UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Mestrado/Doutorado Acadêmico

João Luis Zeni Montenegro

HOPE: A Conversational Agent Based-Model For Pregnancy Health Literacy

São Leopoldo, 2022

João Luis Zeni Montenegro

HOPE : A CONVERSATIONAL AGENT BASED-MODEL FOR PREGNANT HEALTH LITERACY

Thesis presented as a partial requirement to obtain the Doctor's degree by the Applied Computing Graduate Program of the Universidade do Vale do Rio dos Sinos — UNISINOS

Orientador:
Prof. Dr. Cristiano André da Costa

São Leopoldo
2022

M777h Montenegro, João Luis Zeni
HoPE : a conversational agent based-model for pregnant health literacy / by João Luis Zeni Montenegro. – 2022.
136 l. : ill. ; 30 cm.

Thesis (doctor's degree) — Universidade do Vale do Rio dos Sinos, Applied Computing Graduate Program, 2022.
Orientador: Prof. Dr. Cristiano André da Costa.

1. Agente conversacional. 2. Gravidez. 3. Alfabetização em saúde. 4. Deep Learning. 5. Ontologia.6. BERT. I. Título.

CDU 004

Catalogação na Fonte:
Bibliotecária Vanessa Borges Nunes - CRB 10/1556

Dedico esse trabalho aos meus pais Hércules e Luci,
ao meu irmão Lucas, e a minha esposa Marina
por todo suporte durante essa jornada, sem
o apoio de vocês eu não teria conseguido.

AGRADECIMENTOS

Inicialmente, agradeço ao meu orientador Prof. Dr. Cristiano André da Costa, por me dar a oportunidade e o prazer de trabalhar consigo e por todas as orientações que compuseram esse trabalho. Agradeço pelo respeito, compreensão e parceria durante este período.

Fiz muitos amigos na Unisinos nesse período. Agradeço aos meus colegas do laboratório SoftwareLab que acrescentaram muito para o meu trabalho, além de terem se tornado amigos pessoais que levarei para o resto da vida, em especial, Luis Claudio Gubert, Felipe Zeizer e Guilherme Goldschimidt que estiveram desde o início até final da apresentação da minha tese.

Também agradeço a todos funcionários da Unisinos que sempre me trataram muito bem na instituição, em especial a Luciana Aquino e a Bruna Konzen Severo do PPGCA que construí uma amizade fora do ambiente de trabalho

RESUMO

O período gestacional é caracterizado por ser um momento de muita expectativa e ao mesmo tempo crítico para as gestantes devido a um grande contingente de incertezas e dúvidas que atingem a gestante neste período. Os primeiros meses do início da gravidez e do início da maternidade envolvem os estágios de pré natal e pós natal e é conhecido como período dos mil dias do bebê. Durante essa fase, diversos tópicos são explorados pelas gestantes como mitigação de riscos, exercícios físicos, nutrição e outros assuntos que acabam por provocar incertezas e ansiedade neste grupo. Agentes conversacionais vêm sendo utilizados através dos anos como ferramentas de engajamento, suporte e informação em diferentes áreas. Na área da saúde este tipo de ferramenta já vem sendo pesquisada para atuação conjunta a atividades junto a pacientes e médicos. Para esta tese propomos o desenvolvimento, implementação e avaliação de agentes conversacionais baseados no modelo de arquitetura HoPE (Help in Obstetrician for PrEgant) que tem por objetivo atuar na promoção de educação para gestantes através de informações confiáveis. Avaliamos esse modelo por meio de estudos clínicos e experimentos envolvendo a recuperação de informações por uma nova arquitetura para agentes conversacionais. O número de estudos com profissionais de saúde e gestantes ainda são remotos e carecem de mais aprofundamento. As estratégias que elencamos para gerenciamento de diálogo e recuperação de informações são pouco exploradas em conjunto no contexto científico, sendo que não encontramos nenhuma proposta que promova modelos com conceitos similares. A arquitetura desenvolvida tem como principais características a capacidade de recuperação e desambiguação de informações utilizando ontologias e modelos Transformadores. Realizamos quatro avaliações principais que promoveram numerosos insights para estudos neste campo. Em um estudo de caráter quantitativo por meio de questionários semi-estruturados, gestantes e profissionais de saúde participaram de interações junto a agentes conversacionais treinados em dados nutricionais. Os resultados apontaram que ambos os grupos têm percepções positivas sobre a experiência com o agente conversacional e estatisticamente a hipótese nula foi aceita (P-value = 0.713). Em uma segunda avaliação com amostra formada por diferentes gestantes e médicos, verificou-se por meio de uma análise mista que as percepções desses grupos são complementares e positivas sobre o uso de agentes conversacionais na saúde treinados em dados gerais do conteúdo de mil dias na gravidez. A nova amostra de gestantes demonstrou novamente percepção positiva de forma geral sobre novos constructos avaliados (Média Geral = 4.0 Desvio médio = 1.1). Ainda, insights gerados por médicos através da análise qualitativa indicaram algumas melhorias como a inserção de conteúdos sobre COVID-19 e comportamento familiar, além de ajustes na abordagem e linguajar do agente conversacional. Buscamos também o desenvolvimento de um novo tipo de estudo, voltado a avaliação da arquitetura proposta e sua performance frente a tarefas de recuperação de informações. Inicialmente, avaliamos modelos Sentence-BERT pré-treinados em língua portuguesa ajustados a dados de protocolos da saúde que extraímos de protocolos oficiais do Governo do Brasileiro. O modelo BERTimbau treinado em estratégias de aumento de dados, obteve a maior correlação com embeddings gerados pelo corpus de dados da saúde (Spearman:95.55) e foi selecionado como modelo vencedor em nossos experimentos. Usando este modelo, realizamos o segundo estudo que avaliou a performance da arquitetura HoPE para agentes conversacionais. Neste estudo, três métricas principais foram avaliadas: eficácia na recuperação de informações, capacidade da arquitetura para identificar intenções compostas e velocidade de inferência da arquitetura. Para tarefa de recuperação de informações, a arquitetura HoPE obteve um F1-Score de (0.89) sob os dados de teste, um hit score de (90%) na identificação de intenções compostas/únicas sob um conjunto de 10 frases e um desempenho regular na velocidade de recuperação de informação (CPU=2.223, GPU=0.222). Estudos futuros irão

avaliar por meio de estudos clínicos a arquitetura híbrida HoPE para recuperação de informações, a validação para grupos de gestantes de diferentes estratos demográficos e aprofundaram o estudo em mecanismos para identificação de múltiplas intenções em diálogos.

Palavras-chave: Agente conversacional, Gravidez, Alfabetização em saúde, Deep Learning, Ontologia, BERT .

ABSTRACT

The gestational period is a moment of great expectation and critical for pregnant women due to many uncertainties and doubts that affect the pregnant woman. The first months of pregnancy and motherhood, known as the baby's thousand days period, include the prenatal and postnatal stages with pregnant women investigating many topics, including risk management, physical activity, nutrition, and other issues that cause uncertainty and anxiety. Conversational agents have been played a role over the years as engagement, support, and information tools in different areas in the health field for collaborative action with patients and doctors. We propose in this thesis the development, implementation, and evaluation of conversational agents based on the HoPE (Help in Obstetrician for PrEgant) architecture model, which aims to promote literacy in pregnant women through reliable information. We evaluated this model through clinical trials and experiments involving information retrieval. Studies involving clinical trials with health professionals and pregnant women are still remote and need further investigation. The strategies we have listed for managing dialogue and retrieving information are unprecedented in the scientific context, and we have not found any proposal that promotes models with similar concepts. The architecture developed has as main pillars the ability to recover and disambiguate information using ontologies and architectures based on Transformers as a center. We carried out five assessments that provided numerous insights for studies in this field. Initially, we applied a survey to get a general picture of the subject in question. In a quantitative study using semi-structured questionnaires, pregnant women and health professionals interacted with conversational agents trained in nutritional data. The results showed that both groups have positive perceptions about the experience with the conversational agent and statistically the null hypothesis was accepted (P-value = 0.713). A second evaluation with a sample formed by different pregnant women and doctors verifies through a mixed analysis that the perceptions of these groups are complementary and positive, regarding the use of conversational health agents trained in general data of the content of a thousand days in pregnancy. The new sample of pregnant women again showed a positive perception in general about the new constructs evaluated (Overall Mean = 4.0 Mean Deviation = 1.1). Also, insights generated by doctors through qualitative analysis indicated some improvements as the inclusion of COVID-19 content and family behavior, as well as adjustments in the approach and language of the conversational agent. We evaluated the pre-trained Sentence-BERT models in Portuguese, adjusted to health protocol data that we extracted from official protocols of the Brazilian Government. The BERTimbau model, trained in data augmentation strategies, obtained the highest correlation with embeddings generated by the health data corpus (Spearman:95.55) and was selected as the winning model in our experiments. Using this model, we performed the second study that evaluated the performance of the HoPE architecture for conversational agents. Three main metrics were evaluated in this study: information retrieval efficacy, architecture's ability to identify composite intents, and architecture inference speed. For the information retrieval task, the HoPE architecture obtained an F1-Score of (0.89) under the test data, a hit score of (90%) in the identification of composite/unique intents under a set of 10 sentences, and a performance regular in the information retrieval speed (CPU=2.223, GPU=0.222). Future studies will evaluate through clinical studies the hybrid HoPE architecture for information retrieval, validation for groups of pregnant women from different demographic strata, and deepen the study on mechanisms for identifying multiple intentions in dialogues.

Keywords: Conversational Agent, Pregnancy, Health Literacy, Deep Learning, Ontology, BERT

.

