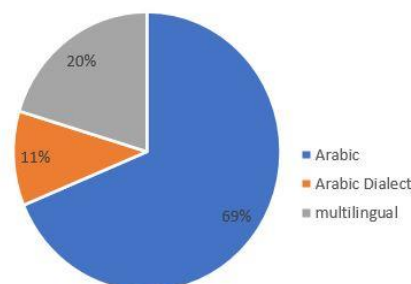


FIGURE 4. Distribution of studies by publication year.

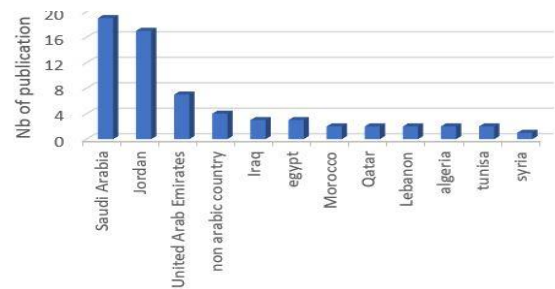


FIGURE 5. Distribution by region of study

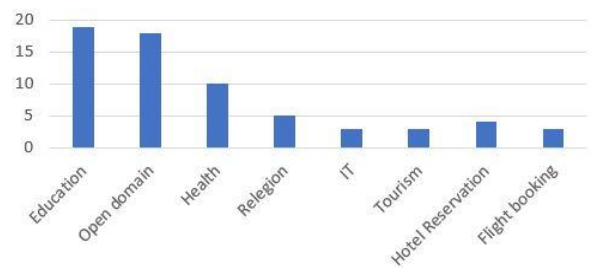


FIGURE 6. Distribution by domain of the study.



FIGURE 7. Distribution By Domain-based Approach

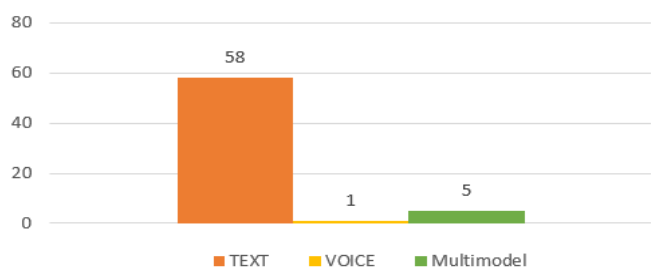


FIGURE 8. Distribution by chatbot language.

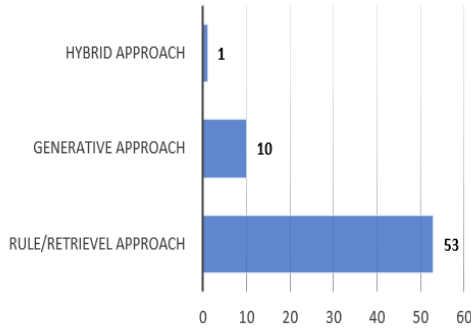


FIGURE 9. Distribution by interaction type.

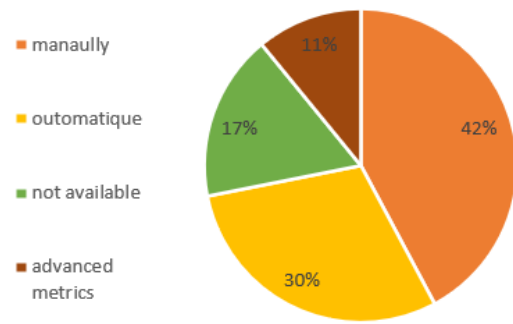


FIGURE 10. Distribution by approach utilized.

FIGURE 11. Distribution based on evaluation metrics.

Table 3. Main characteristics of the included studies (64).

Study Ref	problematic/Challenge	Solution	technique used	Data-set used	Model Performance
[67]	Build an Arabic Chatbot <ul style="list-style-type: none"> Scarcity of Arabic datasets Handle erroneous or misspelled user inputs 	<ul style="list-style-type: none"> Collect data manually Proximity processing with Pool of suggestions 	<ul style="list-style-type: none"> chatterbot Google Colab Twilio to send messages to user 	Not Mentioned (N/M)	Not Available (N/A)
[68]	Build an Arabic Chatbot <ul style="list-style-type: none"> Cover the complexities of the Arabic language Handle erroneous Conserve the context flow Evaluation the Chatbot performance 	<ul style="list-style-type: none"> Using pattern matching PM because it doesn't require intensive processing Utterance validation: to validate the utterance if it is valid or not. Using temporary memory Calculating the ratio of correct answers from the conversation logs 	<ul style="list-style-type: none"> Pattern Matching 	N/M	<ul style="list-style-type: none"> Manually: Ratio of Matched Utterance (RMUT) = 73,56%
[69]	Build an Arabic Chatbot <ul style="list-style-type: none"> Morphological changes occurring on a word through adding affixes to it Handling the user utterance that targets more than one topic inside the same utterance Evaluation the chatbot performance 	<ul style="list-style-type: none"> Using Pattern Matching Rule fire for multiple topics in one utterance to extract all the topics extracted. Text Adapter: used to link the matched rules' responses Conducting conversation logs 	<ul style="list-style-type: none"> Pattern Matching 	N/M	<ul style="list-style-type: none"> Manually: conversation log
[70]	Arabic word generation using Deep learning <ul style="list-style-type: none"> Scarcity of Arabic medical datasets Handling Long Sequences 	<ul style="list-style-type: none"> Collect and process data from Altibbi's platform databases Data augmentation by using n-gram data model 3-gram and 4-gram Using deep neural model LSTM, BiLSTM, and CONV1D 	<ul style="list-style-type: none"> LSTM B-LSTM CONV1D LSTM-CONVD 	<ul style="list-style-type: none"> Altibbi's databases 	<ul style="list-style-type: none"> Accuracy = 60% Loss = 80%

[71]	<p>Chatbot for QA in Arabic</p> <ul style="list-style-type: none"> • Scarcity of Arabic QA datasets • Handle erroneous • Difficulties using Arabic with NLP techniques. 	<ul style="list-style-type: none"> • Collect and process data from the web manually in QA format • Create a default corpus • Keyword matching; The most significant word is the least frequent word, which will have the highest information content 	<ul style="list-style-type: none"> • Pattern Matching • AIML • Keyword matching 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Accuracy = 93% • Recall = 87%
[72]	<p>Arabic Chatbot that provides users with answers related to IT.</p> <ul style="list-style-type: none"> • Scarcity of dataset • Datasets not structured • Multiple keywords are required to accurately retrieve an answer. • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Collect and process data manually • Text preprocessing with NLTK • Keywords matching; Searching for important keywords using AI, and then using those keywords to find matching sentences within the corpus. • Conducting a survey of 6 participant 	<ul style="list-style-type: none"> • NLP • Keyword matching 	<ul style="list-style-type: none"> • Web scraping • Madora data set 	<ul style="list-style-type: none"> • Manually: -user satisfaction : 67%
[73]	<p>Arabic Chatbot that helps teach students Islam by engaging with them in conversations</p> <ul style="list-style-type: none"> • Solve the complexity and ambiguity of processing the Arabic language • Lack of resources, such as an appropriate Arabic Wordnet (AWN) • Handle erroneous • Context flow • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Pattern matching • PM and keyword matching do not require the utterance to be grammatically correct or the input to be complete. • Utterance checking process before keyword matching • Temporal memory • Conducting conversation logs 	<ul style="list-style-type: none"> • Pattern Matching • Keyword matching 	<ul style="list-style-type: none"> • Islamic Database (IDB) • the Arabic Grammar Database (AGDB) • CA Scripts (CAS) • Tutorial Database (TDB) 	<ul style="list-style-type: none"> • Manually: - user satisfaction : 85%
[74]	<p>Develop a seq2seq neural network-based conversational agent that can effectively understand and generate responses in Gulf Arabic dialect</p> <ul style="list-style-type: none"> • Dataset not available and not appropriate for NN model • Build a generative model • Ensure the effectiveness of the generative model and the ability to cover a wide range of topics 	<ul style="list-style-type: none"> • Collect, label and preprocessing data manually • Using Seq2Seq model • Developing a set of templates for the agent to draw upon based on an analysis of the types of queries users were likely to ask the agent 	<ul style="list-style-type: none"> • Seq 2 Seq • Encoder-Decoder 	<ul style="list-style-type: none"> • Collected manually from various online sources 	<ul style="list-style-type: none"> • F1 score • Recall • Precision • Manually
[75]	<p>Build an Arabic chatbot that aids in accessing web information via chat; extracting exact answers from web pages.</p> <ul style="list-style-type: none"> • Difficulties in building models that can handle Arabic • Collecting and preprocessing Datasets • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Train Alice on a new dataset • Use Quran corpus to create AIML files automatically using JAVA and machine learning • Conducting a user satisfaction survey to evaluate the chatbot's performance 	<ul style="list-style-type: none"> • Pandora -bots • Pattern matching • AIML 	<ul style="list-style-type: none"> • Quran 	<ul style="list-style-type: none"> • Manually - user satisfaction
[76]	<p>Build A chatbot in the form of a mobile application that can assist college students.</p> <ul style="list-style-type: none"> • Dataset not available and not oriented toward the chatbot goal • An efficient information retrieval without diving into 	<ul style="list-style-type: none"> • Build their own dataset manually • Use retrieval-based approaches: Text similarity • Use TF-IDF to vectorize data. • Calculate the cosine similarity and Jaccard distance between 	<ul style="list-style-type: none"> • Text similarity • - Cos and Jaccard • TF-IDF 	<ul style="list-style-type: none"> • Manually: - university data 	<ul style="list-style-type: none"> • N/A

	language structure complexity.	the user utterance and questions in the database.	<ul style="list-style-type: none"> • Kotlin • Firebase 	
[77]	Build a Chatbot: "Smart Guidance" that enhances hospitality and tourism services by providing information about the city of Jeddah, Saudi Arabia. <ul style="list-style-type: none"> • Chatbot easy to use and available 24/7 • Dataset not available • Build an efficient model with a small dataset • Cover a wide range of topics not only the predefined in the dataset 	<ul style="list-style-type: none"> • Build Chatbot in a mobile app platform • Build their own dataset manually by using online surveys. • Use RASA as a third-party platform to build the chatbot model. • External API calls: to extract additional information from the web 	<ul style="list-style-type: none"> • RASA • Android studio • API 	<ul style="list-style-type: none"> • Manually: - online surveys • external API <p>N/A</p>
[78]	Enhanced ArabChat represents the culmination of Hijawi's collaborative efforts in constructing ArabChat by amalgamating all of their individual contributions	<ul style="list-style-type: none"> • Using Pattern matching to extract answers • Using a hybrid rule to handle multiple topics in one utterance • Using a decision tree to make a difference between question and non-question utterance. 	<ul style="list-style-type: none"> • Pattern matching • Decision tree • Text adapter • Pandora -bots • Pattern matching • AIML 	<ul style="list-style-type: none"> • same as previous in ArabChat • Manually: -User satisfaction : 70%
[79]	Chatbot that can answer users' utterances with verses from the holy Quran. <ul style="list-style-type: none"> • Datasets collecting • Avoid the problem of morphosyntactic analyses. • Misspelling problems 	<ul style="list-style-type: none"> • Create a Java Program that transcript the Quraan into an AIML file and train Alice on this dataset • support only Classical Arabic as a training language • Creating a default file that works on the concept of least frequent terms is the highest significant element. 	<ul style="list-style-type: none"> • Pandora -bots • Pattern matching • AIML 	<ul style="list-style-type: none"> • Quran • Manually
[80]	Arabic Chatbot that engages with hotel guests and provides responses regarding hotel room reservations and other services. <ul style="list-style-type: none"> • Dataset not available • Engage in a well-structured dialogue • Interpret user responses according to the dialogue context • Evaluating the performance of the chatbot 	<ul style="list-style-type: none"> • Build their own dataset manually • Combines natural language understanding and flexible dialogue control • Using the Government and Binding theory to build GB-parser to understand the structure / the meaning / intuition of user utterance • Dialogue manager to extract the answer / or make an action • Report an experiment with 500 volunteers 	<ul style="list-style-type: none"> • GB theory 	<ul style="list-style-type: none"> • Manually: - hotel information • Manually: - user satisfaction : 93%
[81]	Build a Dialectical chatbot that offers information about the university <ul style="list-style-type: none"> • Dataset not available • handle the character and complexity of the Arabic language • Unknown or misspelling utterance 	<ul style="list-style-type: none"> • Build their own dataset manually • Keyword matching algorithm and a string distance comparison algorithm • Store it in the conversation log and add it later. 	<ul style="list-style-type: none"> • Pattern matching • AIML • word Similarity 	<ul style="list-style-type: none"> • Manually: - College <p>N/A</p>

[82]	<p>Buil Goal oriented Conversational agent to help in passport services</p> <ul style="list-style-type: none"> • Extract the correct answer without diving into Arabic morphology • Extract the context of the conversation • Keep up the flow of the conversation • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Using a pattern matching • Keyword extracting and use of temporal memory • Using a tree engine or scripted knowledge tree • Conducted through a questionnaire 	<ul style="list-style-type: none"> • Pattern Matching • Knowledge Tree 	N/M	<ul style="list-style-type: none"> • Manually -user satisfaction : 95%
[83]	<p>Build first Arabic end-to-end generative model for task-oriented (AraConv)</p> <ul style="list-style-type: none"> • Data set not available • Small amount of Data • Build a Generative model 	<ul style="list-style-type: none"> • Translate the English BiToD dataset • Joint-training: train model with Arabic and another language such as English. • Model based on single multi-lingual Seq2Seq (mSeq2Seq) that uses the pre-trained model mT5. 	<ul style="list-style-type: none"> • Seq2Seq: model mt5 	<ul style="list-style-type: none"> • Arabic-TOD dataset 	<ul style="list-style-type: none"> • joint goal accuracy (JGA) • BLEU metric • API call accuracy • the task success rate • the dialogue success rate • Manually: log checking
[84]	<p>Build a deep learning-based chatbot called ArRASA</p> <ul style="list-style-type: none"> • Chatbot that can understand Arabic using existing platforms 	<ul style="list-style-type: none"> • Tokenization of text • Featurization • Intent categorization • Entity extraction • Using RASA framework as a third-party 	<ul style="list-style-type: none"> • RASA • NLP • NLU 	N/M	<ul style="list-style-type: none"> • Accuracy = 95% • F1 score = 94%
[85]	<p>Building an Arabic Flight Booking Dialogue System Using a Hybrid Approach</p> <ul style="list-style-type: none"> • Dataset not available • Quantity of training data gathered is not sufficiently large. • Keep the flow of conversation and extract correct answers • Complexity of Arabic • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Wizard of Oz technique: interacting with the system manually until having sufficient data • Using Rule-Based approaches. • Using data-driven approaches • Using Wit.ai, as a third-party platform • Conducted through a questionnaire with 21 participants 	<ul style="list-style-type: none"> • Wit.ai • Rule-based • Data-driven 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - User satisfaction
[86]	<p>BOTTA female chatbot, that aims to simulate conversation and connect with Arab users in a Dialectical language</p> <ul style="list-style-type: none"> • Handle the complexity of the Arabic language • Address Arabic challenges in a more controlled setting • Handle misspelling • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Rule-based approaches • Script pattern using pattern matching with AIML • Orthographic normalization to overcome the inconsistent spelling variations of certain characters. • Conducted through a questionnaire with 3 participants 	<ul style="list-style-type: none"> • Pandora -bots • pattern matching • AIML 	<ul style="list-style-type: none"> • Publicly Available under the name: Botta database 	<ul style="list-style-type: none"> • Manually: -user acceptance

