Developing chatbots in the field of healthcare: A systematic review

Haniel Gomes Cavalcante, Thaís de Almeida Barros, Ricardo Antônio Rebouças Celestino, Daniel Gleison Moreira Lira, Francisco Vinicius Nascimento da Silva, Pedro Henrique Araújo de Brito, Mariela I. Cortés [1]

[1]Programa de Pós Graduação em Ciencia da Computação – Universidade Estadual do Ceará
Avenida. Dr. Silas Munguba, 1700 - Itaperi, Fortaleza - CE, 60714-903
{haniel.cavalcante, thais.almeida, ricardo.celestino, daniel.gleison, francisco.vinicius, pedrinho.brito}@aluno.uece.br, mariela.cortes@uece.br

Abstract. *Considering the context of Computer Science, chatbots are computer programs that use Artificial Intelligence techniques to simulate human behavior in dialogues. The use of chatbots applied to the health area has been growing, especially in scenarios for dealing with pandemics, such as COVID-19, as they help to avoid the burden of face-to-face care. Thus, this article proposes a systematic review of the work carried out in this line of research. After the review, it was found which technologies, strategies and frameworks are most used in recent times, as well as which specific areas of health are having more focus on the use of chatbots.*

Resumo. *Considerando o contexto da Ciência da Computação, os chatbots são programas de computador, que utilizam técnicas de Inteligência Artificial para simular o comportamento humano em diálogos. O uso de chatbots aplicados à área de saúde vem crescendo, principalmente nos cenários de enfrentamento de pandemias, como o COVID-19, pois ajudam a evitar o ônus do atendimento presencial. Assim, este artigo propõe uma revisão sistemática dos trabalhos realizados nesta linha de pesquisa. Após a revisão, constatou-se quais tecnologias, estratégias e enquadramentos são mais utilizados nos últimos tempos, bem como quais áreas específicas da saúde estão tendo mais foco no uso de chatbots.*

1. Introduction

Artificial Intelligence (AI) is known as a science dedicated to the study of systems that in any observer's perspective, act with intelligence [Coppin 2015]. In this context, the chatbot or smart agent are different computing system, where the symbolic and connective approaches can act in a collaborative way, aiming for problem solving. [Bernardini et al. 2018]. The basic principle employed in a chatbot consists of an environment that receives questions in human natural language, associates these questions to a knowledge base and finally, emits an answer [Fryer and Carpenter 2006].

Chatbots can be used in many application domains, such as entertainment, business, education and health. Examples of chatbots are the projects ELIZA ([Weizenbaum 1983]), MGONZ ([Humphrys 2009]), PARRY ([Huang et al. 2007]) and ALICE ([Lima 2014]). Regarding chatbot development in the healthcare domain, the

works aim to help the interaction between patients and healthcare professionals in many specialities: psychiatry, psychology, pediatrics, cardiology, and many more.

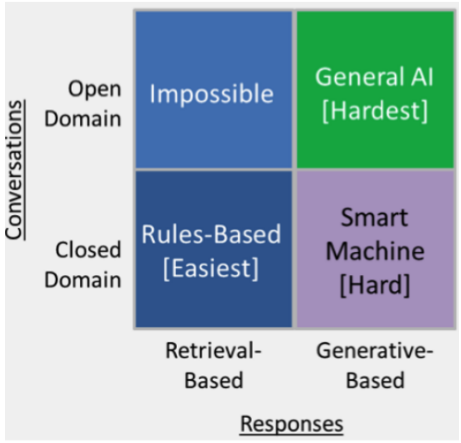
Chatbots used in the healthcare domain can make symptoms mapping and diagnostic predictions, as well as advising instructions for the patient based on machine learning models. By doing so, it is possible to do the care screening remotely and privately, avoiding face-to-face service overload and exposing the patients to unnecessary contact with other patients, especially when dealing with global pandemics, such as COVID-19.

In this context, it is necessary to investigate the use of chatbots in healthcare to determine and identify which techniques, strategies and frameworks are the most used, especially when this use for healthcare is expected to grow in the scientific field, also with empirical evaluations, making it possible to produce systematic reviews of the literature. In this work, research methods, collected results from previous studies and discussions about the development of chatbots in healthcare will be addressed.