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

BERT Stanford Question Answering Dataset

CONSORT Consolidating Standards of Reporting studies

CPU Central Processing Unit

EHR Electronic Health Records

ELMO Embeddings from Language Model

GLUE General Language Understanding Evaluation

GPT-3 Stanford Question Answering Dataset

GPU Graphics Processing Unit

HoPE Help in Obstetrician for PrEgant

HPV the human papillomavirus

JSON JavaScript Object Notation

MRR Mean Reciprocal Rank

MSE Mean squared Error

NLI Natural Language Inference

NLP Natural Language Processing

NLU Natural Language Understanding

OWL Ontology Web Language

PICOC population, intervention, comparison, outcome, and context

POS Part of Speech Tagging

RDF Resource Description Framework

SBERT Sentence-BERT

SMS Short Message Service

SQuAD Stanford Question Answering Dataset

STS Sentence Text Similarity

TF-IDF Term Frequency-Inverse Document Frequency

UTAUT2 Unified Theory of Acceptance and Use of Technology 2

CONTENTS

1 INTRODUCTION

Many technologies are driving considerable changes in the health field, involving physicians, patients, and other health professionals. Among these changes, interactions between humans and robots are gaining importance in assorted areas, such as speech and object recognition and natural language processing. As a result, these technologies have achieved improved user-robot interactions (NOVIELLI, 2010)(HERBERT; KANG, 2018). Furthermore, artificial intelligence applications seek to combine and imitate human functions and traits to solve everyday problems (DIVYA et al., 2018).

The primary definition of conversational agents is related to a computer program or artificial intelligence able to hold a conversation with humans through . Moreover, conversational agents can emulate a human personality to ease their interactions with real users (ABDUL-KADER; WOODS, 2015). Conversational agents use complex semantic and syntactic grammar or highly technical scripting languages to parse user statements in language construction processes (HERBERT; KANG, 2018).

A conversational agent has been used as a training tool for new therapists, providing real-time feedback with less expensive tasks (TANANA et al., 2019). Urethroscopy patients who have complications after kidney stone removal procedures may use conversational agents to learn about complications and counseling (GOLDENTHAL et al., 2019). Making the pediatric transition to physicians has also been a role played by conversational agents. Agents have been present on social media and smartphones (BEAUDRY et al., 2019)).

The conversational agent can be used in pregnant education to identify and mitigate Preconception health risks, thereby assisting African-American women with less education in this area (JACK et al., 2015). Furthermore, systems that use a text messaging service could distribute health information in an automated manner as a lower-cost alternative for low-income pregnant women (ZHANG; BICKMORE; PAASCHE-ORLOW, 2017). In addition, agents could promote breastfeeding education through counseling techniques (ZHANG; BICKMORE; PAASCHE-ORLOW, 2017).

Counseling on social networking reliable sites during the prenatal and postnatal period may improve maternal and neonatal health outcomes. Patient education has been incorporated into several different proposals. Discharges from hospitals are typically a significant process for users, and education via conversational agents can transform patients into self-sufficient decision-makers responsible for all aspects of their care (BICKMORE et al., 2016). (BICKMORE et al., 2010).

Often interactions with conversational agents reveal useful information. In many situations, information can aid humans in making better choices. For instance, accurate information can be beneficial during prenatal and postpartum. Cesarean deliveries accounted for 42 percent of the pregnant women population. According to the Ministry of Health of births in Brazil in 2018, and in most cases, the decision happens due to a patient's lack of information. Numerous stud-

ies have been conducted over the years regarding patient education in general (BICKMORE; PFEIFER; JACK, 2009)(BICKMORE et al., 2010) (JACK et al., 2015) (ZHANG; BICKMORE; PAASCHE-ORLOW, 2017)(TANANA et al., 2019) (PALANICA et al., 2019)(NOURI; RUDD, 2015).

This thesis makes two major contributions: first, it proposes a model architecture based on conversation agents called (Help in Obstetrician for PrEgant) , and second, it evaluates the acceptance of conversational agents in the context of pregnancy through randomized studies.

The HoPE model is a conversational agent hybrid architecture in health that focuses on information delivery to your target audience: pregnant women. The proposal encourages pregnant women to learn about topics that are priorities during the thousand days of pregnancy. We conduct clinical studies assessments, focusing on conversational agent acceptability using a variety of assessment constructs. In addition, we assess the conversational agent architecture to investigate the ability of information retrieval.

Several studies (MILITELLO et al., 2021)(EGEDE et al., 2021)(CHUNG; CHO; PARK, 2021)(JACK et al., 2020) in the field of health computing have addressed the use of conversational agents to support pregnant women in tasks of education, training or facilitation of daily tasks. We recently discovered several findings that help shape pregnant women's behavior when using this technology. Users are very interested in interacting with conversational agents, especially when the service is personalized (FADHIL; SCHIAVO; WANG, 2019).The literacy levels of pregnant women significantly impact the evaluation of conversational agents in usability (EGEDE et al., 2021). Utility perception and hedonism appear to be relevant constructs for the use of conversational agents (CHUNG; CHO; PARK, 2021).Different assessments of conversational agents can be influenced by territory, ethnicity, or social status (JACK et al., 2020). The use of voice mechanisms can help leverage technology for the widespread use of conversational agents as an aid tool in health literacy, which refers to the ability of a user to understand some information (MILITELLO et al., 2021).

However, we identified some gaps in this field that require several interventions to generate more interesting and informative conclusions. From a territorial perspective, we believe there is still a lack of research in Brazil on the use of conversational agents in reliable information delivery. Nutrition is one of the issues that we have identified as of critical importance today. In second place (JÚNIOR et al., 2021), the pregnant teenager's nutritional status was 37.8 percent underweight, 46.9 percent normal weight, 12.1 percent overweight, and 3.3 percent obese in 2008. The study's most recent findings indicate an increasing trend in the prevalence of overweight adolescent pregnant women in Brazil and its regions, consistent with the global pattern observed in children, adolescents, and adults. We propose a quantitative validation study with pregnant women and health professionals to evaluate a conversational agent that provides nutritional information.

Another gap that we identified was in assessment constructs. Many studies use a variety of theoretical models to assess the acceptability and usability of the conversational agent in health

(BICKMORE et al., 2020) (DEVLIEGER, 2021). For example, model was recently cited as the most appropriate theoretical lens for modeling the adoption of health issues (MOKMIN; IBRAHIM, 2021). However, according to our findings, only one study used this technology acceptance model for conversational agents assisting pregnant women. As a result, we conducted a new study using a mixed methodology, this time involving pregnant women and health professionals from the perspective of the UTAUT2 model. The study used a mixed-method analysis to assess pregnant women's conversational agent experiences and compare them to health professionals' qualitative impressions.

We also discovered some gaps in the architecture of conversational agents in health, which we addressed in this study through the development of the HoPE model. The HoPE model aspires to be a comprehensive architecture for identifying user intents, managing dialog, and retrieving information from input utterances. As domain ontologies have become more prevalent in healthcare, their use in architectures involving conversational agents has increased (TEIXEIRA; MARAN; DRAGONI, 2021) (ADIKARI et al., 2022)(ALTI; LAOUAMER, 2021). Over the last five years, we identified approximately 3,100 studies involving this structure with this theme. Simultaneously, transformer-based natural language processing models have achieved outstanding results in information retrieval and semantic similarity tasks (MASS; ROITMAN, 2020) (PADAKI; DAI; CALLAN, 2020), (ESTEVA et al., 2021).Despite this, these models face significant gaps when used as conversational agents in production environments.

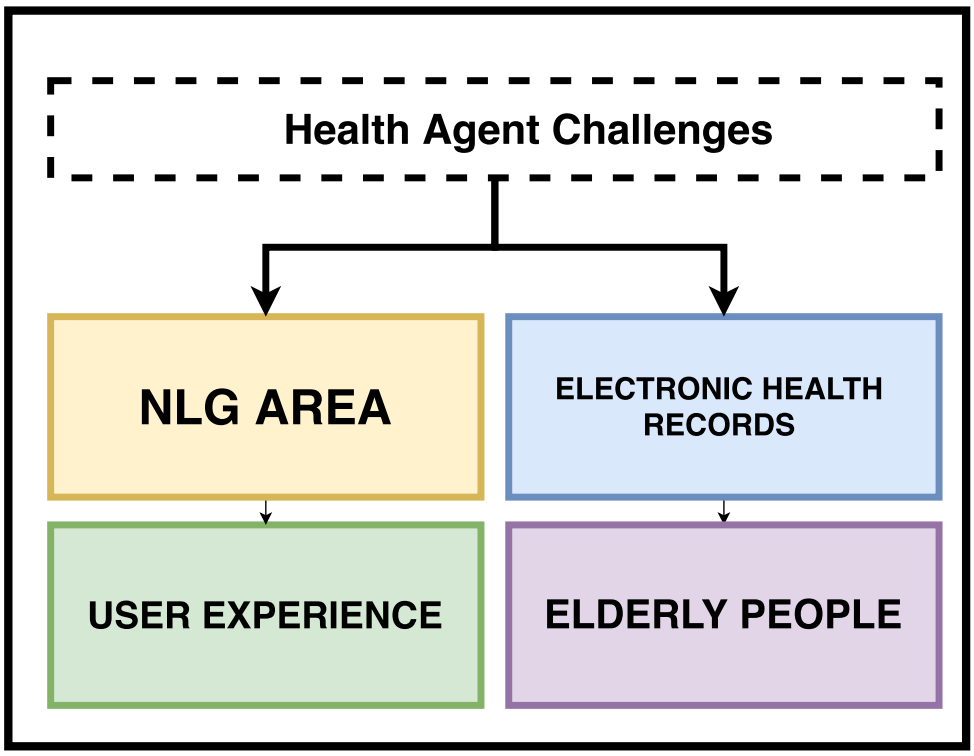
In this thesis, we chose to explore gaps related to deep learning architectures and the use of ontology in conversational agents.

1.1 Motivation

Pregnant women frequently seek information in various formats to fully comprehend their pregnancy status. In these cases, the source depends on the availability and the basic level of education of the person. Other factors, however, are inherent knowledge and the ability to understand the outcome. In difficult moments, a precise source of information could help mothers facing anxiety issues (MUGOYE; OKOYO; MCOYOWO, 2019). The community that studies the conversational agent's application in health has been carrying out several studies to improve the quality of life of human beings through the use of conversational agents (NIKITINA; CALLAIOLI; BAEZ, 2018). Therefore, this is one of the questions we would like to answer in this study: What are the challenges related to conversational agents in health? Some challenges like the usability to improve user experience, the data as a source of real-time information, and new approaches such as the natural language generation. These main areas we previously mapped in follow studies (MONTENEGRO; COSTA; RIGHI, 2019) (MNASRI, 2019) and are shown in Figure 1.