[87]	<p>Build a chatbot for the students that can interact in Arabic and English</p> <ul style="list-style-type: none"> • Dataset not available • The limitless dataset/Knowledge base • Need for precise and accurate responses to a specific task and domain • Ability to understand Arabic dialect and Arabic text containing English words inside. • Handle derived words resulting from inflections and affixes • Lack of Python library for lemmatizing Arabic words • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Built Manually • Use Retrieval-based approaches. • Use Retrieval-based approaches. • Scripting patterns manually. • Using Tokenization and lemmatization. • Use ISRI Stemmer to stem the words • Conducted through the conversation logs 	<ul style="list-style-type: none"> • NLP • Machin learning NN model 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually Accuracy = 75%
[88]	<p>Build a generative model using seq to seq model</p> <ul style="list-style-type: none"> • The Need for a representative dataset to build an accurate model. • Handle different topics and generate new answers 	<ul style="list-style-type: none"> • Built and preprocessed manually • Using Seq2Seq approach to build the model 	<ul style="list-style-type: none"> • Seq2Seq 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • GRU accuracy = 85% • LSTM accuracy = 89%
[89]	<p>Buil a generative task-oriented Chatbot</p> <ul style="list-style-type: none"> • build Chatbot with low-quality data • Avoid the cost and time-intensive data collection and annotation 	<ul style="list-style-type: none"> • Cross-lingual transfer learning: knowledge of high-resource languages (English) is transferred into low-resource languages (such as Arabic) • Using the mT5 transformer model 	<ul style="list-style-type: none"> • Transformer model mt5 	<ul style="list-style-type: none"> • Arabic-TOD dataset 	<ul style="list-style-type: none"> • JGA metric • API Acc metric • the dsr metric • blue score • Manually: - user satisfaction
[90]	<p>Arabic chatbot for flight ticket booking services using pattern-matching approaches</p> <ul style="list-style-type: none"> • Problem of large morphological and derivative diversity for an Arabic word • More control in the conversation • Extract correct answers • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Use Khoja stemmer with improving some of its functionality. • Using Pattern Matching • Conducted through the conversation logs 	<ul style="list-style-type: none"> • Pattern Matching • Stemmin g 	<ul style="list-style-type: none"> • N/M 	<ul style="list-style-type: none"> • Manually: - user satisfaction
[91]	<p>Buil an Arabic chatbot</p> <ul style="list-style-type: none"> • Handl large morphological word syntax • Extract the correct answer without diving into Arabic morphology 	<ul style="list-style-type: none"> • Using Arabic stemmer: ISRIS stemming • Use Pattern matching and scoring approach 	<ul style="list-style-type: none"> • Pattern matching • Scoring approach • Stemmin g 	<ul style="list-style-type: none"> • N/M 	<ul style="list-style-type: none"> • F1 score = 65.5%
[92]	<p>Build a multilingual Receptionist Robot that can operate in a human way.</p> <ul style="list-style-type: none"> • Have a humanoid conversation • Bot fails to respond to personal questions 	<ul style="list-style-type: none"> • Use a flat screen as the face of the bot / answer with text and voice • The bot should initiate the conversation or choose the conversation topic. 	<ul style="list-style-type: none"> • Pattern matching • AIML 	<ul style="list-style-type: none"> • N/M 	<ul style="list-style-type: none"> • N/A

[93]	Build empathy-driven generative chatbot <ul style="list-style-type: none"> • The need for high-quality data • Build a generative model 	<ul style="list-style-type: none"> • Translated from English empathy data set • Adopt the Seq2Seq architecture • Encoder-decoder composed of a (LSTM), (Seq2Seq) with Attention 	<ul style="list-style-type: none"> • Attention • Seq2Seq • LSTM 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Perplexity (PPL) = 38,6 • Bleu score = 0,50%
[94]	Develop an Arabic chatbot that is empathetic-driven and generative. <ul style="list-style-type: none"> • Lack of Arabic datasets suitable for building generative module 	<ul style="list-style-type: none"> • they proposed a transformer-based encoder-decoder initialized with AraBERT parameters 	<ul style="list-style-type: none"> • Transformer • BERT to BERT 	<ul style="list-style-type: none"> • Arabic Empathetic Dialogues 	<ul style="list-style-type: none"> • PPL = 17.0 • BLEU score = 5.58
[95]	Build an Arabic chatbot for autistic children "LANA" <ul style="list-style-type: none"> • The need for high-quality data • Extract the answer without diving into language structure and morphology 	<ul style="list-style-type: none"> • Built and preprocessed manually • PM is effective regardless of language and is straightforward to comprehend • Use cosine similarity algorithm to get the most matching answer to user query 	<ul style="list-style-type: none"> • Pattern matching • Cos similarity 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • N/A
[96]	Build an Arabic Chatbot in Algerian dialect that helps patients get answers about their medical issues" DZchatbot" <ul style="list-style-type: none"> • The need for high-quality data • Build a Generative model 	<ul style="list-style-type: none"> • Collect and preprocess data manually • Use Seq2Seq encoder-decoder 	<ul style="list-style-type: none"> • Seq2Seq • LSTM • BI-LSTM • GRU 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • f1 score • Precision • accuracy
[97]	"IbnSina" an interactive humanoid bot that mimics the personality of ibn Sina <ul style="list-style-type: none"> • The need for high-quality data 	<ul style="list-style-type: none"> • Collected manually via - Book - Quran - Chatterbot corpus 	<ul style="list-style-type: none"> • Pattern matching 	<ul style="list-style-type: none"> • Manually - Via online content 	<ul style="list-style-type: none"> • N/A
[98]	develop hybrid semantic-based and keyword-based to retrieve Medical and Health related topics from the Quran verses <ul style="list-style-type: none"> • The need for high-quality Ontology data • Extract efficient answers 	<ul style="list-style-type: none"> • Built manually • Extract keywords and use SPARQL to query users' utterances 	<ul style="list-style-type: none"> • Semantic keyword matching 	<ul style="list-style-type: none"> • Quraan: - Quran ontology 	<ul style="list-style-type: none"> • Accuracy = 96%
[99]	A chatbot for hotels that assists guests by handling room bookings and inquiries about various services. <ul style="list-style-type: none"> • the need for high-quality data • Handling Arabic language morphology • Control the conversation flow • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Dataset was Built manually • Using Government and Binding theory to restrict the phrase and use it to get the result • Using Dialogue Manager • Conducted through a survey 	<ul style="list-style-type: none"> • Govern me-nt and Binding (GB) 	<ul style="list-style-type: none"> • Manually - Hotel informatio n 	<ul style="list-style-type: none"> • Manually - users acceptance : 92%
[100]	this is a continuation of the progress made in the development of ArabChat by Hijjawi al. <ul style="list-style-type: none"> • the main objective is to Classify user utterances into either questions or non-questions. 	<ul style="list-style-type: none"> • Used Arabic function words such as "هل" "do/does", "كيف" "How" to classify questions and nonquestions utterances with Decision Tree Classifier 	<ul style="list-style-type: none"> • Decision tree 	<ul style="list-style-type: none"> • N/M 	<ul style="list-style-type: none"> • N/A
[101]	build a healthcare chatbot using dialectic corpus <ul style="list-style-type: none"> • the need for high-quality data • extract correct answer while respecting the context 	<ul style="list-style-type: none"> • Data set collected manually • Using Bert to get the most matching response of the user utterance 	<ul style="list-style-type: none"> • Transfor m-ers • Bert 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Accuracy = 95%

[102]	Build a multilingual chatbot that answers student questions called "Jooka" <ul style="list-style-type: none"> • the need of high-quality data • using an existing platform. • Extract correct answers • Handle Arabic complexity • Evaluation the chatbot performance 	<ul style="list-style-type: none"> • Data set collected manually • Built using dialog-flow platform • Intent matching technique • Using a translation model to translate Arabic to English. • Conducted through a survey with 27 participants 	<ul style="list-style-type: none"> • Dialog flow • Entity extraction. • Intent matching 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - user satisfaction
[103]	Utilizing deep learning techniques for text classification and Named Entity Recognition to construct a dialogue system in the field of Arabic home automation. <ul style="list-style-type: none"> • the need for high-quality data • Understanding the user's utterance • Present text data in the models (text to vector) • Tunning the models 	<ul style="list-style-type: none"> • Collected and preprocessed manually • Extracting intent and entity using LSTM and CNN model. • Using AraVec: a pre-trained word embedding model for Arabic • Using Dropout and Early Stopping techniques 	<ul style="list-style-type: none"> • Intent classification: <ul style="list-style-type: none"> • LSTMs • CNNs • Entity extractor: <ul style="list-style-type: none"> • LSTM-bidirectional 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Intent classification: <ul style="list-style-type: none"> - F1 score = 93% • Entity extractor: <ul style="list-style-type: none"> - F1 Score = 94%
[104]	"LABEEB" chatbot that responds to student inquiries in English and in Arabic. <ul style="list-style-type: none"> • The need for training data • Handle the Arabic morphology challenges • Extract an accurate answer • Ability to answer and cover different topics • Ability to interact in vocal mode 	<ul style="list-style-type: none"> • Collected and preprocessed manually • Using NLP techniques to normalize words • Wikipedia API to retrieve the first paragraph and send it as an answer • Microsoft voice recognition to build a chatbot that responds to student utterances. 	<ul style="list-style-type: none"> • NLP • Wikipedia API • Microsoft text-voice recognition. 	<ul style="list-style-type: none"> • Manually 	N/A
[105]	"IbnSina" in mobile platform <ul style="list-style-type: none"> • The need for high-quality data • Covering multiple topics 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Access to online content. 	N/M	N/M	N/A
[106]	Building an Arabic chatbot <ul style="list-style-type: none"> • Automating AIML building • Extract the correct answer 	<ul style="list-style-type: none"> • Using machine learning to build different multilingual AIML files and use them as datasets so the chatbot will learn from them • using ALICE chatbot a third-party platform • Using the first word and the most significant word approaches 	<ul style="list-style-type: none"> • Pandora-bots • Pattern matching • Machine learning 	<ul style="list-style-type: none"> • BNC corpus • Quran 	N/A
[107]	this is a continuation of the progress made in the development of ArabChat by Hijjawi al. <ul style="list-style-type: none"> • Aims to Transform the ArabChat into a mobile app 	<ul style="list-style-type: none"> • ArabChat core implemented in a mobile platform 	<ul style="list-style-type: none"> • Pattern matching 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • RMUT = 72%
[108]	Develop a chatbot that is capable of responding to COVID-19 questions. <ul style="list-style-type: none"> • The need for high-quality data • Extracting the upmost matching answer. • Evaluating the performance of the chatbot. 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Using word similarity and term rare fraction/TF-IDF to look into the cluster about the most matching answer. • Asking medical doctors and a public trial of 308 real users. 	<ul style="list-style-type: none"> • Word similarity 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - user satisfaction : 79%
[109]	"Nabiha" a Chatbot that can engage in conversations with	<ul style="list-style-type: none"> • Using a Java program. • The data was introduced into a 	<ul style="list-style-type: none"> • Pandora-bots 	<ul style="list-style-type: none"> • Manually • Web 	<ul style="list-style-type: none"> • Manually - user

	<p>Information Technology (IT) students at King Saud University, employing the Saudi Arabic dialect.</p> <ul style="list-style-type: none"> • Convert all the data into AIML files • Using existing Platforms to build chatbot • Evaluating the performance of the chatbot. 	<p>Pandorabots platform to build the Arabic dialect chatbot.</p> <ul style="list-style-type: none"> • tested by the students of the IT department 	<ul style="list-style-type: none"> • Pattern Matching • Java • AIML 	<p>scraping</p>	<p>satisfaction</p>
[110]	<p>Arabic question-answering system</p> <ul style="list-style-type: none"> • The need of high-quality data • Determine the context of the question. • Extract the exact answer 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Classify questions that start with "Where" as pertaining to place or location, and questions that start with "When" as relating to time or a date • Using part of speech tagging to identify the most matching answer. 	<ul style="list-style-type: none"> • Information Retrieval • part of speech tagging 	<ul style="list-style-type: none"> • Manually - newspaper 	<p>N/A</p>
[111]	<p>"Ollobot" medical assistant for Arabic users. offers healthcare services that help patients with their physical activities, diet, nutrition, and mental wellness</p> <ul style="list-style-type: none"> • The need of high-quality data • Using an existing platform that provides natural language understanding. • Control the conversation flow 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Using IBM Watson Conversation to build the Arabic dialogue. • click on follow-up choices buttons. 	<p>IBM Watson</p>	<ul style="list-style-type: none"> • Manually 	<p>N/A</p>
[112]	<p>"Rahhal" Arabic chatbot designed to assist travelers exploring Saudi Arabia.</p> <ul style="list-style-type: none"> • The need for high-quality data • Provides more control over the conversation and Reduced Ambiguity • Evaluating the performance of the chatbot. 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Using a button-based interaction approach along with the IBM Watson platform to build the chatbot • Conducted through a survey with 100 participants 	<ul style="list-style-type: none"> • IBM Watson 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - user satisfaction
[113]	<p>Building a new similarity method to calculate and find similarities between questions in Arabic</p> <ul style="list-style-type: none"> • The need of high-quality data • Retrieve the most relevant question matching the user's utterance. 	<ul style="list-style-type: none"> • Collected and preprocessed manually. • Using lexical (string distance) and semantic (original form of the wordnet) similarity • Using cos, Euclid and Jaccard distance to measure the distance between two questions 	<ul style="list-style-type: none"> • lexical and semantic logic • Wordnet 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Precision = 85%
[114]	<p>"SeerahBot" chatbot that offers information about prophet Muhammad</p> <ul style="list-style-type: none"> • The need for high-quality data • Retrieve the most relevant answers to the user's utterance. • Evaluating the performance of the chatbot. 	<ul style="list-style-type: none"> • Manually collected and preprocessed. • Utilized NLTK for preprocessing user utterances. • Extracted valuable keywords using TF-IDF. • Retrieved answers from a predefined corpus. • Conducted through a survey with 14 participants 	<ul style="list-style-type: none"> • NLP • TF-IDF • Keyword matching 	<ul style="list-style-type: none"> • Seerah books in the form of QA pairs 	<ul style="list-style-type: none"> • Manually -user satisfaction

[115]	Robot that helps reduce bullying in school	<ul style="list-style-type: none"> • Real robot that interact with students and teach them about the effects of bullying others and how to avoid it. 	<ul style="list-style-type: none"> • Java • Android studio • Sentiment expression • Voice recognition • image recognition 	N/M	N/A
[116]	<p>dialectal Tunisian chatbot</p> <ul style="list-style-type: none"> • The scarcity of freely available corpora and Their size is relatively limited. • Enable users to interact in vocal mode. • Utilize an existing platform to build the chatbot with more control over the conversation flow. 	<ul style="list-style-type: none"> • Manually preprocessed an existing dataset. • SPEECH RECOGNITION is activated to receive and transcribe users' requests. And, the Speech Synthesis model to generate the corresponding voice • Use RASA as a third-party platform to build the chatbot. 	<ul style="list-style-type: none"> • RASA • CNN • RNN 	<ul style="list-style-type: none"> • the Spoken Tunisian Arabic Corpus (STAC) 	<ul style="list-style-type: none"> • Accuracy = 97% • Manually.
[117]	<p>New Arabic method to calculate similarity between two questions</p> <ul style="list-style-type: none"> • The scarcity of available corpora • Calculate the similarity to extract the most relevant answer 	<ul style="list-style-type: none"> • Manually collected and preprocessed Topical: <ul style="list-style-type: none"> • Textual similarity, lexical similarity, and semantic similarity non-Topical: <ul style="list-style-type: none"> • Rule-based 	<ul style="list-style-type: none"> • Named Entities • bag of words • Word similarity • Cosine • Jaccard • Euclidian distance 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Precision = 85%
[118]	<p>Arabic QA system over linked data</p> <ul style="list-style-type: none"> • The scarcity of available Arabic corpora • Construct an efficient Arabic question-answering system without delving into the complexities of Arabic grammar and morphology. 	<ul style="list-style-type: none"> • Manually collected and preprocessed • Using NLP techniques to preprocess the user question. <ul style="list-style-type: none"> - Tokenization, Normalization, Resource Extraction and Keywords Extraction to build a Keywords List. • Use extracted keywords to formulate the SPARQL query. • SPARQL query is used to extract the answer. 	<ul style="list-style-type: none"> • NLP • SPARQL 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Precision = 71% • Recall = 69% • F-measure = 70%
[119]	<p>"SIAAA-C" Student Interactive Assistant Android Application with Chatbot</p> <ul style="list-style-type: none"> • The scarcity of available datasets and handling the Arabic complexity. • Extract the correct answers • Evaluating performance of the chatbot. 	<ul style="list-style-type: none"> • Created a corpus containing every possible question along with its corresponding answer. • Using Keyword matching technique • Conducted through a survey with 102 participants 	<ul style="list-style-type: none"> • Java • Key Word matching 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - user satisfaction : 80%