2. Theoretical Foundation

Usually, chatbots can be assigned to categories such as in Figure 1 [Kamphaug et al. 2017]. Those that are retrieval-based work with preset answers and can use languages such as AIML (Artificial Intelligence Markup Language) [Wallace 2009] to manually define interaction patterns previously implemented. Generative-based models, in the other hand, have the ability to generate new answers in real time.

Figure 1. Description of possible chatbot types.



Besides that, another classification that can be utilized for these systems is about their universe of knowledge in which it proposes to answer questions [Mollá and Vicedo 2007]. Open domain systems aim to answer general questions, based on open data bases from the web. Restricted domain systems address answers to questions of a certain sector, such as the biomedical domain. The closed domain systems are confined to answer questions about a closed collection of documents, which is usually small. Examples of the latter could be a system to answer questions about a Consumer Protection Code or a state's customs laws.

3. Methodology

For this paper, the purpose of this systematic review is to identify the activities, techniques, methods and tools (*frameworks* and platforms) considered in the devel-

opment of *chatbots* in healthcare. The process used was based on the works of [Sánchez-Gómez et al. 2020] and [Kitchenham and Brereton 2013]. Therefore, the use of this technique has occurred in three phases, as described below.

3.1. Planning

The Planning phase is the once in which the objective of the research must be defined, the way in which the systematic review will be performed, and which criteria will be taken into consideration for the inclusion and exclusion of papers.

3.1.1. Research Objectives

Based on the scenario, as well as in the context that defines the main problem to be addressed in this investigation, the main objective of this research is an analysis of scientific publications that present items of interest related to chatbots in healthcare, aiming to determine which technologies, strategies and frameworks are the most used in the area in order to find out the most adequate and safe ones for the healthcare context.

For the systematic review, searches for primary studies published in journal articles and conference proceedings were done, using the electronic search in digital repositories in the areas of healthcare and technology from 2015 to 2020. This interval was chosen so that only recent articles were analyzed, which means there would be higher probability of representing the state-of-the-art.

3.1.2. Research Questions (RQs)

- RQ1. What are the tools (*frameworks*/platforms) addressed in the research to provide support?
- RQ2. What is the type of response from *chatbot* according to those described in [Kamphaug et al. 2017]?
  - A. Retrieval-based;
  - B. Generative-based;
  - C. Other type.
- RQ3. Does the paper propose the use of any technology for the intelligence of the *chatbot* among the most used ones? (single or multiple selection)
  - A. The article proposes the use of *Machine Learning*;
  - B. The paper proposes the use of *Pattern Matching*;
  - C. The paper proposes the use of Ontologies;
  - D. The paper does not describe the use of any technology for the development of *chatbots*;
  - E. The paper proposes another technology not mentioned in the options above.
- RQ4. What is the adopted/proposed conversational domain according to [Kamphaug et al. 2017] taxonomy?
  - A. The paper proposes the use of Open Domain;
  - B. The paper proposes the use of Restricted Domain;
  - C. The paper proposes the use of Closed Domain;
  - D. The paper does not describe the use of some kind of domain for the development of *chatbots*.

- RQ5. What kind of empirical evaluation was conducted to assess the quality of the *chatbot*? Which aspect of quality was evaluated?
- RQ6. Which subarea of health care is addressed in the paper?
  A. Psychiatry/Psychology;
  B. Pediatrics;
  C. Nursing;
  D. Physiotherapy;
  E. Another.
- RQ7. What are the challenges and limitations identified in the development of *chatbots*?

3.1.3. Search Strategy for Primary Study Selection

For the analysis and selection of primary studies, ACM Digital Library, IEEE Xplore Digital Library, ScienceDirect, Elsevier's Scopus and Springer Link were defined as sources of works, in which the search strings in the title, abstract and keywords were executed in the period 2015-2020. In this sense, the followind searchstring hasbeen defined:

("chatbots" OR "conversational agent" OR "conversational bot" OR "conversational system" OR "conversational interface" OR "chat bot" OR "chatterbot" OR "chat-bot" OR "smartbot" OR "smart bot" OR "smart-bot" OR"virtual coach" OR "virtual agent" OR "embodied agent" OR "relational agent" OR"avatar" OR "virtual character" OR "animated character" OR "virtual human" OR "health") AND ("health") NOT ("systematic review") Title, abstract, keywords. Year: 2015-2020