According to current trends, electronic health records may integrate with conversational agents in the future. Hospital records should serve as a basis for consultation with agents and

Figure 1: Future works trends related to conversational agents in health shown in (MONTENEGRO; COSTA; RIGHI, 2019)(MNASRI, 2019)



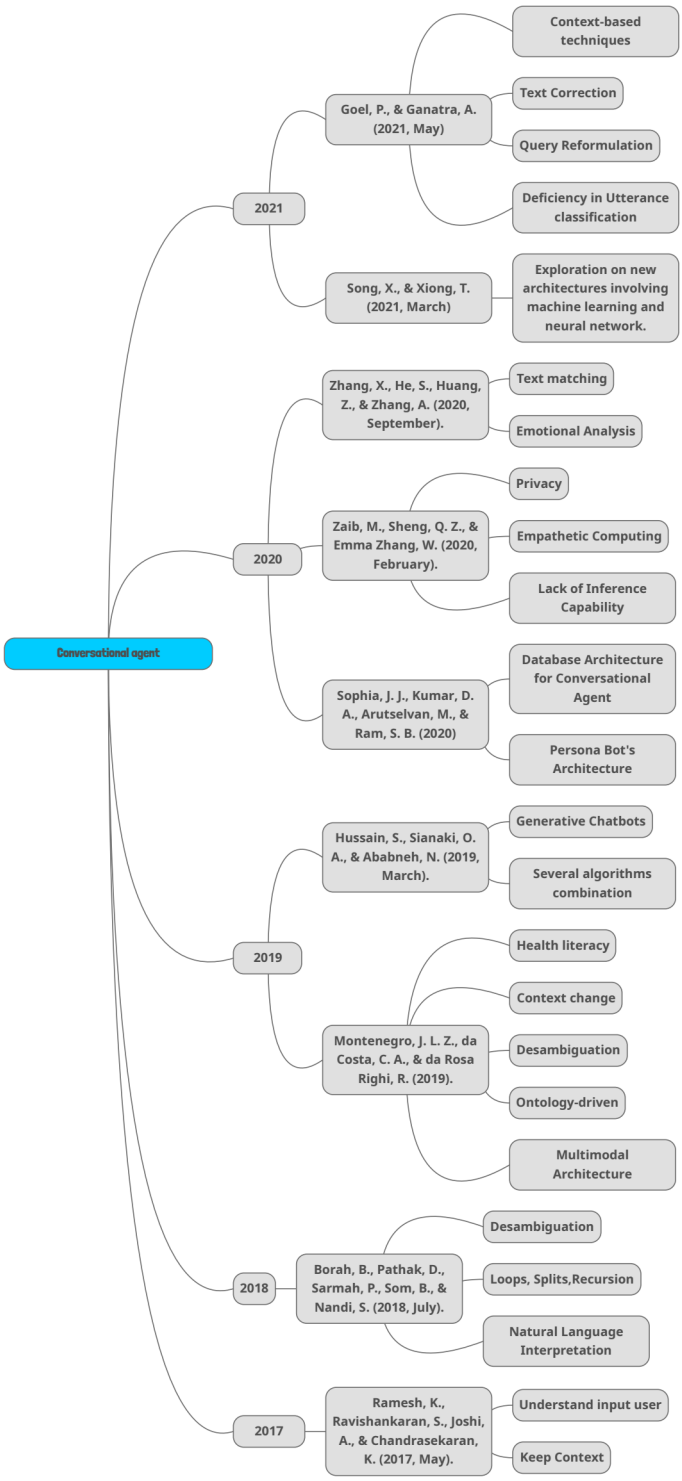
Source: elaborated by the author

should not be in separate stations. The use of EHR jointly by assistants could help prevent disease by providing relevant information to the patient (BICKMORE; PFEIFER; JACK, 2009). Further, the study (TOKUNAGA et al., 2016) improved the care-based quality of the reactions of older adults to agents. Multimodal architecture based on speech interactions for older people remains a further focus of future research (SHAKED, 2017).

Several articles have discussed user-related challenges in interactions and new interfaces (ALESANCO et al., 2017; KOWATSCH et al., 2017; TANAKA et al., 2017b; KASINATHAN et al., 2017). This challenge discussed improvements in patient communication techniques, the emotions used in dialogue, and the interfaces designed to make conversations more realistic. In addition, 3D user interaction interfaces should be more present, offering agent multi-modality (KARPAGAM; SARADHA, 2014). We also look into challenges associated with conversational agent architectures. We chose ten surveys conducted over the last five years to identify the significant gaps in this theme. Some challenges compelled the development of the HoPE model architecture presented in this thesis. We show the main findings in 2.

A challenge often mentioned was disambiguation. Ambiguity is a characteristic of natural language. When a sentence is ambiguous, different words may describe the same simple idea. We constated that conversational agents struggle with loop embedding, split branching, and recursion in the open domain. These strategies become necessary to follow up on prior specific conversations (BORAH et al., 2018).

Figure 2: Gaps identified in nine recent studies involving conversational agent architectures



Source: elaborated by the author

Multimodal interactions are challenges addressed in (MONTENEGRO; COSTA; RIGHI, 2019). User engagement can be increased through architectures that support voice, text, and avatars. Although ontology-based agents have become more prevalent in health, additional research is required. Another aspect of health is the architecture that aims to increase patient literacy through conversational agent uses. Open-domain chatbots may encounter issues processing a disparate utterances collection, necessitating multiple algorithms to ensure the system's accurate response. Another issue is the response generation-based algorithms, which are still in their infancy compared to recovery-based agents (HUSSAIN; ATHULA, 2018).

In terms of the architecture of personal chatbots, it is widely accepted that implementing personalized medicine would save many lives while also increasing public awareness of medical issues. The chatbot's efficiency may be increased by adding additional phrases and expanding its database, allowing the medical chatbot to manage all forms of diseases (SOPHIA et al., 2020). The system can be made more user-friendly by including audio communication as well. Additionally, conversational agents face gaping gaps in terms of privacy and empathy. Architectures that respect user privacy and are sensitive to emotions via text or voice dialogues are unsolved problems (ZAIB; SHENG; ZHANG, 2020). Analyses of dialogue's emotional prism were presented as future trends (ZHANG et al., 2020). Additionally, text-matching techniques deserve to be investigated further for chatbot architectures.

According to (GOEL; GANATRA, 2021)study, the worst aspect of conversational agent architecture development is determining how to keep up with a constantly changing conversation. Context-based strategies can also correct textual errors, another challenge for conversational agent architectures. Additionally, this work observes a difficulty associated with unseen inputs during the training of NLP processes, which makes the model's classification uncertain. Similar gaps were found in (RAMESH et al., 2017).In (SONG; XIONG, 2021),it is noted that implementing new machine learning algorithms and neural networks in conversational agents presents a challenge. Some approaches continue to face difficulties in practice.

1.2 Research Question

Conversational agents can serve as excellent encouragement for pregnant women in challenging times or as a way for the public to access vital and reliable information. Parallel to this, there is a growth in the use of conversational agents within the health context, with new architectures and new approaches. However, we identified some gaps in the literature that still need further investigation. There are many studies involving chatbots in the context of pregnant women worldwide. Among them, few studies aim to explore the views of doctors and pregnant women regarding the use of this tool in the same experiment. Although the number of studies on conversational agents globally has grown in recent years, Brazil's number is still low. We understand that the territory makes a difference in the perception of the assessment of technological tools.

Another gap perceived in our review was related to the architectural perspective. Some models, such as transformer architectures, obtained relative success in recent years for numerous studies. These models are considered important for information retrieval due to their ability to understand the context and resolve ambiguity. However, few transformers models are being used for other techniques in conversational agents. How to use transformers in conversational agent architectures, comparison with traditional architectures, what they can help solve, and performance vs. runtime are some gaps, and further research is needed. Another challenge perceived was within the task of recognizing intentions for "Out-of-scope" sentences. How to handle sentences not expected by the information retrieval model or with multi-intentions in the same sentence? In light of these and other factors raised in this section, we understand our research question to be answered is:

How effective are the HoPE model architecture and conversational agents for pregnant women education?

This research question is divided into four specific questions:

- *What are the pregnant perceptions of the conversational agents support during pregnancy?*
- *What are the health professionals' perceptions of the conversational agents support in pregnancy?*
- *What are the best fine-tuned model performances to deliver health literacy?*
- *How does the HoPE architecture perform for health literacy*

We employ four methodologies to reach the specific issues previously mentioned. These are detailed below:

- A survey to understand the literature on conversational agents in health in general
- The use of clinical studies to better understand the behavior and perceptions of pregnant women, doctors, and other health professionals regarding conversational agents
- The application of embedding correlation tests in Transformer models adjusted to Brazilian pregnancy guidelines data
- A test collection application for assessing the HoPE architecture in accuracy, disambiguation, and speed inference

The main contribution of this work is the development of conversational agent architecture for providing prepartum and postpartum health literacy to pregnant women. The architecture contribution is a hybrid model designed to provide trust knowledge to help with pregnancy concerns. Hybridism refers to two different mechanisms to retrieve information: an ontology for managing dialog and a model based on transformer architectures for retrieving trust information. We also contribute to the study of the technological adoption of conversational agents for pregnant women in the Brazilian context through clinical studies that point to new directions for future work. As secondary contributions, we also listed:

- A fine-tuned Transformer model on Brazilian pregnancy guidelines
- A full hybrid architecture encompassing ontology and a transformer model
- An ontology knowledge base populated with pregnancy guidelines content
- A chatbot trained with pregnancy guidelines information
- A NLI (Natural Language Inference) corpus for sentence text similarity tasks based on pregnancy guidelines sentences

1.3 Document Organization

There are seven chapters in this thesis, including this one. Chapter two covers the main concepts required to understand this thesis. Themes relating to the prenatal and postpartum periods are elucidated. We also examine the health literacy concept, often discussed in the literature on health education. Finally, we examine the Natural Language Processing concepts that drive conversational agent designs, delving into the information retrieval models and ontologies that serve as the foundation for the architecture proposed in this article.