[120]	"SEG-COVID" Chatbot that helps students get information about Covid-19 <ul style="list-style-type: none"> • The scarcity of available dataset and handle the arabic complexity. • Extract correct answer 	<ul style="list-style-type: none"> • Created manually • Using Word matching technique 	<ul style="list-style-type: none"> • Java • Word matching 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually
[121]	Build a QA system based on semantic matching using deep-learning model	<ul style="list-style-type: none"> • Arabic bidirectional encoder representations from transformers (AraBERT) • BERT contextual representation with BiLSTM (BERT-BiLSTM) • hybrid transfer BERT contextual representation with BiLSTM 	<ul style="list-style-type: none"> • BERT • LSTM • BiLSTM 	dataset called: "Tawasul"	<ul style="list-style-type: none"> • Accuracy = 94%
[122]	Build a QA health dataset in the medical domain and build a generative model to answer user's questions. <ul style="list-style-type: none"> • The need for high-quality data • Build a generative model to answer the user's question. 	<ul style="list-style-type: none"> • Manually collected and preprocessed a dataset called "MAQA" consisting of over 430,000 QA pairs. • Use LSTM and Bi-LSTM 	<ul style="list-style-type: none"> • LSTM • Bi-LSTM 	<ul style="list-style-type: none"> • Built Manually "MAQA" 	<ul style="list-style-type: none"> • Blue score = 58%
[123]	Build a system that is capable of categorizing the user act in order to build an efficient Chatbot <ul style="list-style-type: none"> • The need for high-quality data • Building an efficient model 	<ul style="list-style-type: none"> • The corpus was manually built; it comprises 873 sentences. • Using TF-IDF to vectorize text and SVM with n-gram to detect the act of user's query 	<ul style="list-style-type: none"> • TF-IDF • SVM 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Accuracy = 86%
[124]	Develop a chatbot designed to offer tourist guidance. <ul style="list-style-type: none"> • The need for high-quality data • Control over the conversation flow 	<ul style="list-style-type: none"> • The corpus was manually built. • Using RASA to build the chatbot 	<ul style="list-style-type: none"> • RASA 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Manually - user satisfaction
[125]	Dialogue system for automated homes utilizing NLU, Named Entity Recognition, and text classification. <ul style="list-style-type: none"> • The need for high-quality data 	<ul style="list-style-type: none"> • Using AQMAR dataset and collecting other data manually 	<ul style="list-style-type: none"> • LSTM • Bi-LSTM • CNN 	<ul style="list-style-type: none"> • Manually 	<ul style="list-style-type: none"> • Accuracy = 96%
[126]	Build a dialogue system to facilitate the search for Islamic related information in an interactive way. <ul style="list-style-type: none"> • Extracting accurate answers. • Handling the Arabic morphology richness 	<ul style="list-style-type: none"> • NLP to understand and extract KEY WORDS • Using a semantic approach to extract the literary meaning of each word. • Formulate and execute a SPARQL query. 	<ul style="list-style-type: none"> • NLP research • Semantic "Wordnet" • SPARQL 	<ul style="list-style-type: none"> • Quran ontology 	N/A
[127]	Build a system to answer question related to Islam from the Quran and Tafseer books. <ul style="list-style-type: none"> • Evaluating the performance of the chatbot. 	<ul style="list-style-type: none"> • using NLP techniques to extract keywords • Using SVM to classify utterances and extract the most relevant answer. • Conducted through a survey with 5 participants 	<ul style="list-style-type: none"> • NLP • SVM 	<ul style="list-style-type: none"> • QURAN • TAFSEER 	<ul style="list-style-type: none"> • Manually - correct answer = 76%
[128]	Arabic Question Answering system called "AQuASys" <ul style="list-style-type: none"> • Problem of large morphological and derivative diversity for an Arabic word • Understand the syntactic of user question 	<ul style="list-style-type: none"> • Using khoja stemmer • Using NLP techniques and Part of Speech tagging (POS) • Using the keyword search approach to extract correct answers 	<ul style="list-style-type: none"> • NLP 	N/M	<ul style="list-style-type: none"> • Precision = 66.25% • Recall: = 97.5%

- Extract correct answers

[129]	Build a QA system in Arabic	<ul style="list-style-type: none"> • Using NLP techniques and Named Entity Recognition to extract keywords and execute a search to extract the most relevant answer. 	<ul style="list-style-type: none"> • NLP - NER 	N/M	<ul style="list-style-type: none"> • Precision = 83%
[130]	Build a QA system in Arabic that can handle multiple domains.	<ul style="list-style-type: none"> • Using Wikipedia as a knowledge base. • Querying the user's question using lexical patterns to extract the answer. 	<ul style="list-style-type: none"> • Lexical pattern • WIKIPE D-IA API 	<ul style="list-style-type: none"> • Wikipedia 	<ul style="list-style-type: none"> • Accuracy = 63%

4.3 RESULTS OF SYNTHESSES

Following a thorough analysis of the research, we have identified six primary challenges that were covered in the selected studies. A summary of these challenges, along with their respective solutions can be found in Table 4. In this section, we will discuss each problem and explore the various approaches utilized to solve them.

4.3.1 CHALLENGES ENCOUNTERED WITH DATASETS.

4.3.1.1 LIMITED AVAILABILITY OF ARABIC DATASETS.

4.3.1.1.1 COLLECT AND PREPROCESS DATA MANUALLY.

The obstacle of having a limited number of high-quality datasets is widely acknowledged as a major challenge in the progress of Arabic chatbot innovation. Therefore, several studies (n=34) have chosen to overcome this challenge by collecting and preprocessing data manually, via conducting various resources such as books, movie scripts and online sources. In references [67, 70-72, 74, 76, 77, 80-81, 85, 87, 88, 95-99, 101-105, 108, 110, 111, 114, 116-120, 122-125] researchers collected and built their own datasets manually. For instance, in [110] authors have built an Arabic corpus based on newspaper content. In [72, 109] researchers used web scraping techniques to collect data automatically from web pages, parsing their content, and storing the extracted data in a corpus. In [77] authors collected data via online surveys, they created a survey template and asked a diverse group of online users to provide their responses, the participants' input was then preprocessed and organized in the form of a Question-answer corpus. Another approach was used in reference [85] called 'Wizard of Oz' technique, the main idea of this technique is to simulate the behavior of an Arabic chatbot through manual responses made by a person behind the scenes. Participants interacting with the system were unaware that the chatbot's replies were actually being created by a human "wizard", thereby guaranteeing genuine and natural interactions. This approach enabled gathering various conversations and oriented exchanges while enhancing the chatbot's possible responses. Although a significant amount of Arabic content is accessible, generating a top-notch Arabic dataset for Natural Language Processing (NLP) or other data-driven tasks remains a challenge, they necessitate significant refining and preprocessing. Moreover, some chatbots are designed for specific domains and, as a result, require them to collect their unique datasets; for instance, in reference [98] The researchers used the Holy Quran as a dataset, but they first had to transform it into labeled verses before they could train the model, which was a time-consuming process. Similarly, in reference [114], the Seerah, a book that elucidates the Holy Quran, was converted into a question-answer format to serve as a knowledge base.

4.3.1.1.2 TRANSLATE AVAILABLE DATASETS INTO ARABIC.

One potential solution for addressing the limitation of Arabic datasets is to translate datasets from a language with high resources, such as English, into Arabic. A notable example of this is the work done by A. *Fuad et al.* [83, 89], they developed an Arabic dataset called "Arabic-TOD", which includes 30 thousand utterances, by translating the BiToD [134] dataset from English to Arabic. Similarly, T. *Naous et al.* [93] built an Arabic empathy dataset by translating the EmpatheticDialogues [135] dataset into Arabic.

4.3.1.2 DEVELOPING A CORPUS REQUIRES A SIGNIFICANT AMOUNT OF TIME AND EFFORT

Gathering, organizing and labeling data manually necessitates careful human involvement. This undertaking entails significant dedication of time and effort to guarantee the accuracy, variety and significance of the data. In order to overcome these restrictions, some studies have adopted automated techniques to accelerate the process of creating datasets. Researchers In [75, 79, 106, 109] developed a Java program along with machine learning approaches to automate and convert a corpus into AIML (Artificial Intelligence Markup Language) files and use it as a knowledge base to train the chatbot. In [75, 79] researchers were able to transcript the Quraan into AIML file automatically. In

[106] *B. Abu Shawar et al*, successfully developed a Java program that generates various AIML files by utilizing diverse training data sets in multiple languages, including Arabic. In [109] *Al-Ghadhban et al*, constructed an IT dataset by automatically converting textual content into AIML files.

4.3.1.3 BUILD AN EFFICIENT MODEL WITH A SMALL DATASET.

Building a successful model that can mimic human-like conversation is highly reliant on the quality, diversity and volume of the training dataset. However, due to the scarcity of available high-quality datasets, completing this task is a challenging endeavor. The solutions that are proposed in the selected articles in order to meet this challenge are as follows:

4.3.1.3.1 USING RULE-BASED AND RETRIEVAL-BASED APPROACHES.

Researchers in references [77, 85, 87] have proposed addressing the challenge posed by limited datasets through the adoption of Rule-based and Retrieval-based approaches. These approaches are particularly advantageous as they do not necessitate extensive training data, given that they rely on predefined rules and predefined answers rather than trainable models. Techniques such as regular expressions, pattern matching, keyword matching and word distance measurement constitute the foundation of these methods. Furthermore, chatbots based on these approaches are typically task-oriented, created to address specific inquiries and provide accurate information within a particular topic or industry. The ability of these approaches to customize the chatbot's responses to a specific domain makes them highly effective in handling user queries and offering relevant information within that specific domain. Therefore, the majority of the studies adopted these approaches, as shown in Table 4.

4.3.1.3.2 EMPLOYING JOINT TRAINING.

The primary objective of Joint Training is to improve the model's performance by training it on data sets not only from Arabic dialogue but also from high-resource languages such as English and Chinese. This approach leads to better results as it enables the model to learn from a diverse set of language structures and more types of patterns, which will help it better understand and generate new responses [136]. For instance, in [83] *A. Fuad et al* employed the joint training approach to handle the problem of Arabic small datasets, to verify the hypothesis and compare the results of joint training, they built three distinct models called "AraConv models". The first model was trained only on the mono-lingual corpus, specifically in Arabic. The second model was trained on bi-lingual language, incorporating both Arabic and either English or Chinese. The third model was trained on the multi-lingual corpus, encompassing Arabic, English, and Chinese. Various metrics were utilized to compare the results; however, for the sake of conciseness, we will only include the BLEU score metric, which is According to *Sourav Dutta (2019)* [137] "BLEU Score is an algorithm that was primarily developed to evaluate how accurate machine-translated text was" (p. 7). The findings reveal that the AraConv model in the multi-lingual setting (BLEU Score=32.58) outperformed the AraConv model in the mono-lingual setting (BLEU Score=31.05).

4.3.1.3.3 LEVERAGING CROSS-LINGUAL LEARNING.

According to *Pikuliak et al.* [138] "Cross-lingual learning is a paradigm for transferring knowledge from one natural language to another" (p. 1). The main idea of cross-lingual learning is in order to build an efficient chatbot model with a low-resource language such as Arabic, the model should be trained first on a high-resource language such as English [139]. Subsequently, by transferring the acquired knowledge from a language with high-resources, the low-resource language model can benefit from the shared linguistic properties, syntactic structures and semantic information that are common across languages. This approach can help address the issue of insufficient data and enhance the proficiency of low-resource models without necessitating abundant labeled data for training [138]. In [89], *A. Fuad et al* aim to investigate the effectiveness of utilizing cross-lingual transfer learning in constructing an end-to-end Arabic task-oriented dialogue system, by using the multilingual variant of the Text-To-Text Transfer Transformer model (mT5). The researchers employed the Arabic-TOD dataset for both training and testing the model performance. To tackle the challenge posed by the limited Arabic dialogue dataset, they introduced three distinct cross-lingual transfer learning methods: mSeq2Seq, Cross-lingual Pre-training (CPT) and Mixed-Language Pre-training (MLT). The research's key finding is that cross-lingual transfer learning has the potential to enhance the performance of the Arabic system when confronted with small datasets. Furthermore, the study demonstrates that increasing the size of the training dataset leads to improved results.

4.3.1.3.4 APPLYING N-GRAM AUGMENTATION TO THE DATASET.

Another approach addressing the challenge of small datasets is training the model using multiple N-gram representations of the dataset. An N-gram refers to a sequence of n consecutive words or characters in a text [140]. For example, in the expression "Arabic Conversational agent" the 2-gram representation (or bigrams) would consist of "Arabic Conversational " and " Conversational agent" while the 3-gram would be " Arabic Conversational agent" The variable "n" represents the number of words or characters within each sequence. In [70], *M Habib et al.* Aim to build a model that can write medical recommendations in Arabic using Natural Language Generation (NLG) approaches such

as Unidirectional and bidirectional LSTM, one-dimensional convolutional neural network (CONV1D), and a pairing of both these models (LSTM-CONV1D). However, due to the limited availability of Arabic datasets, the researchers employed N-grams as a strategy to enhance the dataset's comprehensiveness. This involved the creation of two distinct versions of training datasets, specifically incorporating 3-gram and 4-gram representations. For instance, applying 2-gram and 3-gram representations to a dataset that has only one sentence with 5 words such as "*Deep learning revolutionized artificial intelligence*", will create a dataset containing 8 sentences and 18 words. In addition to augmenting the dataset, N-gram representation will also enhance the model's contextual understanding [141], by capturing sequential patterns of words within the text and gaining insight into how words interact within phrases and sentences.

4.3.2 HANDEL MISSPELLED WORDS AND OUT-OF-CORPUS QUESTIONS

Natural language used by humans possesses a vast array of forms and structures, making it inherently complex and open to interpretation. Words can often have multiple meanings (lexical ambiguity) and a sentence can be constructed in various structures (syntactic ambiguity) [68]. Moreover, real-life conversations are often grammatically incorrect and confronted with misspellings. Consequently, comprehending users' utterances, particularly in languages like Arabic poses a challenging task. The solutions that are proposed in the selected articles in order to meet this challenge are as follows:

4.3.2.1 PROXIMITY PROCESSING WITH POOL OF SUGGESTIONS

In [67], YM Mohialden et al used a technique called Proximity Processing with a pool of suggestions is a technique utilized for correcting misspellings by presenting alternative terms when the user inputs a term that is misspelled or contains errors. This method involves generating a selection of possible suggestions based on the initially misspelled or erroneous term and subsequently presents the most suitable suggestion to the user.

4.3.2.2 UTTERANCE VALIDATION

M. Hijjawi et al [68], OG. Alobaidi et al [73] and Z. Sweidan et al [119] Addressed the issue of erroneous words by employing the utterance validation process. The main objective of this approach is to verify and validate the correctness, relevance, and suitability of the user's utterances for matching against scripted patterns before proceeding with any kind of preprocessing or matching; if the utterance is recognized as valid, the chatbot engine proceeds to process it; otherwise, the model responds to the user by indicating that his or her utterance is not valid.

4.3.2.3 ORTHOGRAPHIC NORMALIZATION

Ali and Habash [86], used orthographic normalization to standardize the representation of words, thereby overcoming the inconsistent spelling variations of certain characters and also increasing the likelihood of matches between user utterances and prescribed patterns. The authors addressed prevalent errors commonly found in Arabic texts, such as the misspelling of the letter Alif-Hamza (i.e., Alif with Hamza Above (أ), Alif with Hamza Below (إ), Alif with Madda (آ), Alif without Hamza (ا)), which is considered as the most common spelling mistake in Arabic texts. They normalized these different forms into a standardized representation which is Alif without Hamza (ا). Moreover, an additional frequent error involves the confusion between two letters ("Ta-Marbuta" (ة) and "Ha" (ه)) and ("Alif-Maqsurah" (ة) and "Yaa" (ي)) when they appear in the final position of words, this confusion is due to both their visual similarity and the fact that they are pronounced the same in certain cases. To avoid this confusion, they changed every ("Ta-Marbuta" to "Ha") and ("Alif-Maqsurah" to "Yaa") when they appeared in word-final positions. Implementing these transformations enhanced the authors' chatbot's pattern-matching capability and it was, as stated in the article in reference [86] (p .4), "able to overcome 85.1% of the spelling mistakes found in spontaneous Arabic typing".

4.3.2.4 CREATE A DEFAULT CORPUS

Since there is no way to guarantee that a user's utterances entered will be the same as those stored in the chatbot knowledge base, Shawar et al. [71] suggested addressing the out-of-corpus inputs issue (utterances not in the pre-scripted patterns) and inputs with errors by deploying a default corpus, which operates based on the concepts of First Word and Most Significant Word. According to Shawar et al., the main goal of the first word is to classify the question. For example, when a question starts with "who," it often indicates that the question is seeking information about a person or individuals, while when it starts with "where," it typically suggests that the question is seeking information about a specific location or place. The authors demonstrate that the least frequent word in a question holds the most significant information content. For instance, in the question "What is your name?", the least common word is "name", thus the answer should be generated based on it [71]. Another suggestion was proposed by Al-Madi et al. [81], in order to handle the out-of-corpus utterances problem, they recommended storing the user's utterances in a log file. Afterward, an administrator can add an appropriate response to the recorded utterances. The purpose of this process is to gradually enhance the system's knowledge base and improve its capability to handle a wider variety of inquiries as time goes on.

4.3.3 HANDLE THE COMPLEXITIES OF THE ARABIC LANGUAGE

The core engine of a chatbot is its ability to understand and preprocess language in a human-like manner, but natural language processing in Arabic is confronted with many challenges, as illustrated in Section 2. The solutions that are proposed in the selected articles in order to meet this challenge are as follows:

4.3.3.1 EXTRACT ANSWERS WITHOUT DELVE INTO THE LINGUISTIC STRUCTURE.

The Arabic language is known for its rich morphology, high level of ambiguity, and lack of appropriate resources. As a result, building an effective language processing and understanding model is a challenging endeavor. Therefore, many studies suggest using approaches that are scripted manually. Due to the fact that these approaches aim only to detect particular patterns, key terms, or other indicators that suggest the presence or absence of relevant information, they subsequently allow for the retrieval of answers and information from text without the need for intricate linguistic analysis or consideration of the language's grammar, syntax, or structure. As a result, this enables a more straightforward and effective retrieval of answers. Furthermore, they provide more control over language ambiguity. The suggested approaches are Pattern Matching (PM), Key Word Matching (KWM), Text Similarity (TS), and Semantic Word Matching (SWM).