3.1.4. Study Selection Procedures

The selective process of this review began with the execution of the search *strings*. In the second phase, screening took place, where exclusion criteria were applied. At this point, the primary studies were distributed to each researcher according to Table 1. The last phase included a meeting to provide a forum for discussion and consensus among researchers when there were questions for the evaluation of a paper. The purpose of this meeting was to reduce each researcher's bias in order to resolve any doubts in the application of the inclusion/exclusion criteria. In these cases, a complete reading of the doubtful papers was necessary. After this reading, all researchers decided to include or exclude the primary study (PE). The decision was joint to avoid subjectivity. The following inclusion and exclusion criteria were defined to select the primary studies:

Table 1. Distribution per researcher

Phase	Description	Participating researchers
F1	Execution of the search strategies considering the search strings	5 researchers (one for each base)
F2	Screening: exclusion of primary studies dealing with other issues	2 researchers
F3	First Consensus Meeting	All 7 researchers

a) Inclusion Criteria IC1 - Papers applying *chatbots* to healthcare; IC2 - Papers published from 2015 to 2020; IC3 - Papers published in conferences or *journals*;

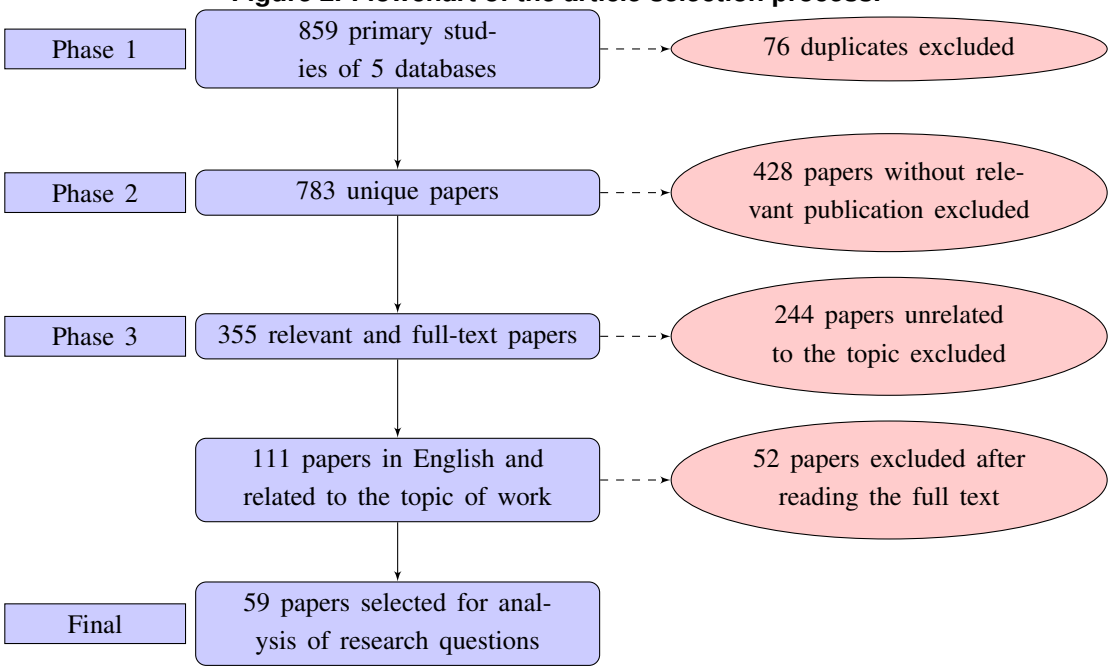
b) Exclusion Criteria EC1 - Papers in languages other than English; EC2 - Papers that have not been published in reputable journals (i.e. journals indexed in the *Journal Citation Reports - JCR*) or prestigious conferences (i.e. conference level A*, A, B and C categorized in the CORE *Conference Ranking*). EC3 - Papers without full text available; EC4 - Papers not related to the development of *chatbots* in healthcare; EC5 - Thesis, books, discussions, opinion papers related to *chatbots*; EC6 - Systematic Reviews.