Chapter three will present research specifics to the proposed project, detailing the state-of-the-art literature on conversational agents in health, expanding its strategies and key outcomes. Chapter fourth details the experimental setup of the HoPE architecture used to overcome the challenges. In this chapter, we present how the architecture and its components work.

The fifth chapter discusses the thesis's methodology and organization. It also specifies the initial resources that will be available, the structure used, and the use cases that we considered during the chatbot development process. Chapter six provides a human judgment evaluation of the chatbot through user testing and discusses the main outcomes, analyzing aspects of the study and comparing it with other research already carried out in the area. Finally, chapter seven summarizes the research findings and suggests new directions for this study in the future.

2 BACKGROUND

This chapter introduces the key concepts that will underlie this project. These principles are directly linked to the proposed architecture model, serving as the basis for its development. The main definitions that we will cover in the chapter are 2.1 Thousand Days, 2.2 Health Literacy, 2.3 Conversational Agent In Health, 2.4 Conversational Agent High-Level Structure, and 2.5 Conversational agents related fields.

2.1 Thousand days period

Based on the Lancet series concept (PATEL et al., 2008),“Thousand Days” identify the first thousand days of life (encompasses the approximate 270 days of pregnancy plus the 730 days of the baby’s first two years) that are critical to the health of the mother and child. During this time, the pregnant woman faces some challenges. Women who are pregnant are more vulnerable to stress. According to data from the US pregnancy risk assessment monitoring system, nearly 75% of postpartum mothers reported at least one major stressful event in the year preceding their baby’s birth in 2009 and 2010 (TRAYLOR et al., 2020).

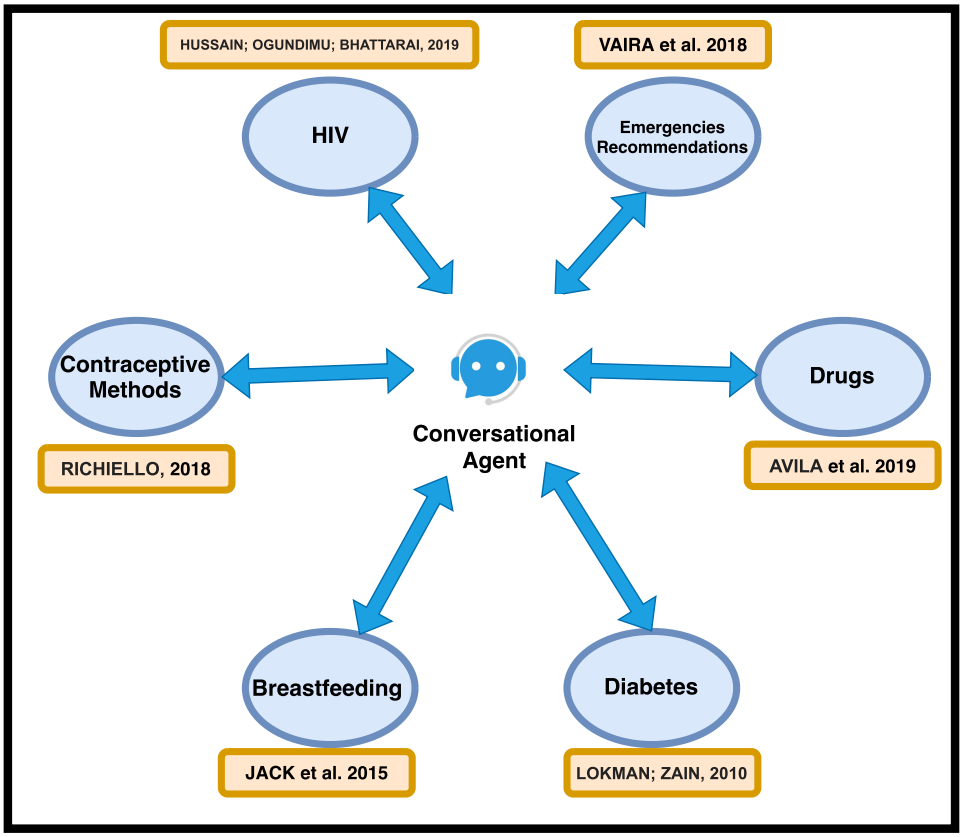
This period is also notorious for high levels of anxiety. If the mother is stressed or anxious during pregnancy, these well-established risk factors for premature birth, low birth weight, and infant health problems may have long-term effects on the offspring. Pregnancy anxiety affects approximately 21% to 25% of expectant mothers (e.g., excessive worry, nervousness, agitation) (MOYER et al., 2020).

When confronted with these and other symptoms, pregnant women seek information to alleviate them. During pregnancy, its frequently used the internet as a source of information (70–97 percent) (BJELKE et al., 2016). According to one study of pregnant women, the internet was frequently used for seeking pregnancy information, verifying information received from health professionals, social networking, social support, and electronic commerce (i.e., e-commerce) (LAGAN; SINCLAIR; KERNOHAN, 2010).

While the internet provides a wealth of information to pregnant women, it is unknown how it affects their decision-making process. Pregnant women are believed to be conflicted and worried during the decision-making process because they do not trust the information they read online (SAYAKHOT; CAROLAN-OLAH, 2016). Examining the internet’s impact on pregnant women can help to enhance their decision-making process. To create meaningful online tools, health care clinicians and web developers must understand how and why pregnant women use the internet while making decisions (LAGAN; SINCLAIR; KERNOHAN, 2010).

The study (CRISS et al., 2015) carried out a survey with some of the main topics and ways pregnant women look for information during the “Thousand Days.” One of the questions refers to the immediacy in the search for some information that needs speed. The vast of women across all demographics family members are a relevant information source about healthy pregnancy

Figure 3: Conversational agents articles in pregnancy area from 2010-2020 aiming Health Literacy



Source: elaborated by the author

and childbirth. Women with children in early childhood reported that they often follow their intuition for decision-making. In addition, all women use the internet to find fast information, using technologies such as Google and other search engines. However, trusted websites, such as health guidelines-based apps, are less popular among the groups surveyed (CRISS et al., 2015).In the following section, we present a concept directing many studies aimed at pregnant patients and meeting the demand for information in the health field.

2.2 Health Literacy

Health literacy is the concept that reflects a range of skills and tools related to the individual's capacity to reach health-related information (RAHMANI et al., 2019). In summary, this definition includes reading, listening, assessing, and decision-making skills and the capacity to apply those skills to health conditions (SAFAIE et al., 2019). Patient education importance is presented at all medicine levels, as shown in 3. .

Hospital discharges are usually an unusual process for users, and the education procedure through conversational agents can turn patients into fully responsible decision-makers for all aspects of their care (BICKMORE; GIORGINO, 2006). Furthermore, descriptions of documents

by a lay client using medical avatars in face-to-face encounters may allow physicians to expand their practices while providing reliable information to patients (BICKMORE; SCHULMAN; YIN, 2009).

In another context, agents are applied to promote nutritional education, thereby overcoming the current limitations of some solutions and allowing for more effective health interventions among users who participate in this field. The health literacy models for low-income people are widespread due to the lowest levels of this population in the health sector (FADHIL; GABRIELLI, 2017).

There are works related to pregnant education through conversational agents. SMS systems may give healthy automated information through mobile phones as an alternative to low-cost pregnant women. A study showed that agents could promote breastfeeding education through therapy techniques. Counseling through social networking sites during the prepartum and postpartum periods could improve maternal and child health outcomes (ZHANG et al., 2014).

Health literacy is relevant during pregnancy because the mother's health habits affect her own and her child's health. Individuals with inadequate health literacy have more emergency department visits and longer hospital stays, poorer healthcare outcomes, and lower utilization of preventative interventions than those with good health literacy. International research applied in Europe, for example, found that 13% of respondents had inadequate health literacy and 47% had limited health literacy, with significant disparities among nations (NAWABI et al., 2021).

2.3 Conversational Agent In Health

Conversational agents with verbal or non-verbal communication features do immersive human-to-machine activities (WARGNIER et al., 2015). Often called relational agents, they can mimic certain human characteristics by using different interactions, such as voice, gaze, hand gestures, and other non-verbal modes (WANG et al., 2015).

Engaged users are critical to the success of a conversational agent, which is defined as an ongoing dialogue between the agent and the user. This process should check and analyze the user's behavior. In this context, conversational agents play a vital role in interaction, using text strategies to maintain a dialog for more time (WARGNIER et al., 2015). In addition, agents are useful tools for human-machine interaction, allowing data input through natural language, processing sentences, and returning accurate answers through text or voice (GALVAO et al., 2004).

Several agents use probabilistic approaches to enhance communication. Natural Language Processing is one of the primary tools for human-to-agent interactions, focused primarily on signal processing, syntactic and semantic discourse analysis, and pragmatics, allowing natural language generation to provide meaningful conversation based on context analysis (GANESH et al., 2008). In this way, conversational agents can successfully create an informal partnership with their users (IFTENE; VANDERDONCKT, 2016). Conversational agent studies have

sought information related to many domains to help mothers do different tasks in the pregnancy area. Conversational agents can be crucial in different healthcare situations. For medical students, agents can be training decision-making skills regarding thromboembolism, using natural language processing techniques and machine learning for interactions (DOLIANITI et al., 2020).

Adults with attention deficit disorders need to be diagnosed and treated as soon as possible. Many patients, however, do not receive adequate therapy due to time, space, and financial constraints. Based on feasibility and usability metrics, it is proposed in (JANG et al., 2021)to investigate the possibility of reducing attention deficit symptoms using a chatbot intervention. In the health field, agents for education have been extensively studied, and several applications in clinical and patient care have been made (WANG et al., 2015). Robots are continually present in elderly care with features that support physicians by providing advice and insights for the elderly patient treatment (HEERINK et al., 2010). Conversational agents act in any environment to gather and use information in their interactions through knowledge engineering processes (LÓPEZ; EISMAN; CASTRO, 2008). This collection process may allow an agent to do different functions based on current necessity, e.g., some studies have used speech and language techniques to detect dementia in patients (TANAKA et al., 2017a)(FRASER; MELTZER; RUDZICZ, 2016)(ARAMAKI et al., 2016).