4.3.3.1.1 PATTERN MATCHING

PM is a process that aims to match a user's utterance with pre-scripted patterns. It involves searching for specific strings within a piece of text to identify all occurrences of those strings within the text [68]. Generally, PM architecture is built in a hierarchical manner, or as a knowledge tree [82]. It begins with defining domains; each domain has different contexts, and each context has its own patterns. Additionally, it employs a technique called wildcards, which is a placeholder symbol used to match a portion of the user's utterance. For instance, in the case of a pattern like "I want to travel to *", the "*" symbol functions as a wildcard, it has the ability to correspond with different words that the user may include after "to" such as "Paris", "London" or "New York". As a result, this wildcard enables the system to identify and react appropriately to various inputs from users that follow a certain pattern. Several studies have employed the PM approach to build their chatbots [68, 73, 82, 86, 90, 91, 95]. The common reason for choosing this approach is that PM relies only on the matching process; it doesn't depend on any grammatical or linguistic details, and can also maintain conversation flow. Thus, it helps in handling the complexities of the Arabic language successfully. By adopting this technique, the authors in [68] were able to build a chatbot that handled 73% of the users' utterances. Similarly, the authors in [90] used AIML to script the patterns, and due to its capability to extract pertinent patterns with a low error rate, they successfully developed a PM chatbot that obtained a user approval rate of 61%. In [95]. Aljameel et al. adopted the PM approach to develop an Arabic Conversational Intelligent Tutoring System, called LANA-I, for children with ASD to enhance their learning experience.

4.3.3.1.2 KEYWORD MATCHING

KWM is a technique that involves extracting a set of keywords from the user's utterances and then comparing them to the predefined questions or answers in the corpus. The ratio of matching keywords indicates the relevance of the response. Many studies [71, 72, 114, 119, 120, 127] have employed this technique as it requires no processing and is solely based on matching techniques. Abu Shawar [71] used this technique to build a Chatbot that answers users' Islamic questions with verses from the holy Qur'an (similarly as [127]) via matching specific keywords. The developed chatbot has a commendable accuracy rate of 93%. Yassin et al. [114] also used KWM to build an information retrieval-based chatbot that offers information about the Prophet's Muhammad (PBUH) biography. Z. Sweidan et al. [119] suggested using, in addition to KWM, a mathematical comparison to retrieve the most relevant answers by comparing the user's question with all questions stored in the corpus and calculating matching ratios at each step of the process, the question with the highest matching ratio is considered to be associated with the most relevant response. Alhassan et al. [72] used the KWM technique along with AI for the purpose of analyzing the user's input and extracting only important keywords, subsequently matching them with the predefined corpus to retrieve the most relevant answer. As a result, they built an IT-related assistance chatbot capable of providing solutions to users' problems with high accuracy.

4.3.3.1.3 TEXT SIMILARITY

TS is an algorithm that aims to find how two sentences or documents are similar based on some mathematical concepts and equations; the greater the commonness, the higher the similarity [142]. There are three main ways to measure the similarity between sentences: the length distance, the distribution distance, and the semantic distance. Moreover, several algorithms can be used to measure TS, such as cosine similarity, Euclidean distance, Jaccard distance, and Levenshtein distance (more detailed information can be found in [143]). The following studies [76, 81, 95, 108, 113, 117] have applied TS to develop an Arabic chatbot. The study by A. Mundher et al. [76] used both Cosine and Jaccard Similarity to build a chatbot that is oriented toward the educational domain. The primary reason for employing this approach is that these studies are task-oriented and based on information retrieval. Most of their objectives are to extract correct and relevant answers from a predefined corpus; thus, applying TS approaches is the most suitable for executing their tasks because TS does not require in-depth linguistic analysis; however, it is

essentially interested in quantifying the similarity or distance between sentences. In [95], Aljameel et al. Proposed an enhancement for boosting the performance of the pattern-matching- based chatbot through the integration of the cosine similarity algorithm. Similarly, in [108], M. Ghaleb et al. Recommended that in order to enhance the TS algorithm, the dataset should be stored in clusters, and normalization procedures (e.g., removing stop words, tokenization, and stemming) should be applied to both the user's question and the questions stored in the corpus before comparing them. Consequently, reducing the calculation time and also the ratio of error. M. Daoud. [113, 117] used hybrid approach that utilizes string similarity and semantic similarity along with machine learning model (SVM [144]). They successfully developed a Question Answer system that produced a good result, achieving a precision rate of 85%.

4.3.3.1.4 SEMANTIC KEY WORD MATCHING USING SPQRQL.

Although text similarity provides favorable results, certain questions require a deeper understanding. This is because a single word can have multiple synonyms, and text similarity relies on a one-to-one word-matching approach that may not capture all aspects of similarity. Therefore, semantic-based techniques can fill the gap. Semantic-based techniques are able to find similarity based on matching the contextual meaning of a keyword, thus providing more relevant and accurate search results [145]. To deploy semantic search, the knowledge base should be stored in the form of ONTOLOGY, which is a formal and structured representation of knowledge that defines the concepts, entities, relationships, and properties within a particular domain and provides a shared understanding of the domain's semantics. To extract information from ontologies and linked data sources, most studies used SPARQL (SPARQL Protocol and RDF Query Language), which is a query language and protocol used to query and manipulate data stored in RDF (Resource Description Framework) format [118]. The following studies [98, 118, 126] have employed semantic search along with SPARQL to extract information from interlinked data. A. Bouziane et al. [118] explained the different steps of extracting an answer, which start with preprocessing the user's question using normalization techniques, after which keywords are extracted, and then a SPARQL query is formulated and executed. The same process was adopted by [118, 126] to answer user's question from the Quraan ontology.

4.3.3.2 USING STAT-OF-ART PLATFORMS

Given the complexity and high ambiguity of the Arabic language, building a high-quality Arabic chatbot that integrates all the NLP, NLU, and NLG approaches is a challenging task, as it requires a deep understanding of the language structure and compatible tools and technologies (e.g., algorithms and models that support Arabic). Moreover, developing an Arabic chatbot from scratch does not guarantee success due to the complex nature of Arabic and the multiple challenges associated with Arabic chatbots (e.g., data sets, algorithms, and evaluation metrics). In addition, the majority of Arabic chatbots are task-oriented. They are designed to answer specific questions or perform specific tasks. Hence, they don't require a high level of language processing. Therefore, many studies opt for existing platforms that provide Arabic language support and offer some level of intelligence to facilitate the development of chatbots. For instance, Alhassan et al. [111] used IBM Watson as a third-party platform to build a chatbot that offers medical assistance to Arabic users. Likewise, Rocha et al. [102] used the DialogFlow platform to develop a university admissions chatbot at the German University in Cairo (GUC) supporting English and Arabic. Al-Ajmi et al. [85] used the Wit.ai platform to develop an Arabic flight booking dialogue system. Researchers in [75, 79, 86, 106, 109] built Arabic chatbots using the Pandorobot platform, and researchers in [77, 84, 116, 124] built Arabic chatbots using the Rasa platform. The utilization of this platform yielded positive outcomes in several studies. For example, in [84], an accuracy of 95% and an F1-score of 94% were achieved. Likewise, [116] reported an accuracy of 97%, while [125] attained an accuracy of 95%.

4.3.3.3 LACK OF APPROPRIATE ARABIC RESOURCES.

Due to the limited availability of appropriate resources in Arabic, including knowledge bases, models, Arabic WordNet, libraries for orthographies normalization, and algorithms supporting Arabic, Rocha et al. [102] proposed building a chatbot that is trained on English datasets, since English is a high-resource language and more advanced than Arabic, and adopting a translation model to handle the users' Arabic utterances. When a user provides an utterance in Arabic, it is first translated into English using the Google Cloud Translation API and then sent to the chatbot engine to be processed and matched to the most suitable answer in the training corpus. After extracting the relevant response, it is translated back into Arabic before being delivered to the user. Adopting this approach enabled researchers to build not only an Arabic chatbot but also a bi-lingual chatbot that understands and interacts in English. Moreover, the developed chatbot showed good performance results. The evaluation metric employed was the System Usability Scale (SUS) [150] and the Chatbot Usability Questionnaire (CUQ) [151], which yielded a mean SUS score of 88.5% and a CUQ score of 87.3%. Another challenge faced in the domain of Arabic chatbots is the diversity of the Arabic language representation (i.e., MSA, CA, and dialectal variations). To overcome this challenge, A. Shawar et al. [79] suggested supporting only the MSA version of the Arabic language. This is because MSA is formal, structured, and offers greater consistency compared to other versions.

4.3.3.4 THE PROBLEM OF LARGE MORPHOLOGICAL AND DERIVATIVE DIVERSITY FOR THE ARABIC WORDS

One of the reasons Arabic is highly ambiguous is due to the derivational nature of its words [90]. Through the addition of affixes, inflections, diacritical changes, and modifications to gender forms, the root of a word in Arabic can result in the creation of multiple words (Section 2). As a result, creating a dataset that includes all word variations or manually scripting all potential Arabic word patterns will be expensive in terms of effort, time, and efficiency. Several authors have recommended the use of word normalization techniques as an alternative to overcome this challenge. For instance, using Lemmatization and Stemming to return words to their original form. The main difference between lemmatization and stemming is that lemmatization transforms words into their original form based on context, whereas stemming reduces words to their base by removing prefixes and suffixes without maintaining the word's meaning [87]. The word normalization method effectively reduces the volume of training data while also allowing for the understanding of words that have not been explicitly trained on. For instance, in [86, 104], researchers used the orthographic normalization technique offered in the Python library, specifically using the NLTK (Natural Language Toolkit) library, to normalize users' utterances by tokenizing text, deleting stop words, and stemming remaining words. In [118], researchers used the MADAMIRA [146] library for the tokenization and normalization steps. In [87, 91], due to the lack of a sound library in Python, researchers used the ISRI stemmer [147]. Moreover, in [110], researchers used Khoja stemmer [148] to extract Arabic roots from their words; likewise, in [90], researchers enhanced Khoja stemmer functionality to stem the word based on the context.

4.3.3.5 ABILITY TO UNDERSTAND TEXTS CONTAINING FOREIGN WORDS INSIDE ARABIC SENTENCES (CODE-MIXING & CODE-SWITCHING).

Owing to the fact that several Arabic-speaking countries are bilingual, with Arabic as the native language and English, French or Spanish as foreign languages, it is frequent for Arabs to incorporate foreign words into their Arabic expressions in their conversations. This linguistic phenomenon is known as Code-Switching [152]. For instance, in Moroccan colloquial interactions, it is common to say "cava واش ننا" (pronounced as "wash na-ta sa-va"), which mixes two languages (i.e., Arabic and French) to convey the meaning "are you okay". This hybridization of languages is more prominently observed on social media platforms, particularly in text messaging and tweeting. Furthermore, the integrated foreign terms are occasionally adjusted in pronunciation or spelling to suit the regional dialect, making them more coherent with Arabic. This phenomenon is commonly known as Code-Mixing [153]. For instance, in some regions, "bus" is transformed into "طوبيس" (pronounced as "tow-bees") or "الباص" (pronounced as "el-bus"). Nonetheless, incorporating foreign terms into the Arabic language can facilitate communication and enhance mutual understanding among individuals, especially when dealing with technical terms. However, when considering the resource-intensive endeavor of building a chatbot capable of emulating human-like conversations, this practice adds an extra layer of complexity to the task. A. Shawar et al. [79] recommended focusing on the MSA version solely due to its structured nature. However, addressing code-switching and code-mixing is essential for creating more engaging, natural and human-like chatbots. To handle this issue in a more sophisticated manner, G. Bilquise et al. [87] suggested collecting all the popular foreign words that can be integrated into Arabic sentences and scripting all potential patterns manually. As a result, the authors were successfully able to build a dialectal chatbot that understands Arabic questions, including English words, with an accuracy of 75%.

4.3.3.6 MAKE DIFFERENCE BETWEEN QUESTION AND NON-QUESTION UTTERANCE

Real conversations don't always respect the question-and-answer format. Therefore, if a chatbot aims to mimic human-like dialogue, it has to be able to engage in conversation regardless of whether the user's input is in the form of a question or a non-question. To address this challenge, Hijjawi et al. [100] suggested that before any further analysis takes place, the user's input should be categorized as either a question or non-question. To achieve this objective, they proposed a methodology that combines extracting Arabic function words (e.g., 'هل' - do, 'كيف' - how) from the user's input and utilizing a Decision Tree Classifier to classify if the input is a question or no. In addition, Hammo et al. [110] suggested that interrogative pronouns (e.g., what, which, who, whose) or interrogative adverbs (e.g., when, where, why, how), are typically used to start question utterances. Therefore, classifying whether the utterance is a question or not is not a challenging task. Furthermore, they recommended employing interrogative pronouns and adverbs to enhance response quality by extracting fluent and coherent responses. This is because they assist in distinguishing whether the question seeks specific information or inquiries about aspects such as time, place, reason, or manner. For instance, if a question starts with "where" it indicates the user is likely seeking specific information about location. If a question starts with "when," it implies an inquiry about time or a schedule.

4.3.4 CONSERVE THE CONVERSATION FLOW

Engaging in fluid and contextually relevant conversations is a challenging task for chatbots, especially in Arabic, due to the complexity and richness of the Arabic morphology. The subsequent section presents various methods and strategies suggested in the chosen article to tackle the difficulty faced when it comes to the conversation flow.

Z. Noori et al. [82] employed a pre-scripted knowledge tree to guide the conversation toward the system's objectives. According to this article, the employed approach empowered the conversational agent with the capability to maintain control over the dialogue flow, achieving a conversation success rate of 85%. Researchers in [101, 103] suggested to use data-driven approaches to address this challenge. Specifically, in [101] T. Wael et al, used Bidirectional Encoder Representations from Transformers (BERT) to build a healthcare assistance chatbot that supports many Arabic dialects (such as Egyptian dialects, Saudi Arabian dialects, and Iraqi dialects). The model reported an accuracy of 95%. In [103] AM. Bashir et al, Adopted the LSTM architecture, recognized for its effectiveness in NLP and its ability to capture long-term relationships within text segments. Furthermore, intent and entity extraction techniques were employed to improve the chatbot's capacity to comprehend user inputs, recognize the user's objective or intention when sending a message or inquiry and select the proper course of action or response to provide a personalized and contextual experience. By using these methods enabled the authors to build a chatbot that achieved an accuracy rate of 96.33%. Another recommended approach involves utilizing temporary memory to store a captured portion of a user's utterance, such as answers, questions, age, names, and user statements, during the conversation in order to use it later in the conversation to enhance the system's perceived intelligence and fluency in engaging with users. Adopting this approach enabled Alobaidi et al. [73] to build a chatbot that produced a good result, achieving a user satisfaction of 85%. Similarly, in [68], the incorporation of temporary memory led to a significant enhancement in user satisfaction, achieving a rating of 96.5%. In [80] A. Moubaidin et al, suggested that the deploying of a Dialogue Manager in the chatbot engine would be beneficial for maintaining the conversation flow. The main role of the dialogue manager is to maintain the context and flow of the conversation by understanding the user's utterance (e.g., asking further questions to fully understand the intent of the user's utterance), keeping track of what the participants (i.e., user and system) are exchanging, and also taking into account the user's history and preferences to adapt responses and interactions to the respective user. By employing this approach, the authors effectively created a chatbot that achieved positive feedback. Specifically, 65% of users reported that their experience with the chatbot was "very good", with a further 25% rating it as "good". Another proposed approach involves constructing a button-based chatbot, as recommended by S. AlHumoud et al [112], By using this approach they developed a tourist Arabic chatbot. In this approach, instead of typing responses, users interact with the chatbot by selecting predefined options or buttons. This approach offers flexibility and efficiency, reducing the time required for typing, minimizing errors and offering more coherent conversation. The tourist chatbot received favorable feedback, with 87% of users rating its conversational ability as "excellent".

4.3.5 HANDLING CORPUS LIMITATIONS AND COVER WIDE RANGE OF TOPICS

Considering the scarcity of Arabic resources, especially datasets, a significant number of studies collected datasets manually, as shown in Table 4. Although covering all topics within the same domain or covering many domains in the same chatbot is a time-consuming task of data collecting and preprocessing. Therefore, some studies suggest that, in addition to the knowledge base, leveraging external sources of information can enhance the capabilities of the chatbot. In [77] R. Alotaibi et al, developed a tourism-oriented chatbot. In order to cover a wide range of different potential users' questions (e.g., locating nearby hotels) and to provide up-to-date and accurate information (e.g., the current weather or news), they incorporated additionally to the training dataset, external API call functions. These functions extract relevant responses from the web, which are then transmitted to the user. Similarly, in [104] Y. Almuradha, implemented a function that uses the Wikipedia API to retrieve the first paragraph for any needed definition and display it to the user. Likewise, in [105] N. Mavridis et al, enhanced the performance of the IbnSina robot (a robot that mimics the real scientist Ibn Sina), so it could become a more resourceful and exciting educational robot by enabling it to access online content such as Wikipedia.

4.3.6 CHATBOT EVALUATION METRICS

The current state of chatbot evaluation metrics highlights that there is no standard methodology adopted by researchers to assess chatbot performance, owing to the interactive and open-ended nature of conversations, this is particularly notable in the case of Arabic chatbots, which is associated with a lower level of research activity. Therefore, many methods, techniques, platforms and approaches are being suggested and recommended to overcome this challenge [154]. However, there is no consistency across evaluation metrics.