3.2. Execution

In phase 1, the automatic search was performed in each digital library. Thus, all documents returned by search queries were included in this phase. In phase 2, articles were considered only in English and with full texts. Papers that not related to the subject were excluded. This exclusion phase included eliminating duplicate documents, as well as reading the title and abstract. In case of doubt about any article, this paper was included preliminarily. The final decision was considered and evaluated in the next phase. In the third and final phase, the researchers reviewed all articles where there was any uncertainty, performing a complete reading of the article. After the end of this phase, the final list of primary studies that would be analyzed according to the previously defined research questions was defined.

In the first phase, 859 papers related to the theme of this research were retrieved. Applying the exclusion criteria defined for phase 2, it was possible to identify 355 primary studies. After the consensus meeting of phase 3, 111 relevant articles concerning the objectives of this systematic review remained. After a complete reading of the texts, 59 articles were defined for analysis of the research questions. A flowchart with all the phases can be seen in Figure 2.

Figure 2. Flowchart of the article selection process.



4. Results Analysis

The final stage of the systematic review was the results analysis. In this stage, each of the 59 articles was analyzed to answer the research questions. Table 2 shows the number of articles per selection phase and per research base. The list of selected articles, include authors, year of publication and database, is available at Github [1].

Table 2. Number of articles per selection phase

Database	Phase 1	Phase 2	Phase 3	Final
ACM	65	61	19	10
IEEE	249	178	66	7
ScienceDirect	160	159	110	6
Scopus	255	255	146	35
SpringerLink	130	130	14	1
Total	**859**	**783**	**355**	**59**

The answers to RQ1. "What are the tools (*frameworks*/platforms) addressed in the research to provide support?" were as diverse as possible. Here, we can list tools, such as: *Google Dialog (5%), Facebook Messenger (5%), AWS Polly (2%), FAtiMA (3.5%), Skills* from *Alexa* (3.5%), among others. In 13.5% of the files reviewed, no tools were reported.

In RQ2. "What is the *chatbot* response type?", we can see that the *chatbot* response types are well divided into: retrieval-based (49.2%) and generative-based (39%). We have few types of responses that differ from these two ways (1.7%), and in approximately 10% of the cases, the authors do not inform the type of response.

In RQ3, "Does the article propose the use of any technology for the intelligence of the *chatbot* among the most used?", we have a percentage of more than 1/3 of the articles that do not reference any technology. Next, we have the use of *Machine Learning* (23.7%), the use of ontologies (13.6%), the use of *Pattern Matching* (10.2%), and the use of Natural Language (8.5%). Therefore, it is possible to verify the use of the most diverse technologies for the intelligence of the *chatbot*.

For RQ4. "What conversational domain is adopted/proposed?", we can observe that the conversational domains adopted are usually either closed domain (57.6%) or restricted domain (25.4%). In only a little over 3% of the files, an open domain is proposed, and in 13.6% of the articles, no domain was proposed or adopted.

To answer RQ5. "What kind of empirical evaluation was performed to evaluate the quality of the *chatbot*? Which aspect of quality was evaluated?", we observe that there are several forms of empirical evaluation regarding the quality of *chatbots*. The most common is the analysis based on the collected data, and may or may not use statistical methods, which occurs in approximately 35.5%. The other most common type of evaluation is based on the *feedback* provided by the participants with approximately 29%, and in some cases (13.5%) both types of empirical evaluation are used. Thus, the empirical evaluation is performed in approximately 78% of the articles, whether it is based on the analysis of the collected data, such as the application of statistical methods, K-FOLD cross-validation, BLEU score and BERT score, or based on the *feedback* provided

[1] https://github.com/danielgleison/chatbots-health

by the participants, where, in most cases, a questionnaire is applied to assess the level of satisfaction, empathy or comfort in the interaction with the *chatbot*. In addition to these evaluation methods, some articles use both types, achieving a broader evaluation of the tool. In these cases, data such as changes in patient anxiety levels, the degree of intimacy and bonding between the *chatbot* and the patient, and the user's desire to interact with the *chatbot* again are evaluated. In 22% of the articles, there was no or no report of whether there was any empirical evaluation.