Primary care nurses are under constant stress because of the growing number of patients. The chatbot can help nurses and patients by providing evidence-based content to support care management. In (ROQUE et al., 2021) describes the development process of a BOTCURATIVO chatbot to assist non-specialists with wound treatment by providing a step-by-step guide to wound dressing recommendations for each type of wound. Other vaccine information, such as HPV (the human papillomavirus), was tested using a conversational agent guided by an OZ Wizard protocol (AMITH; ROBERTS; TAO, 2019). CoachAI engages its users in health-related activities via a conversation medium to promote a healthy lifestyle. The method employs a clustering algorithm to assist physicians in assigning activities to patients and grouping patients together (FADHIL; WANG; REITERER, 2019). Several types of research were conducted in obstetric care using conversational agents for obstetric support (CHUNG; CHO; PARK, 2021). A Q&A knowledge database-based chatbot (Dr. Joy) was developed and tested for perinatal women, focused on obstetric and mental health care, using a text-mining technique and contextual usability testing (UT).

2.4 Conversational Agent High-Level Structure

A Conversational Agent architecture can be composed of several modules and have different characteristics. Pattern matching is one of the modeling options which can be used. Representative stimulus-response blocks are essential to pattern matching. Responses are generated in answer to a user's input (stimulus). However, this method has the drawback of producing an-

swers that are entirely predictable, repetitive, and devoid of individuality. In addition, there is no way to keep track of previous responses, which can lead to endless back and forth.

The conversational agent normally uses natural language processing modules in its structure to develop into a more dynamic conversation. Natural Language Processing (NLP) is a concept related to comprehending and controlling natural language text or speech to accomplish efficient tasks. Machine translation, natural language text processing, summarization, user interfaces, multilingualism, and information retrieval are examples of NLP applications (BIRD; KLEIN; LOPER, 2009). NLP is a significant area of computing that focuses on maximizing the efficiency of user interaction with computers (JAIN; KULKARNI; SHAH, 2018).

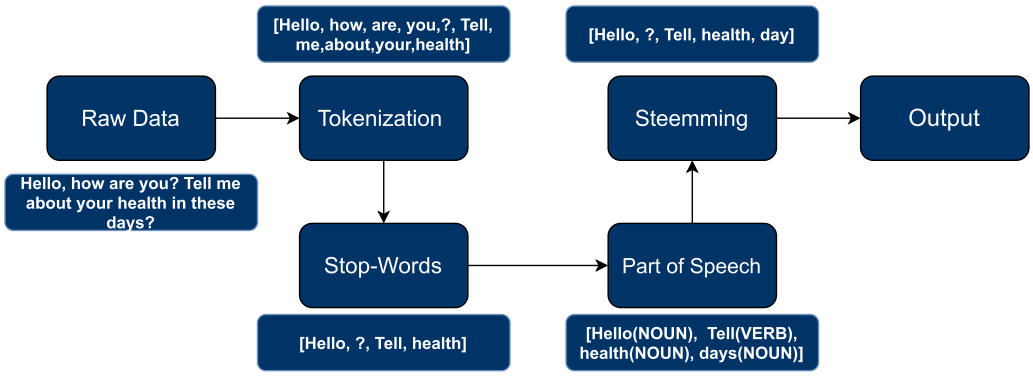
Vector representations are used in health studies with combination techniques to capture analytical and semantic relationship properties between words. Numerous trials utilizing word embedding have been conducted for common natural language processing (NLP) tasks such as information extraction (IE), information retrieval (IR), sentiment analysis, question answering, and text summarization (WANG et al., 2018).

Many studies (MORA et al., 2021)(NIKFARJAM et al., 2019)(ZHANG et al., 2021) have shown that the preparation pipeline for health-related texts varies. These pipelines carry out a variety of functions that vary depending on the problem, data, and data type. Natural Language Processing, as exemplified by (CIAMPI et al., 2020),may play a role in architectures that seek to retrieve information from unprocessed documents, such as the development of corpora, the treatment of input messages, and the generation of responses. Ambiguity is a significant difficulty that Natural Language Processing faces when dealing with conversational agents. Ambiguity refers to the phenomenon of having many interpretations for the same information, which occurs in all human languages at all levels. This feature enables a high degree of diverse semantic and syntactic data and is widely regarded as one of the most difficult aspects of human interaction via language (COHEN; DEMNER-FUSHMAN, 2014).

There are several possible processing methods within NLP pipelines. It is normal to find processes like Tokenization, Stop-Words, Part of Speech Part of Speech Tagging (POS), and Stemming processing commonly in conversational agents as shown in Figure 4. Tokens are collections of characters that together form a semantic unit. Tokenization is the process of splitting these strings (ZIA; RAZA; ATHAR, 2018). Stopwords frequently appear in many different natural language documents or parts of the text in a document but carry little information about the part of the text they belong to. Hence, removing stop-words can increase the signal-to-noise ratio in unstructured text and thus increase the statistical significance of terms that may be important for a specific task. Example of Stopwords include "each", "about", "such", and "the" (SARICA; LUO, 2021).

POS determines the most likely syntactic category for each word occurrence in a sentence. Tagging creates a relevant intermediate representation that can be used for tasks such as speech synthesis, information retrieval, and machine translation (BR; KUMAR, 2012). Stemming is often used in natural language processing applications, including information retrieval, POS,

Figure 4: Natural Language Processing Pipeline for a conversational agent



Source: elaborated by author

syntactic parsing, and machine translation. A morphological procedure aims to turn a word's inflected forms into its root form (ABDUL-KADER; WOODS, 2015).

Inside of natural language processing, the NLU is a process often used in conversational architectures. NLU NLU is usually composed of two steps: slot-filling and intent detection. Typically, these systems treat intent recognition as a classification task over the whole sentence while slot labeling is addressed using a sequence labeling approach. Because it entails identifying the most significant slots and then extracting the utterance's overall intent, slot-filling is a popular method for resolving the NLU problem. While the slots are defined as the entities and relations required for a specific activity, the user's intention determines the purpose. For instance, "Can I continue to exercise while pregnant?" the objective (do exercises) appears as intent, while the actions are slots (physical activity, "gym"), (a situation, "pregnant") (ABRO et al., 2020).

The conversational agent NLU is typically also made up of fallback intents. This type of intention is triggered when the system does not feel safe responding to an utterance. Most approaches use this moment to apologize to the user for not understanding the answer or not knowing about the statement's content (HU et al., 2021). There are two common approaches in this field: the knowledge-guided approach and the statistical approach. In this thesis, we provide both. Chapter five discusses these approaches (REN; WANG; LIU, 2020).

In general, conversational agent systems are made up of three main components: intents, entities, and contexts. An intent links what a user says and what the conversational agent should do. Conversational agent actions are the actions taken by the conversational agent in response to user inputs triggered by specific intents. In general, intent detection is a subset of sentence classification in which one or more intent labels are predicted for each sentence (BANSAL; KHAN, 2018)(RAMESH et al., 2017). A tool for extracting parameter values from natural language inputs is the entities. Entities can be defined by the system or the developer (BANSAL; KHAN, 2018)(RAMESH et al., 2017). We also have domain entity extraction, also known as a slot-filling problem, formulated as a sequential tagging problem in which sentences are extracted and tagged with domain entities. An object's context is stored in a string called a con-

text, which the user can refer to or discuss. Using an example, a user may refer to a previously defined object in his next sentence. It's possible to type in "Turn on the fan." A user may say "Switch it off" as their next input, and the intent "switch off" will be invoked in the context "fan" (BANSAL; KHAN, 2018)(RAMESH et al., 2017).

2.5 Conversational agents related approaches

In this section, we discuss concepts related to conversational agent architectures. We present sections covering architectures used for information retrieval, which frequently is applied to conversational agents. Finally, we discuss ontology's structures used in semantic systems.

2.5.1 Information Retrieval Architectures: BM25

Although the term "information retrieval" appears to encompass a broad range of subjects, it is most frequently used to relate to the retrieval of narrative data. Information retrieval systems can process letters(KADRI; NIE, 2006), newspapers (LARKEY; BALLESTEROS; CONNELL, 2002), medical summaries (GOEURIOT et al., 2016), and other contents. When referring to a broader range of activities, such as document or text retrieval, the term "information retrieval" may be used as (IR) (HERSH; HERSH; WESTON, 2020). One of the state-of-art IR models is BM25. The BM25 model has different variations (LV; ZHAI, 2011) and is a popular benchmark for information retrieval tasks. Word vectors attempt to reduce the problem's complexity by moving away from TF-IDF techniques, which require us to one-hot-encode the entire vocabulary to work with them successfully (SAFDER; HASSAN, 2019).

$$Query = (x_1, ...x_n)$$

$$BM25(w_i, Query) = \frac{(k+1)c(w_i, Query)}{c(w_i, Query) + k(1 - b + b\frac{|d|}{avdl})}$$

The BM25 document score is calculated for the term frequency in the document *(f_i*,|*d*| is the length of the document in words, and is the average document length in the text collection from which documents are drawn, given a query *(x_1, ...x_n)* containing keywords. *(K)* and *(b)*are free parameters usually chosen without advanced optimization. *(Q_i)* is the number of documents that contain.

2.5.2 Information Retrieval Architectures: Transformers

Machine learning systems have frequently employed word embedding systems as a feature, enabling new techniques to contextualize raw text data (MIKOLOV; YIH; ZWEIG, 2013). Recent dataset enhancements such as GLUE (WANG et al., 2019) and SQuAD (RAJPURKAR et al., 2016)have driven the development of NLU systems based on statistical approaches and

embeddings.