By conducting research on Arabic chatbots, it becomes evident that the evaluation metrics employed are still in their early stages of development. The majority of approaches used are human-based, mainly due to their ability to assess the semantic relatedness of the conversations [127]. In [114], Yassin et al., evaluated their chatbot performance based on the ISO 924 standard, which is a set of international standards for human-computer interaction (HCI) and usability [149]. They emphasize that the three most important characteristics of an interactive system are effectiveness, efficiency and satisfaction. To measure these concepts, they conducted a questionnaire with 14 participants. Noori et al. [82] suggested that both subjective and objective evaluations should be conducted. However, as there are no benchmark metrics, they conducted a survey of 13 questions that combined subjective and objective measures. The majority of the selected articles assessed their chatbot's performance by either creating a survey to gather user feedback and calculate user satisfaction, such as in [72, 75, 80, 82, 85, 86, 99, 102, 109, 114, 119, 127], or by conducting conversation logs,

such as in [73, 90]. In [68, 69, 78, 100, 107], Hijjawi et al., proposed a technique in which they studied the resulting dialogue to compute the Ratio of Matched Utterances (RMUT), considering answer’s accuracy and context respect. Moreover, some studies recommended relying only on automatic metrics, such as accuracy, precision, recall, F1-score, Bleu-score, and perplexity, as indicated in Table 3. This is because these metrics are considered efficient and quantifiable. Additionally, they provide a general overview of the chatbot engine model's performance, indicating whether it is performing well or not. Nevertheless, avoid bias that may be introduced by human evaluators’ as they are affected by individual preferences and subjectivity in their judgments.

Nonetheless, some studies recommended using both human and automatic evaluations to examine the chatbot's performance, benefiting from the advantages of both approaches. For instance, in [87] Bilquise et al., they assessed the model's performance using the accuracy metric, specifically assessing the intent extraction model accuracy. Additionally, they conducted a human evaluation with 37 participants to gather feedback on their chatbot experience. Likewise, in [72]. NA. Alhassan et al., used Precision, Recall and F1-score as automatic metrics and a survey with 6 participants to evaluate the performance of their chatbot manually. Similarly, in [108] Ghaleb et al., in addition to automatic metrics such as accuracy and recall, fetched the conversation histories of users and chatbots to assess the correctness and fluency of the chatbot’s answers. Similarly, in [72] NA. Alhassan et al., used Precision, Recall and F1-score as automatic metrics and a survey with 6 participants to evaluate the performance of their chatbot.

Although significant effort has been devoted to creating a consistent and efficient set of assessment criteria for evaluating chatbot performance, it is worth noting that a significant number of studies (n=14) did not use any metrics, as shown in Table 4.

Table 4. Summary of Challenges and Solutions found in the selected articles.

Challenges	Solution	Articles Reference
1. Challenges encountered with datasets.	Limited availability of Arabic datasets	Collect and process data manually [67] [70-72] [74] [76] [77] [80] [81] [85] [87] [88] [95-99] [101-105] [108] [110] [111] [114] [116-120] [122-125]
		Translate English dataset to Arabic [83] [89] [93]
	Developing a corpus requires a significant amount of time and effort.	Create AIML file automatically [75] [79] [106] [109]
2.Orthography handling and out-of-corpus questions	Building a chatbot with a small dataset	Using Retrieval-Based/Rule-Based approaches [85] [87] [77]
		Joint-training [83]
		N-gram [70]
		Cross-lingual transfer learning [89]
	Handling erroneous or misspelled user inputs	proximity processing with Pool of suggestions [67]
	Utterance validation [68] [73] [119]	
	Create a default corpus [71]	
	Orthographic normalization [86]	
Out of corpus utterances	save it to add it later [81]	
Extract answers	pattern matching [68] [73] [82] [86] [90] [91] [95]	

	without delve into the complexities of Arabic grammar.	key word matching	[71] [72] [114] [119] [120] [127]
		text or word distance (Cos and Jaccard)	[76] [81] [95] [108] [113] [117]
		Semantic key word match using SPQRQL	[98] [118] [126]
	the problem of large morphological and derivative diversity for Arabic words	Orthographic normalization	[86] [104] [115] [118]
		using ISRI stemmer	[87] [91]
		using KHOJA stemmer	[90] [110]
3. Handle the complexities of the Arabic language	Using existing platform to build chatbot	Pandorabots	[75] [79] [86] [106] [109]
		Rasa	[77] [84] [116] [124]
		Wit.ai	[85]
		Dialogue flow	[102]
		IBM Watson	[111]
	Code-mixing and Code-switching.	Scripting patterns manually	[87]
	Lack of an appropriate Arabic resources	using Translation to English support only MSA	[102] [79]
	Make difference between question and non-question utterance	decision tree	[100]
		key word extraction "where" "when" ...	[110]
4. Conversation flow	Conserve the conversation context flow and provide more control over the conversation	temporarily memory	[68] [73]
		Using pre-scripted patterns or scripted knowledge tree	[82]
		Dialogue manager	[80]
		use data-driven approaches	[101] [103]
		clickable follow-up choice buttons	[112]
5. Covering multiple topics	Handling Corpus Limitations	using API to external websites.	[77] [104] [105]
		Conducting conversation logs: the ratio of correct answers.	[68] [69] [73] [78] [87] [90] [100] [107] [108]
		Creating a survey to gather user feedback and calculate user satisfaction	[72] [75] [80] [82] [85] [86] [99] [102] [109] [114] [119] [127]
		Accuracy	[70] [71] [84] [87] [88] [96] [98] [101] [108] [116] [121] [123]
6. Chatbot			[124] [130]

Evaluation Metrics	Evaluating the chatbot performance	Precision	[72] [96] [113] [117] [118] [128] [129]
		Recall	[71] [72] [118] [128]
		F1-score	[72] [84] [91] [96] [103] [118]
		BLEU metric	[83] [89] [93] [94] [122]
		Perplexity	[94] [95]
		Not available	[76] [77] [81] [92] [95] [97] [100] [104] [105] [106] [110] [111] [115] [126]

4.4 RISK OF BIAS IN STUDIES

The assessment results of the studies analyzed in this review are summarized in Table 5. According to the quality assessment criteria we established, the majority of the selected articles (n=57) were categorized as good quality, while five (n= 5) articles received a medium quality rating, and two (n=2) articles were deemed low-quality. Table 5 offers a detailed breakdown of the evaluation outcomes for each checklist item, forming the basis for the quality rating assigned to each article.

Table 5. Quality Assessment results

Article Ref	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Score
67	1	1	1	1	0	0	0	1	1	0	1	0	0	53 %
68	1	1	1	1	0	1	0	1	1	1	1	1	1	84 %
69	1	1	1	1	0	1	0	1	1	1	1	1	0	76 %
70	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
71	1	1	1	1	1	1	0	0	1	1	1	1	0	76 %
72	1	1	1	1	1	1	1	1	1	0	1	1	1	92 %
73	1	1	1	1	1	1	0	1	1	1	1	1	1	92 %
74	1	1	1	1	1	1	1	1	1	0	1	0	1	84 %
75	1	1	1	1	1	1	0	1	1	1	1	1	0	84 %
76	1	1	1	1	0	0	0	1	1	1	1	0	1	69 %
77	1	0	1	1	1	0	1	1	1	1	1	0	1	76 %
78	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
79	1	0	0	1	1	0	0	1	1	1	0	0	0	46 %
80	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
81	1	0	1	0	0	0	0	1	1	1	0	1	0	46 %
82	1	1	1	1	1	1	1	1	1	0	1	0	1	84 %
83	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
84	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
85	1	0	1	1	1	1	1	1	1	0	1	0	1	76 %
86	1	1	1	0	1	1	0	1	1	1	0	1	1	76 %
87	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
88	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
89	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
90	1	1	1	1	0	1	0	1	1	0	1	1	1	76 %
91	1	1	1	1	0	1	1	1	1	0	1	1	1	84 %
92	1	0	1	1	0	0	1	1	1	1	1	0	0	61 %
93	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
94	1	1	1	1	1	1	1	1	0	0	1	1	1	84 %
95	1	1	1	1	0	0	0	1	1	1	1	1	1	76 %
96	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
97	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
98	0	1	1	1	1	1	0	1	1	1	1	0	0	69 %

99	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
100	1	1	1	1	0	0	0	1	1	1	1	1	1	76 %
101	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
102	1	0	1	1	0	1	0	1	1	1	1	0	1	69 %
103	1	1	1	1	1	1	1	1	1	1	1	0	1	92 %
104	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
105	1	1	1	1	1	0	0	1	1	1	1	0	0	69 %
106	1	1	1	1	1	0	0	1	1	1	1	0	1	76 %
107	1	1	1	1	1	1	0	1	1	1	1	0	0	76 %
108	1	1	1	1	1	1	1	1	1	1	1	0	1	92 %
109	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
110	1	1	1	1	1	0	0	1	1	1	1	1	0	76 %
111	1	0	1	1	0	0	0	1	1	1	1	0	0	53 %
112	1	0	1	1	1	1	0	1	1	0	1	0	0	61 %
113	1	1	1	1	1	1	0	1	1	0	1	1	1	84 %
114	1	1	1	1	1	1	0	1	1	0	1	0	1	76 %
115	1	1	1	1	0	0	1	1	1	0	1	1	1	76 %
116	1	0	1	1	1	1	1	1	1	0	1	1	1	84 %
117	1	1	1	1	1	1	0	1	1	0	1	1	1	84 %
118	1	1	1	1	0	1	1	1	1	1	1	1	1	92 %
119	1	1	1	1	0	1	1	1	1	1	1	0	1	84 %
120	1	1	1	1	0	1	1	1	1	1	1	0	1	84 %
121	1	1	1	1	1	1	1	1	1	0	1	1	1	92 %
122	1	1	1	1	1	1	1	1	1	1	1	1	1	100 %
123	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
124	1	0	1	1	1	1	1	1	1	1	1	0	1	84 %
125	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
126	1	1	1	1	1	0	0	1	1	1	1	0	1	76 %
127	1	1	1	1	1	1	0	1	1	1	1	0	1	84 %
128	1	1	1	1	0	1	0	1	1	1	1	0	1	76 %
129	1	1	1	1	0	1	0	1	1	1	1	0	1	76 %
130	1	1	1	1	1	1	0	1	1	0	1	0	1	76 %

5. DISCUSSION

The purpose of this study was to conduct a systematic review of the literature on Arabic chatbots to gain a better understanding of their current status, challenges encountered, solutions proposed and future potential. Two broad research questions were specified in relation to the objectives of this SLR. To answer these questions 64 article were examined. In this section we will be discussing the different solution proposed to overcome the challenges encounter developing an Arabic chatbot, investigating their advantageous and disadvantageous.

5.1 DATASETS AVAILABILITY

Many studies have addressed the challenge of limited available datasets by manually collecting data. Although the solution is fulfilling and aiding the researchers in overcoming this challenge, it does not significantly contribute to the advancement of the Arabic chatbot field because, in addition to the amount of the built datasets, which are usually small, not variance and limited by human capacity, the created datasets are customized and tailored exclusively to the chatbot's objectives and the tasks it needs to perform effectively. This customization ensures that the words, sentences, examples and expressions that users are likely to use included in the dataset are relevant and align with the chatbot's intended functionality. Therefore, using existing datasets may require more preprocessing, even within the same domain with the same objectives.

Moreover, some researchers have recommended training chatbot on English datasets and using translation model for Arabic sentences. This involves translating Arabic utterances into English, processing, generating responses, and then translating them back into Arabic before being sent to the user. Although this solution is effective, it eliminates the need to process Arabic content and handle its morphology complexities. Additionally, training Chatbot on English could be advantageous, given that there are sufficient resources, including datasets, pre-trained model, and API which can enhance the its capability. However, translating Arabic utterances into English (translating between two languages in general) will impact the chatbot by causing a loss of nuances, idiomatic expressions, and cultural context, potentially affecting the quality of the conversation and making it incoherent. Furthermore, the process of translation introduces

complications and delays in the interactions with chatbots, which may not be suitable for tasks requiring real-time responses or high-speed processing.

Therefore, it is crucial to focus on creating appropriate Arabic datasets. This would enable researchers to devote their time and resources to building and advancing sophisticated models, rather than wasting time on collecting, pre-processing and labelling small datasets.

In light of the challenges posed by low-quality datasets, we propose the following recommendations: Firstly, it is imperative to collect all the existing small, low-quality datasets, categorize them according to their respective domains, and enrich the dataset by incorporating additional data. Moreover, standardize all the data to create a high-quality dataset suitable for exploitation by all researchers. Secondly, when constructing a rule-based or retrieval-based chatbot, it is advisable to integrate a lexical database. This integration empowers the chatbot to respond to users' utterances that do not strictly conform to predefined patterns. Thirdly, consider using movie scripts, books and play scripts as a knowledge base. This will enhance the development of high-quality datasets as they contain valuable dialogues that closely resemble real human conversations, especially when building AI-based or generative chatbots. Additionally, benefits from other high-level language datasets by translating them into Arabic with preprocessing them beforehand to maintain cultural nuances and contextual relevance in Arabic conversations. Finally, researchers should give consideration to dialect datasets, as they enable chatbots to mimic human-like conversations and improve their understanding of user utterances. This avenue of research deserves further exploration, whereby researchers can delve into the development of sophisticated dialect datasets.

5.2 APPROACHES EMPLOYED

The scarcity of Arabic datasets and the complexity of Arabic morphology have led the majority of researchers to use pre-scripted rule approaches (i.e., Retrieval-based and Rule-based approaches), as they neither require massive data nor delve into grammatical or linguistic details, in addition they have demonstrated their effectiveness in information extraction and responding to predefined queries. However, these approaches have some disadvantages that warrant consideration.

Both rule-based and retrieval-based approaches typically employ either pattern matching or sentence similarity algorithms. Although pattern matching algorithms are considered one of the most successful methods for developing Chatbot, the main disadvantage of this approach is the extensive time and effort needed to script all possible patterns and cover the entire spectrum of linguistic variations. This difficulty stems from the fact that users can phrase their utterances in various ways, making it challenging to predict and accommodate all potential interactions with the chatbot. On the other hand, sentence similarity algorithms are more effective as they reduce the scripting effort to a minimum. However, given the limited availability of Arabic language resources, such as Arabic WordNet and Lexicon Corpora, this presents a significant challenge. Thus, when deploying these algorithms on Arabic texts, researchers primarily rely on lexical matching rather than semantic matching. Nonetheless, because of the Arabic nature, which is known for its synonym diversity and rich vocabulary, relying only on lexical matching may not aid the objective of building a high-quality chatbot because this approach involves comparing words or phrases based on their exact spelling or form without considering their meaning. Consequently, even if the user's utterance is contextually the same as in the corpus, the chatbot may not find a match.

Therefore, working on providing a high-quality Arabic lexicon and making it available for researchers can be considered a valuable contribution to Arabic chatbot research.

Furthermore, pre-scripted rule approaches are not designed to utilize historical and memorized data. Thus, the generated responses usually depend on the current questions. As a result, chatbots using these approaches may gradually lose the flow of conversation as they lack the ability to recall previous interactions or adapt responses based on the evolving discussion. Consequently, interacting with such chatbots can feel robotic and disjointed to users since they miss out on the natural fluidity and coherence typically found in human conversations.

On the other hand, more advanced chatbot models that incorporate machine learning techniques such as RNN, LSTM, and SEQ2SEQ can make use of historical data and also memorize important features during the conversation to provide responses that are more natural and relevant within the given context. This enhances overall conversational experiences by offering a higher level of contextual understanding. However, the primary limitation of these approaches is the requirement for high-quality and massive datasets. Additionally, there is the inherent complexity associated with training these models with Arabic content, as they are usually designed and optimized for English.

In order to leverage the advantages of both approaches, Hybrid approaches are particularly well-suited, especially in the context of Arabic. Hybrid approaches combine pre-scripted techniques, which do not demand massive datasets, with machine learning model, as they are capable of preserving conversation context and coherence. However, among the selected articles, only one study was identified that employed hybrid approaches. Therefore, there should be greater emphasis on developing Arabic chatbots using hybrid methods.

Many researchers have opted to use third-party platforms like IBM Watson or Pandorabots for developing their chatbots. While these platforms reduce time and simplify chatbot development, they have limitations. Notably, not all languages are supported; for example, some chatbot platforms do not support Arabic (e.g., Google's DialogFlow platform). Moreover, these platforms do not produce independent chatbots, users need to have accounts on social media

platforms such as Facebook, Slack and WeChat in order to interact with the bot. Additionally, these platforms are using public cloud services that are not free, thus developers are required to renew subscription fees. Furthermore, the chatbot developed using these platforms will always be constrained by the techniques they offer. For instance, if the platform only offers pattern matching for chatbot development, the resulting chatbot will be restricted to this specific approach. Consequently, errors may be increased, especially in the context of Arabic, where not all methods are applicable. Nonetheless, it is noteworthy that using existing platforms may not be considered a significant and valuable contribution to the field of Arabic chatbot research.

To enhance the research in this area, it is advisable for researchers to focus on building their own chatbots using advanced models and techniques, as this allows for greater customization and adaptation to specific language structure, grammar, and context requirements.