In RQ6, "Which subarea of the healthcare domain is addressed in the article?" it was observed that the main subarea addressed in the articles is psychiatry/psychology, with approximately half of the occurrences. The subarea of clinical medicine comes in 2nd place with 13.6%, and clinical analysis along with health and quality of life are tied as the third most addressed with 6.8%. In addition, several other areas are addressed, such as obstetrics, pediatrics, endocrinology, and physical therapy.

To answer RQ7, "What are the challenges and limitations identified in the development of *chatbots*?", based on the complete reading of the papers, excerpts were extracted in which the authors identified possibilities for improvement, as well as limitations in the development of *chatbots*. We detected that, in approximately 27% of the analyzed papers, no challenges or limitations regarding the development of *chatbots* had been reported. In the others, it was verified that the biggest challenge reported would be the need to make the *chatbot* more friendly and attractive to the user, which was reported in almost 34% of the cases. In addition, in 70% of the cases, it was reported that it was necessary to increasingly improve the conversation between the *chatbot* and the user so that the communication would flow better. In addition, other challenges/limitations cited were: "Examine the effect of a mental health chatbot on mood in a postpartum population", "Acquire real-life data to improve the algorithm", and "Understand how the use of emotional language influences interaction".

5. Threats to Validity

In this section, we discuss possible threats to the validity of our study. We identified threats to the identification of primary studies, as we addressed possible limitations in the process of searching for articles that could lead to the absence of related literature and a major challenge of the work was the existence of articles that did not focus on agent engineering per se, but on its interface or only in the empirical evaluation, leaving, in some cases, research questions unanswered. To minimize these threats, leading digital libraries in computing were considered to reduce publication bias. Another threat concerns data extraction, related to possible problems in the data collection phase, such as the subjectivity of the researcher who performs this collection. To reduce this risk, the extraction of information was carried out by a researcher and reviewed by all, in cases of uncertainty.

6. Conclusion and Future Work

This work performed a systematic review of articles in the healthcare area, where some type of conversational agent is used. Through the proposed research questions, it was possible to determine that most used *chatbot* reponse type are retrieval-based and generative-based, that the most used technologies for the intelligence of *chatbot* are Machine Learning, Ontologies and Pattern Matching that are responsible for around 50%. We can also

note that for *chatbots*, closed domain and restricted domain are used in more than 80% of the analyzed papers and that psychiatry/psychology is the subarea of health that is most addressed in the context. This work can be used as a reference for developers looking to implement a conversational agent in the field of healthcare and who want to know the most commonly used technologies, as well as examine different options for strategies and approaches. It is also of interest to researchers, as we map what is being researched in the area. For future work, other systematic literature reviews can be conducted, this time with a greater focus on a particular sub-area of the field of healthcare or with the use of a particular technology or standard, since this work was quite comprehensive.

References

Bernardini, A. A., Sônego, A. A., and Pozzebon, E. (2018). Chatbots: An analysis of the state of art of literature. In *Workshop on Advanced Virtual Environments and Education*, volume 1, pages 1–6.

Coppin, B. (2015). *Inteligência artificial*. Grupo Gen-LTC.

Fryer, L. and Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10(3):8–14.

Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428.

Humphrys, M. (2009). How my program passed the turing test. In *Parsing the Turing Test*, pages 237–260. Springer.

Kamphaug, Å., Granmo, O.-C., Goodwin, M., and Zadorozhny, V. I. (2017). Towards open domain chatbots—a gru architecture for data driven conversations. In *International Conference on Internet Science*, pages 213–222. Springer.

Kitchenham, B. and Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and software technology*, 55(12):2049–2075.

Lima, L. A. (2014). Estudo de implementação de um robô de conversação em curso de língua estrangeira em ambiente virtual: um caso de estabilização do sistema adaptativo complexo.

Mollá, D. and Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics*, 33(1):41–61.

Sánchez-Gómez, N., Torres-Valderrama, J., García-García, J. A., Gutiérrez, J. J., and Escalona, M. J. (2020). Model-based software design and testing in blockchain smart contracts: A systematic literature review. *IEEE Access*, 8:164556–164569.

Wallace, R. S. (2009). The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Weizenbaum, J. (1983). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 26(1):23–28.