The majority of these benchmarks, on the other hand, imply that the model has access to a large amount of manually labeled data. As a result, the few-shot setting has received attention as a critical component of testing NLU performance. Most NLU research on few-shot learning evaluates it using randomly selected subsets of existing datasets (MUKHERJEE et al., 2021).

Compared to embeddings learned from scratch, NLU's strategy to transfer pre-trained neural language representations improve downstream task scores (ALOMARI et al., 2021).More recent work, including but not limited to (ALAMBO et al., 2021) (BOUDJELLAL et al., 2021), has added end-to-end fine-tuning of language models for downstream tasks extraction of contextual word representations, expanding on these ideas even further. As a result of this advanced engineering and large compute availability, state-of-the-art NLU's Transformers architecture has evolved from word embeddings to transferring language models with billions of parameters achieving unprecedented results across NLP tasks (LIU et al., 2020).

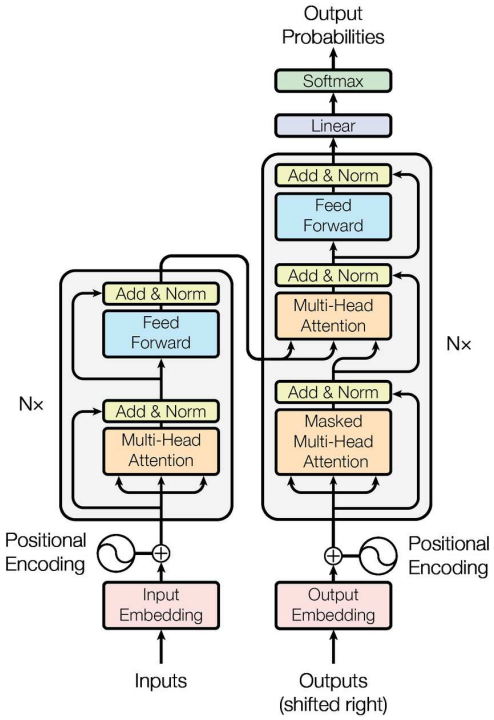
Original Transformer is a six-layered encoder-decoder model that generates a target sequence based on the encoder's output. The encoder and decoder have a high level of self-attention and a feed-forward layer. By adding an attention layer between it and the encoder, the decoder can map its relevant tokens to the encoder for translation purposes. Self-attention enables the look-up of remaining input words at various positions to determine the significance of the currently processed word. This is done for all input words to improve the encoding and context understanding(SINGH; MAHMOOD, 2021). We present an illustration of this architecture in Figure 5.

Regarding the evolution of word embedding systems, transformers address a significant issue with previous architectures: the models were static, which meant that each word had a single vector regardless of context. This causes a slew of issues, not the least that all possible meanings of a polysemic word will share the same image. Contextualized word representations and context-aware word vectors were generated by transformer architectures (PETERS et al., 2018), (DEVLIN et al., 2018) (RODRIGUES et al., 2020). Many architectures such as GPT-3 (Generative Pre-Training Transformer 3) (FLORIDI; CHIRIATTI, 2020), BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2018), and ELMO (Embeddings from Language Model) (PETERS et al., 2018) have already been used frequently in recent studies. The following section aims to introduce the BERT architecture and its Sentence-BERT derivation.

2.5.2.1 BERT

The BERT (Bidirectional Encoder Representations from Transformers) model architecture reached cutting-edge tasks such as text extraction, question answering, and entity recognition (LEE et al., 2020). BERT, unlike other models, does not provide a single word embedding after your training for each word. Given the complete sentence, it provides a model that generates

Figure 5: State-of-art transformers architecture



Source: elaborated by (VASWANI et al., 2017)

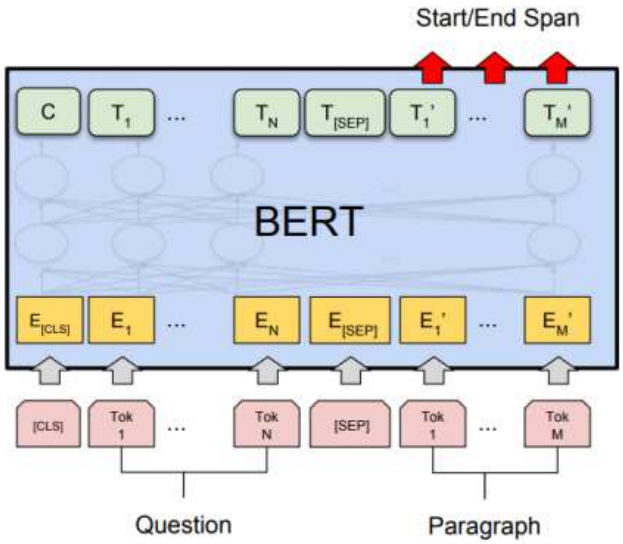
a word integration for every word within the context of the sentence. This means that the sentences "I came to the bench today" and "I am sitting on the park bench" offer the term "bank" with distinct representations in each of them (DEVLIN et al., 2018).The BERT architecture is designed so that during training and predicting embedding, both previous and next words are taken into account. Attention methods are employed to retain the most and least significant word information in the phrase (DEVLIN et al., 2018).

The model architecture proposed by BERT is shown in Figure 6. Being a Transformer Encoder, BERT is a Bidirectional model because of the complexity of the encoder self-attention within the Transformer architecture. BERT provides advanced methods for obtaining contextualized word embeddings for many NLP approaches. Moreover, what does that mean? BERT-based NLP models outperformed previous state-of-the-art results in 11 different NLP tasks, including a question-answer, when the BERT paper was released in 2018 (DEVLIN et al., 2018).

A pre-trained BERT model acts taking into account their context: the last word in the secret state of the Transformer's Encoder (VASWANI et al., 2017) (DEVLIN et al., 2018). As has been evaluated in other articles (LEE et al., 2020) (WANG et al., 2016), the BERT model obtained good results when used for text mining in the medical literature.

While BERT has achieved new state-of-the-art outputs for many downstream natural language processing tasks, the models still findings are insufficient for some tasks involving time execution. One reason for this is the mechanism by which multiple sentence pairs must be checked during inference, which can sometimes result in a slow process (REIMERS; GUREVYCH,

Figure 6: Example of Bidirectional Encoder Representations from Transformers for Question and Answering



Source: elaborated by Develin et al.2018

2019) (ROGERS; KOVALEVA; RUMSHISKY, 2021).

In a conversation with humans, the time response of conversational agents is critical (AMITH et al., 2019). Next section, we present Sentence-BERT (SBERT)(REIMERS; GUREVYCH, 2019), a derivation of BERT that overcomes this type of difficulty.

2.5.2.2 Sentence-BERT

Regarding the different methods and architectures of sentence embedding, the Siamese architecture triplet network presents a valid alternative to derive embeddings from semantically significant sentences that can use cosine similarity measure (REIMERS; GUREVYCH, 2019). Siamese structures of neural networks applied to pre-trained BERT models have often been associated with semantic research tasks, the similarity of sentences, and information retrieval (SOUZA et al., 2020) (REIMERS; GUREVYCH, 2019) (GUO et al., 2020). During the preliminary investigation, Siamese recurrent and convolutional architectures are used for discovering sentence pairs similarities in Chinese, outperforming recurrent's architectures accuracy (WANG et al., 2019).

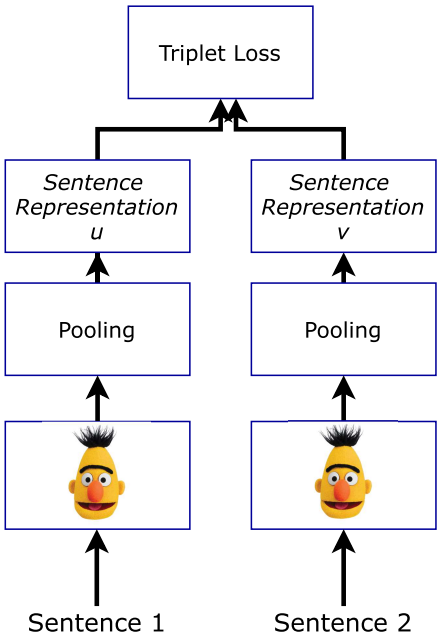
The efficacy of these strategies with BERT was also demonstrated in the (YANG et al., 2019b) study, which proposed a method that uses a pre-training model to encode texts separately and then interacts with the representation vectors to generate attention weights and new vectors, allowing to be pooled and aggregated.

SBERT works over a siamese structure applied in pre-trained BERT models and has been applied for information retrieval, semantic research, translation, and summarizing. Information retrieval works on the following principle: if you feed a short text string and a longer document,

it will return a numeric value between 0 and 1, indicating how closely the two are related. The SBERT model's semantic embeddings run into trouble when dealing with a small number of large documents. For this reason, it is crucial to a fine-tuning process to enrich this model for more accuracy.

The standard training process uses a new Bi-Encoder (SBERT) on the labeled target dataset (REIMERS; GUREVYCH, 2019). It works passing sentence pairs (A / B) in a neural network where each sentence (A / B) yields the embedding you and v, as shown in Figure 7. Cosine similarity is used to calculate the similarity of these embedding and the result comparison to the gold similarity score. This enables the network to be fine-tuned and recognize sentence similarity. The training on data is limited to the upper layers of the pre-trained model to perform "characteristic extraction," which enables the model to utilize the representations of each model (REIMERS; GUREVYCH, 2019).