5.3 HUMAN LIKE CONVERSATION

Simulating human conversations can be a challenging task for chatbots. Many studies have used pre-scripted rule approaches to build chatbots in the form of question-answers. Although these approaches can be beneficial as they provide more control over the conversation, the fact that they rely only on predefined responses and do not generate new responses based on the evolving context means that the conversation may start to feel repetitive, resembling a loop.

Moreover, human conversations are dynamic, flexible interactions with varied purposes, not strictly question-answer-oriented. They often take informal, open-ended forms driven by social or communicative goals, which may differ from the structured and goal-oriented nature of chatbot interactions. This difference in conversational dynamics is an important consideration in chatbot design and development to create a more authentic and human-like chatbot.

The current state of the field highlights a strong demand for generative chatbots that can mimic human-like conversation. Achieving this requires building chatbots based on more advanced methods, such as LSTM, Seq2Seq, and Transformers, which have the capability to learn and generate new responses while respecting the conversation context.

Furthermore, human conversations exhibit distinctive characteristics, such as the personas of the participants, emotional expression, empathy and ethical considerations. These elements are essential contributors to the coherence and fluency of a conversation. However, the pre-scripted rules approach often does not encompass all of these features, as they tend to prioritize extracting the correct answer over providing an answer suitable for the user's specific situation or emotional state, which can result in less natural or less emotionally nuanced interactions.

Additionally, numerous studies have utilized web scraping, questionnaires, surveys, or online resources to collect datasets. The primary drawback of these methods is that chatbots trained on such datasets incorporate opinions and responses from diverse participants, which can significantly diminish their ability to maintain a consistent persona. Therefore, before training the chatbot on the collected dataset, it should undergo preprocessing to ensure consistency in persona. This approach will enhance the chatbot's ability to engage users in more satisfying and fluent conversations. Moreover, incorporating emotional and empathy datasets can significantly improve the conversation's coherence, contextuality, and overall user experience, making it less mechanical and more satisfying. Therefore, further research in these areas is imperative to advance the field of Arabic chatbots.

Nevertheless, it's essential for the chatbot to support the common language used by the majority of users. Even though many Arabic native speakers use dialectal Arabic, the majority of built chatbots tend to support only MSA and CA. This preference is due to the structural nature and the availability of more resources for MSA and CA in comparison to dialectal Arabic. Therefore, it is essential to consider extending the capabilities of the chatbot to include dialectal Arabic in addition to MSA or CA. This expansion will enhance the chatbot's ability to respond to a broader range of user utterances and minimize error rates. By training the chatbot on the various Arabic dialects, it can provide more accurate and contextually relevant responses, ultimately increasing user satisfaction.

5.4 VOICE BASED CHATBOT

Voice-based chatbots are considered as an advanced version of text-based chatbots, they have many advantages, such as offering more natural and conversational interaction for users, reducing the time required writing and reading and minimizing errors arising from typographical mistakes. Despite the considerable attention and progress that this type of chatbot has received in the English language, its development in the Arabic remains at its infancy stages. Among the 64 articles reviewed, only six (n=6) articles were dedicated to the development of voice-based chatbots. In [92], the authors employed voice exclusively for delivering responses, the chatbot received user input in text, processed it, generated a response in a text format then used a model to convert the text into voice before being sent to the user. In [104], the researchers integrated the pre-built model Microsoft Azure Cognitive Services - Speech Service, to develop a voice-based chatbot. Similarly, in [116] utilized Microsoft Azure for speech recognition, in conjunction with the RASA platform, to construct a voice-based chatbot. Notwithstanding their recognized contributions, the voice-based chatbots that were developed utilizing external models or pre-existing techniques are not considered a valuable contribution to the Arabic chatbot field, because none of the authors created a customized voice model tailored to address the unique linguistic features and challenges presented by the Arabic language.

This bias can be attributed to the scarcity of available labeled voice-based datasets and the intricacies involved in training models on voice data, particularly within the context of the Arabic language. It is noteworthy that the majority of models enabling chatbots to use voice are primarily designed for the English language. Therefore, further research in these areas is imperative to advance the field of Arabic chatbots.

5.5 EVALUATION METRICS AND USERS' DATA CONFIDENTIALITY

Evaluating the performance of a chatbot is regarded as one of the most important steps, as it provides insights into the results, functionality, fluency and quality of the chatbot. However, in light of the absence of standardized methods for evaluating chatbots, the majority of authors opt for human-based metrics via reviewing conversation logs and investigating participant feedback regarding their interactions with the chatbot. Although these techniques are deemed suitable for evaluating chatbots as they offer genuine feedback on real conversations, it is imperative to consider some noteworthy points.

Firstly, the representativeness of the participant sample, quantity and diversity should be ensured. The majority of studies that employed human-based metrics did not provide detailed information regarding the representativeness of the utilized sample. Moreover, numerous studies employed non-representative samples. For instance, some studies tested their chatbot performance with only six ($n = 6$) participants. Furthermore, when using surveys or questionnaires, the specific methodologies employed are frequently left undisclosed and tend to lack representativeness. Therefore, whether utilizing surveys or conversation log techniques to assess the performance of a chatbot, it becomes imperative to follow a rigorous and well-defined methodology and make it available to researchers.

Secondly, the confidentiality of users' data is of paramount importance. In order to evaluate the chatbot's performance, certain studies have examined the resulting conversation logs to calculate metrics such as the percentage of accurate responses, contextual coherence and fluency. However, the majority of these studies do not seek to obtain explicit user permission, which poses a potential breach of user data confidentiality. This concern becomes especially pertinent when the chatbot in question requires access to sensitive information or is designed to provide empathetic responses. Users may not be willing to share their thoughts and opinions without their explicit permission. Moreover, some studies integrate external APIs to augment their chatbots with a broad range of topics. Thus, maintaining the privacy of user data while accessing external resources is crucial. Therefore, it is imperative to consider this issue meticulously when building or evaluating chatbots using human-based metrics.

Thirdly, some studies have evaluated their chatbot performance based on the metrics of the developed model. It is essential to note that relying solely on automatic metrics may not provide a comprehensive indicator of the actual chatbot's performance. While these metrics indicate how well the model performs on the training dataset, they may not accurately predict the model's behavior in real-world conversational settings. Therefore, for a comprehensive assessment, it is advisable to consider additional metrics.

In this context, we propose the incorporation of the following metrics in addition to the existing ones for a comprehensive evaluation of chatbot performance: **a).** The number of iterations required to extract the targeted answer. For further clarification, consider the following scenario: Some chatbots have the ability to capture and extract all the required information from a single sentence (e.g., Chatbot 1: "Please provide me with your personal information"), while others extract information from each separate sentence (e.g., iteration (1): Chatbot 2: "What is your name?"; iteration (2): Chatbot 2: "What is your age?"). Although both versions can capture and extract information, Chatbot 1's approach is more natural, simulating human conversations and it is also more time-efficient. **b).** The ability to comprehend untrained data. Due to the richness of Arabic vocabulary, scripting all patterns or training a chatbot to account for every sentence variation is an insurmountable task. Thus, chatbots employing lexicon-based or ontology-based techniques to comprehend user utterances not present in the training data are considered advantageous. **c).** The fluency, contextuality and coherence of the chatbot's responses are also vital factors to consider because the quality of the user experience is influenced by the chatbot's ability to engage in meaningful and natural conversations. **d).** The persona of the chatbot, its ability to respond with empathy and its capacity to remember, orient and adapt the conversation to different user profiles, emotional states, or contexts are key to providing a personalized and supportive experience and user-friendly conversations. **e).** The time spent processing, extracting and generating responses is also an important factor, considering the real-time and instantaneous nature of human conversations. These considerations contribute to a more meaningful and successful conversation with the chatbot.

Lastly, exclusive reliance on either human-based or automatic metrics for assessing chatbot performance may fall short of addressing the full spectrum of evaluation needs and may not capture the authentic performance of the developed chatbot. Thus, incorporating both human-based metrics and automatic metrics can provide a more holistic and accurate assessment of the chatbot's performance. Moreover, the adopted metrics should adhere to a rigorous methodological framework because this approach serves to not only minimize potential bias but also contributes to the advancement of the Arabic chatbot field by promoting standardized and reliable evaluation metrics.

Strengths and limitations of this review

To the best of our knowledge, this is the first systematic literature review that discusses the challenges faced in developing an Arabic chatbot and provides an in-depth investigation of the proposed solutions. This review provides

useful guidance for researchers by identifying the key factors that contribute to the development of intelligent, human-like Arabic chatbots. It also outlines suitable approaches to achieve this goal. Nonetheless, our review has certain limitations. Firstly, due to restricted access, certain articles were not included in the review, which may result in the exclusion of relevant information. Secondly, due to the vast range of challenges outlined in the articles, we concentrated solely on the most significant ones. Exploring further challenges could provide more substantial insights into the research field. Thirdly, this systematic review examines the research questions without conducting empirical research. Future research should include empirical contributions for a more comprehensive study of the development of Arabic chatbots.

Implications of the results for future research

This study presents the latest research findings and categorizes previous studies on Arabic Chatbots. It offers a structured and comprehensive understanding of Arabic language features, challenges hindering the development of intelligent Arabic chatbots, and discusses solutions proposed by researchers in the selected articles. Additionally, it highlights suitable approaches for building an efficient chatbot. The study also suggests new ideas to advance the Arabic chatbot domain, bridging the gap with high-resource language chatbots like English. Furthermore, this SLR contributes to the Arabic Chatbot field by providing a thorough analysis of different approaches in the reviewed articles. It covers their applications, advantages, disadvantages and results. Therefore, this will enable researchers to gain insights into the approaches they intend to employ, streamlining their efforts toward innovating and developing new models and approaches to enhance the Arabic chatbot field, rather than reusing existing ones.

6. CONCLUSION

In this study, we presented a systematic literature review of Arabic chatbots, investigating and discussing the challenges of this field and the proposed solutions. The PRISMA systematic review protocol was used to analyze 64 articles from well-known digital databases: Scopus, Science Direct, Web of Science, PubMed, SpringerLink, IEEE Xplore, ACM, Ebsco, and ICI. The snowballing technique was also used to discover supplementary relevant research. Additionally, this study includes articles from 2000 to 2023. This study aims to contribute to the Arabic chatbot research field by presenting a comprehensive in-depth analysis, exploring a spectrum of topics regarding the development of Arabic chatbots, exploring the datasets used, the techniques and approaches employed, the challenges, investigating and comparing the proposed solutions, and the evaluation processes used to measure the chatbot's performance. The findings show that the major challenge encountered by all studies was the scarcity of high-quality datasets. To overcome this challenge, most of these studies involved collecting and preprocessing data manually; however, this approach is limited by human capacity, and as a result, the built data sets are generally not representative. Additionally, dialectal datasets, Arabic lexicon anthologies, and Arabic empathic datasets are still in their infancy stages. Due to their importance, they enable a chatbot to respond in a more humanized and emotionally intelligent manner; thus, more effort should be dedicated to this research direction. Moreover, the complexity of Arabic morphology led many researchers to adopt pre-scripted rules approaches, and other researchers opted for the use of third-party platforms such as RASA as they facilitate the development process and also use external APIs to enhance their chatbot capability. Only a few were building generative chatbots. Hybrid approaches, on the other hand, prove to be effective; thus, more effort should be directed into developing Arabic chatbots using hybrid approaches. Furthermore, due to the absence of standardized methods for assessing chatbot performance, this field is an open research direction.

This SLR offers a range of insightful recommendations for forthcoming research endeavors, presenting an opportunity for researchers to further advance the field of chatbot research.

Funding

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

- [1] M. Hijjawi, H. Qattous, and O. Alsheiksalem, "Mobile Arabchat: An Arabic Mobile-Based Conversational Agent," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 10, 2015, doi: 10.14569/IJACSA.2015.061016.
- [2] Y. Wardat, M. A. Tashtoush, R. AlAli, and S. Saleh, "Artificial Intelligence in Education: Mathematics Teachers' Perspectives, Practices and Challenges," *Iraqi Journal For Computer Science and Mathematics*, vol. 5, no. 1, pp. 60–77, Jan. 2024, doi: 10.52866/IJCSM.2024.05.01.004.
- [3] A. S. Hashim, A. Aminu Muazu, M. A. M. Yusof, and N. I. Arshad, "Development of Robot to Improve Learning of Programming Skills among Students," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 3, pp. 1–11, Jun. 2023, doi: 10.52866/IJCSM.2023.02.03.001.
- [4] Matt Moran, "25+ Top Chatbot Statistics For 2024: Usage, Demographics, Trends," *Startup Bonsai*. Accessed: Jun. 04, 2023. [Online]. Available: <https://startupbonsai.com/chatbot-statistics/>
- [5] Bharathi Ramadass, "The Truth About Chatbots," *Forebs*. Accessed: May 24, 2024. [Online]. Available: <https://www.forbes.com/sites/servicenow/2022/01/21/the-truth-about-chatbots/?sh=592d2882797d>
- [6] G. Caldarini, S. Jaf, and K. McGarry, "A Literature Survey of Recent Advances in Chatbots," *Information (Switzerland)*, vol. 13, no. 1, Jan. 2022, doi: 10.3390/info13010041.
- [7] I. Shah, S. Jhawar, A. Khater, A. Jacob, and Dr. G. Potdar, "Chatbot Development Through the Ages : A Survey," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 9, no. 3, pp. 262–271, Jun. 2023, doi: 10.32628/CSEIT2390329.
- [8] T. Hu et al., "Touch your heart: A tone-aware chatbot for customer care on social media," In *Proc. Conference on Human Factors in Computing Systems*, vol. 2018-April, Apr. 2018, doi: 10.1145/3173574.3173989.
- [9] A. S. Alsheddi and L. S. Alhenaki, "English and Arabic Chatbots: A Systematic Literature Review," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 662–675, Oct. 2022, doi: 10.14569/IJACSA.2022.0130876.
- [10] S. AlHumoud, A. Al Wazrah, and W. Aldamegh, "Arabic Chatbots: A Survey," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 535–541, 2018, doi: 10.14569/IJACSA.2018.090867.
- [11] Z. N. Abdulkader and Y. F. M. Al-Irhayim, "A Review of Arabic Intelligent Chatbots: Developments and Challenges," *Al-Rafidain Engineering Journal (AREJ)*, vol. 27, no. 2, pp. 178–189, Sep. 2022, doi: 10.33899/RENGJ.2022.132550.1148.
- [12] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouni, A. Abd-alrazaq, and M. Househ, "Arabic chatbot technologies: A scoping review," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100057, Jan. 2022, doi: 10.1016/J.CMPBUP.2022.100057.
- [13] E. S. AlHagbani and M. B. Khan, "Challenges facing the development of the Arabic chatbot," *SPIE digital library*, vol. 10011, pp. 192–199, Jul. 2016, doi: 10.1117/12.2240849.
- [14] M. I. A. Almurayh, "The Challenges of Using Arabic Chatbot in Saudi Universities," *IAENG Int J Comput Sci*, vol. 48, no. 1, 2021.
- [15] 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.
- [16] O. I. Obaid, A. H. Ali, and M. G. Yaseen, "Impact of Chat GPT on Scientific Research: Opportunities, Risks, Limitations, and Ethical Issues," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 4, pp. 13–17, Sep. 2023, doi: 10.52866/IJCSM.2023.04.04.002.
- [17] K. Crockett, J. O'Shea, and Z. Bandar, "Goal Orientated Conversational Agents: Applications to Benefit Society," Included in the following conference series: *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, vol. 6682, pp. 16–25, 2011, doi: 10.1007/978-3-642-22000-5_3.
- [18] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Mag*, vol. 27, no. 4, pp. 12–12, Dec. 2006, doi: 10.1609/AIMAG.V27I4.1904.
- [19] P. Johri, S. K. Khatari, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural Language Processing: History, Evolution, Application, and Future Work," In *Proc. Proceedings of 3rd International Conference on Computing Informatics and Networks*, vol. 167, pp. 365–375, 2021, doi: 10.1007/978-981-15-9712-1_31.
- [20] John Hutchins, "The history of machine translation in a nutshell," 2006.
- [21] D. G. Bobrow, "Natural Language Input for a Computer Problem Solving System," *Massachusetts Institute of Technology (MIT) Libraries*, Mar. 1964.
- [22] J. Weizenbaum, "ELIZA-A computer program for the study of natural language communication between man and machine," *Commun ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, doi: 10.1145/365153.365168.