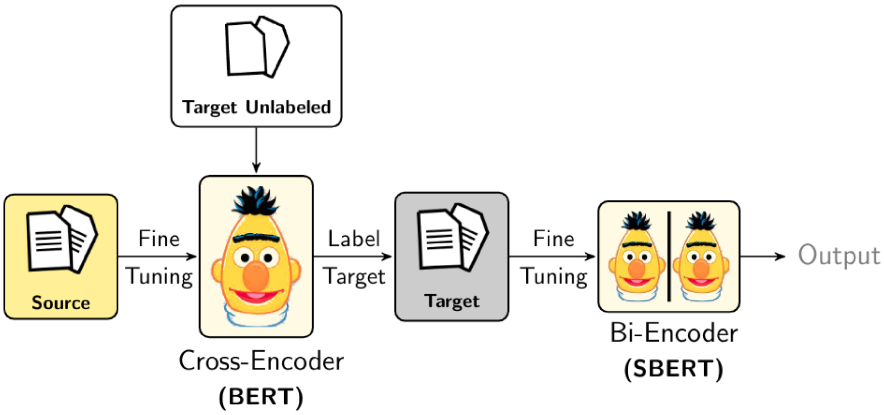
Figure 7: The Bi-Encoder structure provided by SBERT to fine-tune data to pre-trained models



Source: elaborated by Reimers et al.2019

We can train SBERT on data sets comprised of a few pairs using a data augmentation strategy (1k-3k). SBERT augments annotated or unannotated sets and significantly improves results in models fine-tuned with a few data points (REIMERS; GUREVYCH, 2019). The data augmentation training process is depicted in Figure 8. An annotation-enabled BERT cross-encoder is used to train the data. As a result, this encoder (THAKUR et al., 2020) is used to train unlabeled sentences.

Figure 8: SBERT design for the data augmentation strategy to create new sentence pairs for fine-tuned dataset (REIMERS; GUREVYCH, 2019).



Source: elaborated by Develin et al.2018

2.5.3 Knowledge Base

The development and design of ontology is a complex process that requires knowledge management and subject matter experts. Ontologies have grown in popularity in recent years, thanks largely to semantic web development. Ontology systems rely on knowledge acquired through specific formal language, represented in multiple ways. These systems are intended to generate new knowledge from existing data(QIAN; LIANG; DANG, 2009) (MALIZIA et al., 2010). The study (NOY; MCGUINNESS et al., 2001) defines an ontology as a specific domain concept with annotations in its structure relating to the domain's elements.

The ontological structure deals with a concept composition (Classes), the properties of each concept (Properties or Slots), instances (Individuals), and the limitations (Role Restrictions). A class defines concepts that form part of a given domain. These classes have many types, named subclasses. Properties are responsible for explaining the features of the vaccinations as reasons for taking, possible adverse effects, and other details. The Knowledge Base creation occurs by defining individual instances and properties of these classes by filling in specific value information and limiting more properties (NOY; MCGUINNESS et al., 2001).

The ontology could share terms and definitions. Even though the ontology alignment generates two distinct original ontologies, these are incorporated into the link between their equivalent terms. Compatible ontologies can use each other's knowledge via these connections. Ontology mapping generates expressions that link domain knowledge terms, which results in the formal structure. This mapping can be used to relocate data instances, combine and integrate schemes, and carry out other tasks. The integration process creates a unique structure by assembling, extending, specializing, or adapting other ontologies from various subjects (NOY; MCGUINNESS et al., 2001).

Data is stored in two forms to model knowledge: A more complex structure is OWL Ontol-

ogy Web Language representation, which maps all things that the agent can infer around a domain, and a simpler structure is *RDF* Description Framework which straightforwardly specifies facts and relationships (NOY; MCGUINNESS et al., 2001). The *SPARQL* is a query language used in ontologies for knowledge extraction that connects the RDF structure of an ontology to the SQL language of a typical database (BAKOUAN et al., 2018) (RAJOSOA et al., 2019).

Three types of *RDF* data exist IRIs, blank nodes, and literals (CONSORTIUM et al., 2014). All information in *RDF* is represented as triples of the type ***(s, p, o)***, where *s* denotes the subject, *p* denotes the predicate, and *o* denotes the object. Each collection of *RDF* triples can be visualized graphically as an edge-labeled graph, with nodes representing subjects and objects and edges labeled with the appropriate predicates. (RAJI; SURENDRAN, 2016). As a result, collections of triples are frequently referred to as *RDF* graphs. Among the various representations, the (PICALAUSA; VANSUMMEREN, 2011) 4-tuple structure can be used to define the *SPARQL* Query ***Q***.

(query-type, dataset-clause, pattern P, solution-modifier)

Q is based on the graph pattern ***P***, which searches the input *RDF* dataset for defined subgraphs. It generates a set of (many) mappings, each of which connects variables to items. Optionally, the dataset clause specifies the *RDF* dataset for pattern matching. If it is not present, the query processor determines which dataset to use in its absence. Optionally, you can rearrange the pattern-matching mappings and return only a subset of the mappings (e.g., mappings *L* to 10). As a result, there is a list *L* of mappings. The query-type then defines the *SPARQL* query's actual result: select queries return projections of mappings from *L* (PICALAUSA; VANSUMMEREN, 2011).

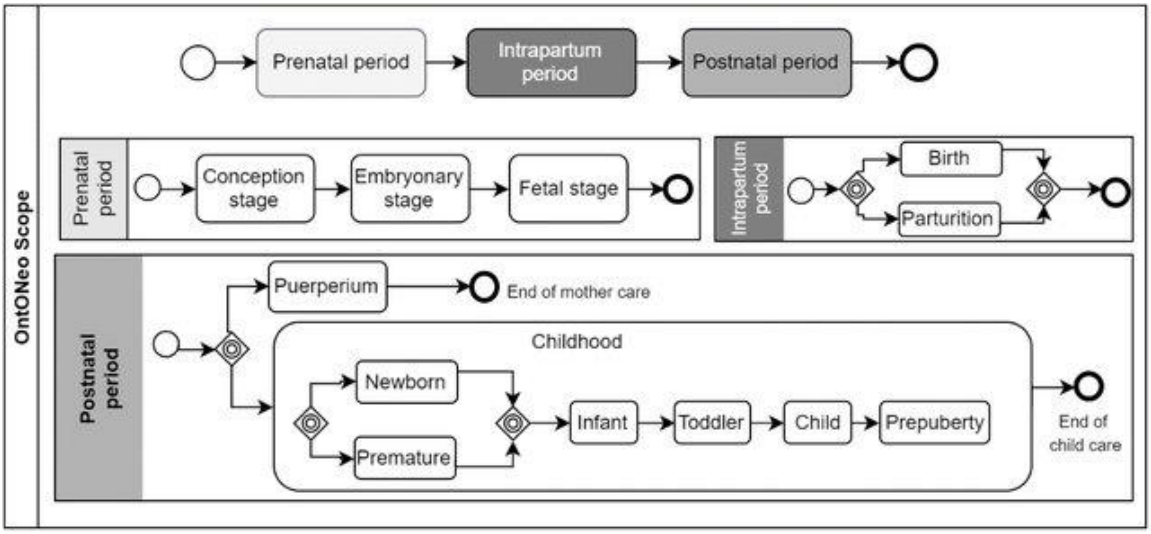
The structure of *SPARQL* queries is defined by: ***Select queries:*** return (in sequence) projections of mappings from *L* ; ***Ask queries:*** produce a boolean true if the graph pattern P was found in the input *RDF* dataset, false otherwise. ***Construct queries:*** build a new collection of *RDF* triples based on the mappings in L ; ***Describe queries:*** generate a set of *RDF* triples that describe the IRIs and blank nodes discovered in *L*.

SPARQL searches can be built with additional syntaxes such as UNION, OPTIONAL, FILTER, REGEX, and point concatenation. The query below attempts to determine whether or not there is a relationship between two entities (classes) via an object property.

2.5.3.1 OntONeo: Obstetric and Neonatal Ontology

OntONeo is a healthcare domain ontology that represents knowledge from electronic health records (EHRs) used in the care of pregnant women and their babies. This source of information is more personal, as it contains relevant information from the pregnant woman's medical record

Figure 9: OntONeo ontology representation scope



Source: elaborated by Farinelli et al. 2018

Table 1: OntONeo metrics

Class	N
Classes	1,797
Individuals	17
Properties	452
Maximum depth	13
Maximum number of children	27
The average number of children	3
Classes with a single child	236
Classes with more than 25 children	3
Classes with no definition	625

(EMYGDIO; ALMEIDA, 2019). The OntoNeo ontology covers its content until prepuberty, as shown in Figure 9. Therefore, this study intends to act only with information from a pregnancy thousand days. This study intends to act only with information from a pregnancy thousand days.

The OntONeo design and development are guided by OBO Foundry principles, which seek to create a set of interoperable ontologies for describing biological and biomedical reality. It is being developed incrementally and iteratively over time, with the scope of each iteration predetermined. Each new version of the ontology adds new entities and relationships. The following Table 1 summarizes the current ontology metrics available.

According to the authors of OntONeo ontology (FERNANDA, 2018), the graph may not be applicable in other much different contexts because they were developed using examples from specific EHRs. On the other hand, the ontology content is rich and focused on the domain entity's representation of a real-life EHR.

2.6 Final Remarks

In the previous sections, we presented the theoretical basis for the technical principles involved in the HoPE architecture. We hope that these concepts have elucidated the proposed architecture's contents. Over the years, these and other concepts have gained strength in architectures for conversational agents. The following section aims to bring an overview of studies focused on architectures, modes of interaction, main objectives, contexts, and challenges related to conversational agents in the field of health.

3 RELATED WORK

In this chapter we present a review focused on conversational agents in health, carried out with the aim of understanding this topic in general through studies published in this field.

3.1 Research Questions

In this section, we seek to describe general and specific questions to guide this research.

General questions:

GQ1 What is the taxonomy for conversational agents in health?

GQ2 What is the state of the art related to conversational agents in health?

GQ3 What are the challenges related to conversational agents in health?

Specific Questions:

SQ4 What are the main environment contexts where conversational agents in health act?

SQ5 What are the main dialogue components used by conversational agents in health?

SQ6 In terms of architecture, what are the main systems and techniques used?

The main concerns and topics related to conversational agents in health are highlighted through the questions above. First, we considered three questions as being general owing to their comprehensiveness. GQ1 concerns the main definitions and classifications of conversational agents in health to assist in generating a taxonomy. Second, GQ2 is a classification of relevant articles considered by our criteria as state of the art. Last, GQ3 is related to challenges and open questions involving conversational agents in health.We also listed three specific questions related to conversational agents in health, proposing discussions related to particular topics within the central questions. SQ4 seeks to identify the main goals, domains and contexts related to conversational agents in health. The focus of SQ5 is on main dialogue components used by conversation agents in the health field. Finally, SQ6 seeks to summarize the systems and techniques used in conversational agent architectures in the health field.