- [23] 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.
- [24] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in *Proc. IFIP Advances in Information and Communication Technology*, Springer, 2020, pp. 373–383. doi: 10.1007/978-3-030-49186-4_31/FIGURES/3.
- [25] B. A. Shawar and E. Atwell, "Using dialogue corpora to train a chatbot," in *Proceedings of the Corpus Linguistics*, Lancaster University, Jan. 2003, pp. 681–690.
- [26] K. Moore et al., "A comprehensive solution to retrieval-based chatbot construction," *Comput Speech Lang*, vol. 83, p. 101522, Jan. 2024, doi: 10.1016/J.CSL.2023.101522.
- [27] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A Survey of Design Techniques for Conversational Agents," in *Proc. Communications in Computer and Information Science*, Springer, Singapore, 2017, pp. 336–350. doi: 10.1007/978-981-10-6544-6_31.
- [28] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A Survey on Conversational Agents/Chatbots Classification and Design Techniques," in *Proc. Advances in Intelligent Systems and Computing*, Springer, Cham, 2019, pp. 946–956. doi: 10.1007/978-3-030-15035-8_93.
- [29] 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.
- [30] A. Montejo-Ráez, S. María Jiménez-Zafra, A. Fuad, and M. Al-Yahya, "AraConv: Developing an Arabic Task-Oriented Dialogue System Using Multi-Lingual Transformer Model mT5," *Applied Sciences* 2022, Vol. 12, Page 1881, vol. 12, no. 4, p. 1881, Feb. 2022, doi: 10.3390/APP12041881.
- [31] A. Fuad and M. Al-Yahya, "Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5," *Mathematics* 2022, Vol. 10, Page 746, vol. 10, no. 5, p. 746, Feb. 2022, doi: 10.3390/MATH10050746.
- [32] S. M. Yassin1 and M. Z. Khan, "SeerahBot: An Arabic Chatbot About Prophetâ€™s Biography," *International Journal of Innovative Research in Engineering and Management*, vol. 9, no. 2, pp. 89–97, Apr. 2021, doi: 10.21276/IJIRCST.2021.9.2.13.
- [33] B. A. Shawar and E. Atwell, "ALICE Chatbot: Trials and Outputs," *Computación y Sistemas*, vol. 19, no. 4, pp. 625–632, Dec. 2015, doi: 10.13053/cys-19-4-2326.
- [34] R. Schwitter, F. Rinaldi, and S. Clematide, "An Arabic chatbot giving answers from the Qur'an," in *Question-Answering workshop of TALN 04: XI Conference sur le Traitement Automatique des Langues Naturelles*, Fès, Morocco: ATALA, 2004, pp. 451–460.
- [35] 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 *Proc. 2021 9th International Conference on Information and Education Technology*, ICIET 2021, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 139–144. doi: 10.1109/ICIET51873.2021.9419656.
- [36] A. S. Alsheddi and L. S. Alhenaki, "English and Arabic Chatbots: A Systematic Literature Review," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 662–675, Oct. 2022, doi: 10.14569/IJACSA.2022.0130876.
- [37] 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.
- [38] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput Intell Mag*, vol. 9, no. 2, pp. 48–57, 2014, doi: 10.1109/MCI.2014.2307227.
- [39] M. McShane, "Natural Language Understanding (NLU, not NLP) in Cognitive Systems," *AI Mag*, vol. 38, no. 4, pp. 43–56, Dec. 2017, doi: 10.1609/AIMAG.V38I4.2745.
- [40] Anush Fernandes, "NLP, NLU, NLG and how chatbots work," *mystery write, connect, Inspire*. Accessed: May 20, 2023. [Online]. Available: <https://yourstory.com/mystory/fac4d8fd9f-nlp-nlu-nlg-and-how>
- [41] B. A. Alazzam, M. Alkhatib, and K. Shaalan, "Artificial Intelligence Chatbots: A Survey of Classical versus Deep Machine Learning Techniques," *Information Sciences Letters*, vol. 12, no. 4, pp. 1217–1233, Apr. 2023, doi: 10.18576/isl/120437.
- [42] 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, Jan. 2021, doi: 10.1109/ACCESS.2021.3049732.
- [43] W. Maeng and J. Lee, "Designing a Chatbot for Survivors of Sexual Violence: Exploratory Study for Hybrid Approach Combining Rule-based Chatbot and ML-based Chatbot," in *5th Asian CHI Symposium 2021*, Association for Computing Machinery, Inc, May 2021, pp. 160–166. doi: 10.1145/3429360.3468203.
- [44] S. El-Kateb et al., "Arabic WordNet and the Challenges of Arabic," in *Proc. BCS International Academic Conference*, 2006.
- [45] M. M. Najeeb, A. A. Abdelkader, and M. B. Al-Zghoul, "Arabic Natural Language Processing Laboratory serving Islamic Sciences," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 3, 2014, doi: 10.14569/IJACSA.2014.050316.

- [46] M. K. Al-Ajlouny, "GENDER IN ENGLISH AND ARABIC," *Journal of International Scientific Publications: Language, Individual & Society*, vol. 8, pp. 236–244, Aug. 2014, [Online]. Available: <https://www.scientific-publications.net/en/article/1000339/>
- [47] M. Alawneh, N. Omar, and T. Sembok, "MACHINE TRANSLATION FROM ENGLISH TO ARABIC," in *Proc. International Conference on Biomedical Engineering and Technology*, 2013.
- [48] M. Mansour, "The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus," *Int J Humanit Soc Sci*, vol. Vol. 3, no. 12, pp. 81–90, Jun. 2013.
- [49] FarghalyAli and ShaalanKhaled, "Arabic Natural Language Processing," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 9–10, Dec. 2009, doi: 10.1145/1644879.1644881.
- [50] Z. N. Abdulkader and Y. F. M. Al-Irhayim, "A Review of Arabic Intelligent Chatbots: Developments and Challenges," *Al-Rafidain Engineering Journal (AREJ)*, vol. 27, no. 2, pp. 178–189, Sep. 2022, doi: 10.33899/RENGJ.2022.132550.1148.
- [51] M. M. Biltawi, S. Tedmori, and A. Awajan, "Arabic Question Answering Systems: Gap Analysis," *IEEE Access*, vol. 9, pp. 63876–63904, 2021, doi: 10.1109/ACCESS.2021.3074950.
- [52] K. Sundus, F. Al-Haj, and B. Hammo, "A Deep learning approach for Arabic text classification," in *Proc. 2nd ICTCS 2019*, Amman, Jordan: Institute of Electrical and Electronics Engineers Inc., Oct. 2019. doi: 10.1109/ICTCS.2019.8923083.
- [53] B. A. Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 6, no. 1, pp. 37–43, Mar. 2011, doi: 10.3991/IJET.V6I1.1502.
- [54] A. Moubaidin, O. Shalbak, B. Hammo, and N. Obeid, "Arabic Dialogue System for Hotel Reservation based on Natural Language Processing Techniques," *Computación y Sistemas*, vol. 19, no. 1, pp. 119–134, Mar. 2015, doi: 10.13053/cys-19-1-1962.
- [55] S. Mayhew, T. Tsygankova, and D. Roth, "ner and pos when nothing is capitalized," in *Proc. EMNLP-IJCNLP 2019*, Association for Computational Linguistics, Mar. 2019, pp. 6256–6261. doi: <https://doi.org/10.48550/arXiv.1903.11222>.
- [56] S. P. L. A. Thawra Kadeed Shadi Abras, "Construction of Arabic Interactive Tool Between Humans and Intelligent Agents," in *Proc. The 16th International Conference on Human-Computer Interaction*, Springer, Creta Maris, Heraklion, Crete, Greece, 2014.
- [57] M. Mustafa et al., "A Comparative Survey on Arabic Stemming: Approaches and Challenges," *Intell Inf Manag*, vol. 9, no. 2, pp. 39–67, Mar. 2017, doi: 10.4236/IIM.2017.92003.
- [58] 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 *Proc. 2017 IEEE International Conference on CIVEMSA*, Annecy, France, Jul. 2017, pp. 24–29. doi: 10.1109/CIVEMSA.2017.7995296.
- [59] N. A. Alhassan, A. Saad Albarrak, S. Bhatia, and P. Agarwal, "A Novel Framework for Arabic Dialect Chatbot Using Machine Learning," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1844051.
- [60] K. Versteegh, "Linguistic Contacts Between Arabic and Other Languages," *Arabica*, vol. 48, no. 4, pp. 470–508, Dec. 2001, doi: 10.1163/157005801323163825.
- [61] E. T. Rother, "Systematic literature review X narrative review," *ACTA Paulista de Enfermagem*, vol. 20, no. 2, 2007, doi: 10.1590/S0103-21002007000200001.
- [62] T. Martínez-Ruiz, J. Münch, F. García, and M. Piattini, "Requirements and constructors for tailoring software processes: A systematic literature review," *Software Quality Journal*, vol. 20, no. 1, pp. 229–260, Jun. 2012, doi: 10.1007/S11219-011-9147-6/FIGURES/13.
- [63] M. J. Page et al., "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N160.
- [64] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. A. Babar, and H. Tang, "Quality assessment of systematic reviews in software engineering: A tertiary study," in *Proc. ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2015. doi: 10.1145/2745802.2745815.
- [65] Corporation for Digital Scholarship, "Zotero | Your personal research assistant." 2006. Accessed: Jan. 12, 2023. [Online]. Available: <https://www.zotero.org/>
- [66] T. E. Vanhecke, "Zotero," *Journal of the Medical Library Association : JMLA*. Accessed: Jun. 14, 2023. [Online]. Available: [/pmlc/articles/PMC2479046/](https://pmlc/articles/PMC2479046/)
- [67] 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," *Webology*, vol. Volume 18, no. No. 2, pp. 946–954, Dec. 2021, doi: 10.14704/WEB/V18I2/WEB18365.
- [68] M. Hijjawi, Z. Bandar, K. Crockett, and D. McLean, "ArabChat: An arabic conversational agent," in *Proc. 6th International Conference on CSIT , IEEE Computer Society*, 2014, pp. 227–237. doi: 10.1109/CSIT.2014.6806005.

- [69] M. Hijjawi, Z. Bandar, and K. A. Crockett, "A Novel Hybrid Rule Mechanism for the Arabic Conversational Agent ArabChat," *Global Journal on Technology*, no. 08, pp. 185–194, 2015, [Online]. Available: <http://awercenter.org/gjt>
- [70] M. Habib, M. Faris, R. Qaddoura, A. Alomari, and H. Faris, "A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context," *IEEE Access*, vol. 9, pp. 85690–85708, 2021, doi: 10.1109/ACCESS.2021.3087593.
- [71] B. A. Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 6, no. 1, pp. 37–43, 2011, doi: 10.3991/IJET.V6I1.1502.
- [72] N. A. Alhassan, A. Saad Albarrak, S. Bhatia, and P. Agarwal, "A Novel Framework for Arabic Dialect Chatbot Using Machine Learning," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1844051.
- [73] O. G. Alobaidi, K. A. Crockett, D. O'Shea, and T. Jarad, "Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education," vol. 2, Jul. 2013.
- [74] T. Alshareef and M. A. Siddiqui, "A seq2seq neural network based conversational agent for gulf arabic dialect," in *Proc. 21st International, ACIT*, Giza, Egypt: Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ACIT50332.2020.9300059.
- [75] B. A. Shawar and E. Atwell, "Accessing an Information System by Chatting," in *Proc. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2004, pp. 407–412. doi: 10.1007/978-3-540-27779-8_39.
- [76] Z. Mundher, W. K. Khater, and L. M. Ganeem, "Adopting Text Similarity Methods and Cloud Computing to Build a College Chatbot Model," *Journal of Education and Science*, vol. 30, no. 1, pp. 117–125, Mar. 2021, doi: 10.33899/EDUSJ.2020.127244.1079.
- [77] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamadi, "AI Chatbot for Tourist Recommendations: A Case Study in the City of Jeddah, Saudi Arabia," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 14, no. 19, pp. 18–30, Nov. 2020, doi: 10.3991/IJIM.V14I19.17201.
- [78] M. Hijjawi, Z. Bandar, and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 2, 2016, doi: 10.14569/IJACSA.2016.070247.
- [79] A. Shawar, "An Arabic chatbot giving answers from the Qur'an," in *TALN04: XI ConProc. fference sur le Traitement Automatique des Langues Naturelles*, Fez, Morocco, Jan. 2004.
- [80] A. Moubaidin, O. Shalbak, B. Hammo, and N. Obeid, "Arabic Dialogue System for Hotel Reservation based on Natural Language Processing Techniques," *Computación y Sistemas*, vol. 19, no. 1, pp. 119–134, Jan. 2015, doi: 10.13053/CYS-19-1-1962.
- [81] 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 *Proc. ICIT 2021*, Amman, Jordan: Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 592–597. doi: 10.1109/ICIT52682.2021.9491699.
- [82] Z. Noori, Z. Bandarl, and K. A. Crockett, "Arabic goal-oriented conversational agent based on pattern matching and knowledge trees," in *Proc. the World Congress on Engineering*, London, U.K, Jul. 2014.
- [83] A. Montejo-Ráez, S. María Jiménez-Zafra, A. Fuad, and M. Al-Yahya, "AraConv: Developing an Arabic Task-Oriented Dialogue System Using Multi-Lingual Transformer Model mT5," *Applied Sciences* 2022, Vol. 12, Page 1881, vol. 12, no. 4, p. 1881, Feb. 2022, doi: 10.3390/APP12041881.
- [84] G. Dudek, R. P. Lopes, and M. Alruily, "ArRASA: Channel Optimization for Deep Learning-Based Arabic NLU Chatbot Framework," *Electronics* 2022, Vol. 11, Page 3745, vol. 11, no. 22, p. 3745, Nov. 2022, doi: 10.3390/ELECTRONICS11223745.
- [85] A. H. Al-Ajmi and N. Al-Twairsh, "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.
- [86] D. A. Ali and N. Habash, "Botta: An Arabic Dialect Chatbot," in *Proc. COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, 2016, pp. 208–212.
- [87] G. Bilquise, S. Ibrahim, and K. Shaalan, "Bilingual AI-Driven Chatbot for Academic Advising," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 50–57, Oct. 2022, doi: 10.14569/IJACSA.2022.0130808.
- [88] 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/METRICS.
- [89] A. Fuad and M. Al-Yahya, "Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5," *Mathematics* 2022, vol. 10, no. 5, p. 746, Feb. 2022, doi: 10.3390/MATH10050746.
- [90] T. S. P. Shadi Abras and Lona Alani, "Construction of Arabic Interactive Tool Between Humans and Intelligent Agents," in *Proc. he 16th International Conference on Human-Computer Interaction*, Creta Maris, Heraklion, Crete, Greece: Springer, 2014.

- [91] N. O. Alshammari and F. D. Alharbi, "Combining a Novel Scoring Approach with Arabic Stemming Techniques for Arabic Chatbots Conversation Engine," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 4, Jan. 2022, doi: 10.1145/3511215.
- [92] M. Makatchev et al., "Dialogue patterns of an Arabic robot receptionist," in *Proc. 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Osaka, Japan: Institute of Electrical and Electronics Engineers (IEEE), Apr. 2010, pp. 167–168. doi: 10.1109/HRI.2010.5453213.
- [93] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven Arabic Conversational Chatbot," in *Proc. The Fifth Arabic Natural Language Processing Workshop*, Barcelona, Spain: Association for Computational Linguistics, 2020, pp. 58–68.
- [94] T. Naous, W. Antoun, R. A. Mahmoud, and H. Hajj, "Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data," pp. 164–172, Mar. 2021, [Online]. Available: Arxiv : arXiv:2103.04353
- [95] 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 *Proc. CIVEMSA*, Annecy, France: Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 24–29. doi: 10.1109/CIVEMSA.2017.7995296.
- [96] A. Boulesnane, Y. Saidi, O. Kamel, M. M. Bouhamed, and R. Mennour, "DZchatbot: A Medical Assistant Chatbot in the Algerian Arabic Dialect using Seq2Seq Model," in *Proc. International Conference on Pattern Analysis and Intelligent Systems*, Oum El Bouaghi, Algeria: Institute of Electrical and Electronics Engineers Inc., Nov. 2022. doi: 10.1109/PAIS56586.2022.9946867.
- [97] L. Riek and Z. Ahmed, "Ibn Sina Steps Out: Exploring Arabic Attitudes Toward Humanoid Robots," in *Proc. Second International Symposium on New Frontiers in Human-Robot Interaction*, De Montfort University, Leicester, UK, Apr. 2010.
- [98] M. A. M. Safee et al., "Hybrid Search Approach for Retrieving Medical and Health Science Knowledge from Quran," *International Journal of Engineering & Technology*, vol. 7, no. 4.15, pp. 69–74, Oct. 2018, doi: 10.14419/IJET.V7I4.15.21374.
- [99] A. Moubaidin, O. Shalbak, B. Hammo, and N. Obeid, "Arabic Dialogue System for Hotel Reservation based on Natural Language Processing Techniques," *Computación y Sistemas*, vol. 19, no. 1, pp. 119–134, Jan. 2015, doi: 10.13053/CYS-19-1-1962.
- [100] M. Hijjawi, Z. Bandar, and K. Crockett, "User’s utterance classification using machine learning for Arabic Conversational Agents," in *Proc. 5th International Conference on CSIT*, Amman, Jordan, Sep. 2013, pp. 223–232. doi: 10.1109/CSIT.2013.6588784.
- [101] T. Wael, A. Hesham, M. Youssef, O. Adel, H. Hesham, and M. S. Darweesh, "Intelligent Arabic-Based Healthcare Assistant," in *Proc. NILES 2021 conference*, Giza, Egypt: Institute of Electrical and Electronics Engineers Inc., Nov. 2021, pp. 216–221. doi: 10.1109/NILES53778.2021.9600526.
- [102] W. El Hefny, Y. Mansy, M. Abdallah, and S. Abdennadher, "Jooka: A Bilingual Chatbot for University Admission," *Advances in Intelligent Systems and Computing*, vol. 1367 AISC, pp. 671–681, 2021, doi: 10.1007/978-3-030-72660-7_64.
- [103] A. M. Bashir, A. Hassan, B. Rosman, D. Duma, and M. Ahmed, "Implementation of A Neural Natural Language Understanding Component for Arabic Dialogue Systems," in *Proc. The 4th International Conference on Arabic Computational Linguistics*, Dubai, United Arab Emirates: Elsevier, Jan. 2018, pp. 222–229. doi: 10.1016/J.PROCS.2018.10.479.
- [104] A. Yahya, "LABEEB: Intelligent Conversational Agent Approach to Enhance Course Teaching and Allied Learning Outcomes attainment," *Journal of Applied Computer Science & Mathematics*, vol. 13, no. 1, pp. 9–12, Apr. 2019, doi: 10.4316/JACSM.201901001.
- [105] N. Mavridis, A. Aldhaheri, L. Aldhaheri, M. Khanii, and N. Aldarmaki, "Transforming IbnSina into an advanced multilingual interactive android robot," in *Proc. 2011 IEEE GCC Conference and Exhibition*, Dubai, United Arab Emirates, Apr. 2011, pp. 120–123. doi: 10.1109/IEEEGCC.2011.5752467.
- [106] B. A. Shawar and E. S. Atwell, "Using corpora in machine-learning chatbot systems," *International Journal of Corpus Linguistics*, vol. 10, no. 4, pp. 489–516, Nov. 2005, doi: 10.1075/IJCL.10.4.06SHA/CITE/REFWORKS.
- [107] M. Aljanabi, M. Ghazi, A. H. Ali, S. A. Abed, and C. Gpt, "ChatGpt: Open Possibilities," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 1, pp. 62–64, Jan. 2023, doi: 10.52866/20IJCSM.2023.01.01.0018.
- [108] M. Ghaleb et al., "Mining the Chatbot Brain to Improve COVID-19 Bot Response Accuracy," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2619–2638, Sep. 2021, doi: 10.32604/CMC.2022.020358.
- [109] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic Dialect Chatbot," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 452–459, 2020, doi: 10.14569/IJACSA.2020.0110357.

- [110] B. Hammo, H. Abu-Salem, S. Lytinen, and M. Evens, "QARAB: A Question Answering System to Support the Arabic Language," SEMITIC@ACL, pp. 1–11, 2002, doi: 10.3115/1118637.1118644.
- [111] A. Fadhil and A. AbuRa'Ed, "OloBot - Towards A Text-Based Arabic Health Conversational Agent: Evaluation and Results," in Proc. RANLP, Varna, Bulgaria: Incoma Ltd, Sep. 2019, pp. 295–303. doi: 10.26615/978-954-452-056-4_034.
- [112] S. Alhumoud et al., "Rahhal: A Tourist Arabic Chatbot," in Proc.- 2nd International Conference of Smart Systems and Emerging Technologies, SMARTTECH , Riyadh, Saudi Arabia: Institute of Electrical and Electronics Engineers Inc., Aug. 2022, pp. 66–73. doi: 10.1109/SMARTTECH54121.2022.00028.
- [113] M. Daoud, "Novel approach towards Arabic question similarity detection," in Proc. ICTCS, Amman, Jordan: Institute of Electrical and Electronics Engineers Inc., Oct. 2019. doi: 10.1109/ICTCS.2019.8923102.
- [114] S. Mahamat Yassin and M. Zubair Khan, "seerahbot-an-arabic-chatbot-about-prophets-biography," International Journal of Innovative Research in Computer Science & Technology, vol. 9, no. 2, pp. 89–97, 2021, doi: 10.21276/ijircst.2021.9.2.13.
- [115] H. Elgibreen, S. Almazayad, S. Bin Shuail, M. Al Qahtani, and L. Alhwiseen, "Robot Framework for Anti-Bullying in Saudi Schools," in Proc. - 4th IEEE, IRC , Taichung, Taiwan: Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 166–171. doi: 10.1109/IRC.2020.00033.
- [116] I. Zribi and L. H. Belguith, "TOWARD DEVELOPING AN INTELLIGENT PERSONAL ASSISTANT FOR TUNISIAN ARABIC," Jordanian Journal of Computers and Information Technology, vol. 8, no. 4, pp. 318–318, Sep. 2022, doi: 10.5455/JJCIT.71-1652434864.
- [117] M. Daoud, "Topical and Non-Topical Approaches to Measure Similarity between Arabic Questions," Big Data and Cognitive Computing 2022, Vol. 6, Page 87, vol. 6, no. 3, p. 87, Aug. 2022, doi: 10.3390/BDCC6030087.
- [118] A. BOUZIANE, D. BOUCHIHA, N. DOUMI, and M. MALKI, "Toward an Arabic Question Answering System over Linked Data," Jordanian Journal of Computers and Information Technology, vol. 4, no. 2, pp. 102–102, May 2018, doi: 10.5455/JJCIT.71-1514749838.
- [119] 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," Computer Applications in Engineering Education, vol. 29, no. 6, pp. 1718–1742, Nov. 2021, doi: 10.1002/CAE.22419.
- [120] 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 Proc. 9th International Conference on ICIET , Okayama, Japan: Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 139–144. doi: 10.1109/ICIET51873.2021.9419656.
- [121] W. Alshammari and S. Alhumoud, "TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT with BiLSTM," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3198955.
- [122] M. Abdelhay, A. Mohammed, and H. A. Hefny, "Deep learning for Arabic healthcare: MedicalBot," Soc Netw Anal Min, vol. 13, no. 1, pp. 1–17, Dec. 2023, doi: 10.1007/S13278-023-01077-W/FIGURES/10.
- [123] L. Sherkawi, N. Ghneim, and O. Al Dakkak, "Arabic speech act recognition techniques," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 17, no. 3, Feb. 2018, doi: 10.1145/3170576.
- [124] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamadi, "AI Chatbot for Tourism Recommendations A Case Study in the City of Jeddah, Saudi Arabia," International Journal of Interactive Mobile Technologies, vol. 14, no. 19, pp. 18–30, 2020, doi: 10.3991/IJIM.V14I19.17201.
- [125] 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 Comput Sci, vol. 142, pp. 222–229, Jan. 2018, doi: 10.1016/J.PROCS.2018.10.479.
- [126] F. Bendjamaa and T. Nora, "A Dialogue-System Using a Qur'anic Ontology," in Proc. 2nd International Conference on EDiS , Oran, Algeria: Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 167–171. doi: 10.1109/EDIS49545.2020.9296437.
- [127] H. Abdelnasser et al., "Al-Bayan: An Arabic Question Answering System for the Holy Quran," in Proc. of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar: Association for Computational Linguistics (ACL), Oct. 2014, pp. 57–64. doi: 10.3115/V1/W14-3607.
- [128] M. AL-HARBI, "AQuASys : A Question-Answering System For Arabic," Computer Science, Linguistics, 2013.
- [129] Y. Benajiba, P. Rosso, and A. Lyhyaoui, "Implementation of the ArabiQA Question Answering System's components," in Proc. Workshop on Arabic Natural Language Processing, 2nd ICTIS, Fez, Morocco, Apr. 2007.
- [130] O. Trigui, L. Hadrich Belguith and P. Rosso, "Arabic Cooperative Answer Generation via Wikipedia Article Infoboxes," Research in Computing Science, vol. 132, no. 1, pp. 129–153, Dec. 2017, doi: 10.13053/RCS-132-1-11.

- [131] M. J. Page et al., “PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews,” *BMJ*, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N160.
- [132] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouni, A. Abd-alrazaq, and M. Househ, “Arabic chatbot technologies: A scoping review,” *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100057, Jan. 2022, doi: 10.1016/J.CMPBUP.2022.100057.
- [133] N. A. Alhassan, A. Saad Albarrak, S. Bhatia, and P. Agarwal, “A Novel Framework for Arabic Dialect Chatbot Using Machine Learning,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1844051.
- [134] Z. Lin et al., “BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling,” arXiv:2106.02787, Jun. 2021.
- [135] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, “Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset,” in *Proc. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Association for Computational Linguistics (ACL)*, Nov. 2018, pp. 5370–5381. doi: 10.18653/v1/p19-1534.
- [136] H. Lu, S. Bao, H. He, F. Wang, H. Wu, and H. Wang, “Towards Boosting the Open-Domain Chatbot with Human Feedback,” *Proc. of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 4060–4078, Aug. 2022, doi: 10.18653/v1/2023.acl-long.224.
- [137] Sourav Dutta, “Evaluating a neural multi-turn chatbot using BLEU score,” Saarland University Department of Computational Linguistics. Mar. 31, 2019. doi: 10.13140/RG.2.2.13374.84808.
- [138] M. Pikuliak, M. Šimko, and M. Bielíková, “Cross-lingual learning for text processing: A survey,” *Expert Syst Appl*, vol. 165, p. 113765, Mar. 2021, doi: 10.1016/J.ESWA.2020.113765.
- [139] Z. Liu et al., “Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables,” in *Proc. EMNLP-IJCNLP, Hong Kong, China: Association for Computational Linguistics*, Nov. 2019, pp. 1297–1303. doi: 10.18653/V1/D19-1129.
- [140] D. Jurafsky and J. H. Martin, “Book Review: *Speech and Language Processing (second edition)* by Daniel Jurafsky and James H. Martin,” *Computational Linguistics*, vol. 35, no. 3, pp. 463–466, Sep. 2009, doi: 10.1162/COLI.B09-001.
- [141] T. Brychcín and M. Konopík, “Semantic spaces for improving language modeling,” *Comput Speech Lang*, vol. 28, no. 1, pp. 192–209, Jan. 2014, doi: 10.1016/J.CSL.2013.05.001.
- [142] D. Lin, “An Information-Theoretic Definition of Similarity,” *Icml*, vol. 98, no. 1998, 1998.
- [143] J. Wang and Y. Dong, “Measurement of Text Similarity: A Survey,” *Information* 2020, Vol. 11, Page 421, vol. 11, no. 9, p. 421, Aug. 2020, doi: 10.3390/INFO11090421.
- [144] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, “A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting,” *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, Feb. 2023, doi: 10.1007/S40745-021-00344-X/METRICS.
- [145] A. R. Yauri, R. A. Kadir, A. Azman, and M. A. A. Murad, “Quranic verse extraction base on concepts using OWL-DL ontology,” *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 23, pp. 4492–4498, Dec. 2013, doi: 10.19026/RJASET.6.3457.
- [146] A. Pasha et al., “MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic,” in *Proc. the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland: European Language Resources Association (ELRA)*, 2014, pp. 1094–1101.
- [147] D. H. Abd, W. Khan, K. A. Thamer, and A. J. Hussain, “Arabic Light Stemmer Based on ISRI Stemmer,” in *Proc. Huang, DS., Jo, KH., Li, J., Gribova, V., Premaratne, P. (eds) Intelligent Computing Theories and Application. ICIC 2021. Lecture Notes in Computer Science()*, Springer, Cham, 2021, pp. 32–45. doi: 10.1007/978-3-030-84532-2_4.
- [148] M. N. Al-Kabi, “Towards improving Khoja rule-based Arabic stemmer,” in *Proc. 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT*, Amman, Jordan, 2013. doi: 10.1109/AEECT.2013.6716437.
- [149] A. Abran, A. Khelifi, W. Suryan, and A. Seffah, “Usability meanings and interpretations in ISO standards,” *Software Quality Journal*, vol. 11, no. 4, pp. 325–338, 2003, doi: 10.1023/A:1025869312943/METRICS.
- [150] J. R. Lewis and J. Sauro, “The Factor Structure of the System Usability Scale,” in *Proc. Kurosu, M. (eds) Human Centered Design. HCD 2009. Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer, 2009, pp. 94–103. doi: 10.1007/978-3-642-02806-9_12.
- [151] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. McTear, “Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?,” in *Proc. ECCE, Association for Computing Machinery, Inc*, Sep. 2019, pp. 207–214. doi: 10.1145/3335082.3335094.
- [152] C. Myers-Scotton, “Code-Switching,” in *The Handbook of Sociolinguistics*, John Wiley & Sons, Ltd, 2017, pp. 217–237. doi: 10.1002/9781405166256.CH13.
- [153] M. Aljanabi, (2024). Assessing the Arabic Parsing Capabilities of ChatGPT and Cloude: An Expert-Based Comparative Study. *Mesopotamian Journal of Arabic Language Studies*, (2024), 16–23. <https://doi.org/10.58496/MJALS/2024/002>

[154] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in Proc. 7th International Conference on Information and Education Technology, Association for Computing Machinery, Mar. 2019, pp. 111–119. doi: 10.1145/3323771.3323824.

[155] S. Elkateb et al., "Building a WordNet for Arabic," in Proc. the Fifth International Conference on Language Resources and Evaluation , Genoa, Italy: European Language Resources Association, May 2006.

APPENDIX:

a			b			
Arabic Word	Form	Type	Alphabet	Shape in words		
				Ending	Middle	Beginning
سَجَّلَ	Registered	Verb	ا	ل	ل	ا
سَجَّلَ	Scoured	Verb	ب	ب	ب	ب
سَجَّلَ	Added	Verb	ت	ت	ت	ت
سَجَّلَ	Enrolled	Verb	ث	ث	ث	ث
سَجَّلَ	Submitted	Verb	ج	ج	ج	ج
سَجَّلَ	Recorded	verb	ح	ح	ح	ح
سَجَّلَ	Wrote	verb	خ	خ	خ	خ
سَجَّلَ	Joined	verb	د	د	د	د
سَجَّلَ	Post	verb	ذ	ذ	ذ	ذ
سَجَّلَ	Throw	verb	ر	ر	ر	ر
سَجَّلَ	Read constantly	verb	ز	ز	ز	ز
سَجَّلَ	Poured	verb	س	س	س	س
سَجَّلَ	File	noun	ش	ش	ش	ش
سَجَّلَ	Record	noun	ص	ص	ص	ص
سَجَّلَ	Pouring	noun	ض	ض	ض	ض
سَجَّلَ	Being registered	Passive verb	ط	ط	ط	ط
سَجَّلَ	Being scored	Passive verb	ظ	ظ	ظ	ظ
سَجَّلَ	Being added	Passive verb	ع	ع	ع	ع
سَجَّلَ	Being enrolled	Passive verb	غ	غ	غ	غ
سَجَّلَ	Being submitted	Passive verb	ف	ف	ف	ف
سَجَّلَ	Being recorded	Passive verb	ق	ق	ق	ق
سَجَّلَ	Being posted	Passive verb	ك	ك	ك	ك
			ل	ل	ل	ل
			م	م	م	م
			ن	ن	ن	ن
			ه	ه	ه	ه
			و	و	و	و
			ي	ي	ي	ي

FIGURE S1. (a) Different derivations can the word "سَجَّلَ"-register/ generate (taken from [14]).
 (b) Arabic letters change shape based on their position.

Word in English	Word in Arabic dialects										
	Arabic (Standard)	Libya	Morocco	Algeria	Egypt	Jordan	Levantine	Iraqi	Hijazi	Jizani	Tunisia
now	الآن	توة	دابا	ضورك	دا الوقت	هالا،هسا	هالا،هسا	هسة	دحين	ذخينة	توا
speak	تكلم	تكلم	تكلم، هذر	أهذر	اتكلم	إحكى	حكى	أحجي	تهزج	اتكلم	فكرونه
Turtle	سلحفاة	فكرونه	فكرون	فكروان	سلحفاة	كر كعة	سلحفاة	رگه/سلحفاة	سلخفة	سلخفة	فكرونه
rain	مطر	مطر	شنا	نو	مطر	مطر	مطر، شتي	مطر	مطره	مطر	شنا/مطر
lazy	كسول	تمبال،كسلان	كسول،معجاز،	فيلان	كسلان	كسلان	عجزان، كسلان	كسلان	كسلان	كسلان	بخيل
want	يريد	بيي	باغي/بيغي	بيغي،	عاوز	بدو	بدو	يريد	بيغي	بيغي	يحب
A lot	كثير	هلية	بالزاف	بزاف	قطعة	بسة	بسينة	بزونه	بسة	بسة	برشة
Cat	قطعة	قطوسة	قطه/مشه	قطعة	قطعة	بسة	بسينة	بزونه	بسة	بسة	قطوسة
How are you ?	كيف حالك؟	شنو حالك؟	كي داير ؟ لا باس؟	كيف حالك؟	از اترك؟	كيفك؟، شلونك؟	كيفك؟	شلونك؟	كيفك؟	ايش حالك	شنوا
What?	ماذا	شنو	اشنو/شنو	واش؟	ايه	ايش	شو	شنو	ايش	ماهو	شنو/اوشيا
Why?	لماذا؟	وعلاش/عليش	غلاش/لاش	علاش/علاه	ليه	لئيش	لئيش	لئيش/لئشو/لئوش	ليه/لئيش	لئيش	غلاش

FIGURE S2. Comparison of Words in Classical Arabic and Various Arabic Dialects.