3.2 Search Strategy

The initial process of this research applied a research query to academic and scientific databases to help answer our questions. The process of query construction was based initially on the authors' previous experience on the subject, correlating known terms such as synonyms, acronyms, and words inserted in the same context. Figure 10 presents the research string used to search the selected databases.

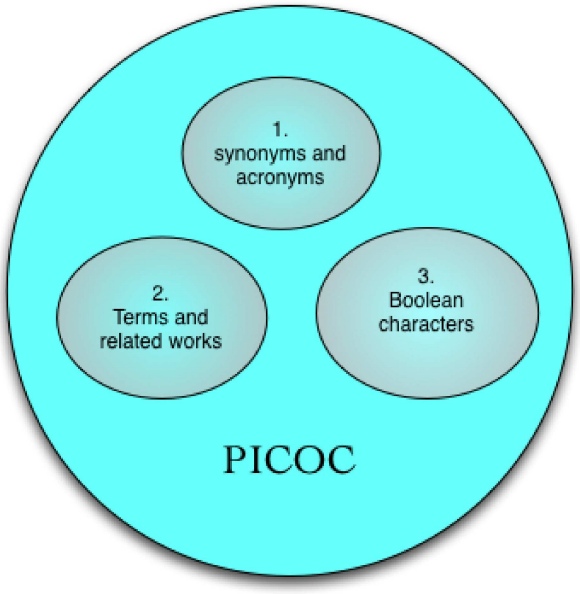
We used the PICOC methodology (population, intervention, comparison, outcome, and context) proposed by (PETTICREW; ROBERTS, 2008) to refine our research string. Figure 11 demonstrates the steps taken to refine the research string by following the PICOC method.

Figure 10: Search string used for database queries

Search String
String: ("Conversational agents" OR "Chatbots" OR "Embodied conversational agent" OR "Relational agent" OR "Intelligent Virtual Agents") AND ("Health" OR "Healthcare") AND ("Hospital")

Source: elaborated by the author

Figure 11: PICOC Method Steps for Literature Review



Source: elaborated by the author

3.3 Article Selection

The article selection process used filtering methods to improve the results that fit our main objectives. The process follows these steps: removal of duplicates, application of exclusion criteria, removal of impurities, and abstract and text filtering. In this context, we initially removed duplicate articles present in various databases. Once we found all related articles, we removed the studies that did not address our criteria. For this purpose, we used the terms of population and intervention derived from the PICOC method as follows:

- Exclusion criteria 1: Articles do not address "Conversational Agents" and related acronyms (population criterion I)
- Exclusion criteria 2: Articles do not address "Health," "Healthcare," or similar words (intervention criterion II)

The next step focused on impurity removal, deleting theses, thesis, and books, focusing only on scientific articles from journals and conferences. We also removed articles that were three pages or less in length.

The final process applied filters to abstracts and full texts in order to select the ideal corpus. We applied a text filter, selecting only articles with similar content to the main topics selected for this article, reading all texts and focusing on an article proposals, methods, and architectures. As suggested by (ZAVERI et al., 2016), three reviewers analyzed all the studies, verifying aspects important to our selection, such as interactions, dialogues, and architecture concepts. Moreover, these reviewers compared their article choices and based on mutual agreement, selected a final list of articles.

3.4 Quality Assessment

The corpus quality was a concern of this research. As a prerequisite, we verified the quality of selected articles through the presence of the following characteristics: the research proposal, context, literature review, related work, methods, results, conclusion, and future work.

We conducted a quality analysis of conferences and journals using the h-index evaluation metric, which is a single-number criterion assessing the productivity and the impact of a scientific journal and conference. Specifically, we used the h5-index, as calculated by Google Scholar [1]. The h5-index is the h-index for articles published in the last 5 complete years for a specific journal or conference. It is considered the largest number h, such that h articles published in 2008-2021 have at least h citations each. For instance, an h5-index is 10 when in the last 5 years 10 published articles have at least 10 citations each one (GOOGLE, 2019). As a criteria, we delimited a h5-index score (equal or higher than 5) to an article be accepted in

[1] <http://scholar.google.com>

Table 2: Quality Assessment of article structure and related questions.

Section	Description	Research Questions
Open Content		
	Title	GQ1, GQ3, SQ4, SQ6
	Abstract	All questions
	Keywords	GQ1, SQ5, SQ6
Article Content		
	Introduction	All questions
	Method	All questions
	Results	All questions
	Discussion	All questions
	Conclusion	All questions

this review. Additionally, we included two more indexes, the SJR and best quartile [2] for journal publications and the CORE [3] for articles in conferences, aiming at expanding the quality assessment.However, some conferences and journals did not have a classification in these indexes. One major factor for that is due to some journal and conferences being newer, sometimes with few editions. We opted to maintain in the survey articles of these publications that attain the minimum h5-index score, because they were significant to our discussion and many times the publication was related to the field of conversational agents.

3.5 Data Extraction

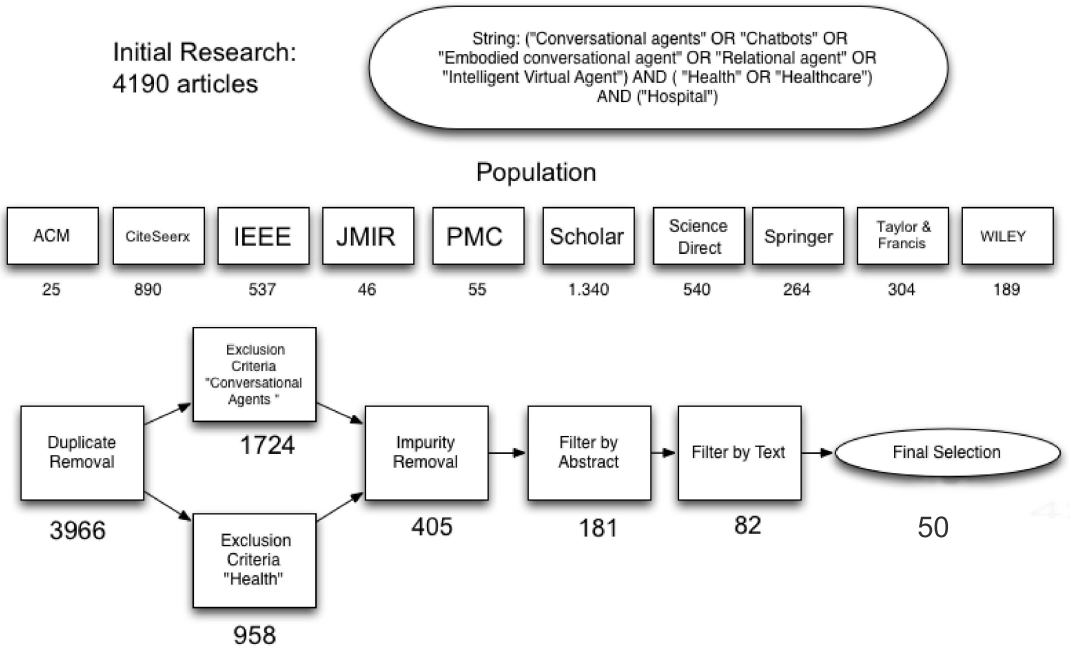
This section consisted of analyzing the relationship between the research questions proposed and the selected studies, verifying which articles could answer each question. Correlations of the questions according to the article contents are listed in Table 2. Thereby, this method improves focus on articles that answer our questions.

3.6 Recruitment

We answered our questions in separate sections based on 40 articles remaining after filtration. We seek to contribute through discussions on the field of conversational health agents, highlighting the main challenges, models, techniques, contexts, and current goals of this subject in recent research.

[2] <http://www.scimagojr.com>
[3] <http://portal.core.edu.au>

Figure 12: Article selection process



Source: elaborated by the author

3.7 Conducting the Search Strategy

We chose ten databases (ACM, CiteSeerx, IEEE, JMIR, PMC, Scholar, Science, Springer, Taylor e Francis, and Wiley) to evaluate articles for this research. Our criterion was the relevance of these databases regarding health literature. The process of selection of articles in each database used the procedures mentioned previously to select studies published over the past 10 years. Finally, we opted to exclude patents and citations, and only selected articles published in English.

3.8 Article Selection

Our search process is detailed in Figure 12, showing all processes used for exclusion and filtration.

Initially, our search string found 4190 articles in different databases. We first removed duplicate studies, resulting in 3966 studies remaining. We excluded 1724 studies that did not specifically address conversational agents , and 958 studies not related to "health" or "healthcare" were excluded, resulting in 405 articles remaining.

The next stage involved removing impurities such as theses, thesis, books, and articles of three pages or less. In this stage, the remaining 405 articles were filtered by abstract, which removed articles that did not address our subject, leaving 181 remaining studies. We then followed the guidelines by (ZAVERI et al., 2016) to further filter articles by abstract and full text. We the proceed with the quality assessment, based on our previously defined criteria. Following

these analysis, we further eliminate 2 articles that did not attain the minimum expected quality, resulting in a final corpus of 50 articles. We present the final corpus, divided in journal articles (Table 3) and articles published in conference proceedings (Table 4). The tables also show the indexes considered in the quality evaluation, h5-index, SJR and quartile for journal publications, and CORE for articles in conferences.

3.9 Data Extraction and Answers to the Research Questions

In this section, we discuss and answer the general and specific questions asked in this study.

GQ1 What is the taxonomy for conversational agents in health?

To obtain a better comprehension of conversational agents in health, we created a taxonomy, as shown in Figure 13. We organized the taxonomy in nodes, each one covering a central concept related to the area. The nodes are further detailed in terms of attributes, describing how the state of the art in the field addresses each of the characteristics of a conversational agent in health.

The primary purpose of this taxonomy is to categorize recurrent ideas by organizing concepts and connections. We organized the taxonomy characterizing an analysis of agent features and at the same time considers contexts and original proposals.

We have begun the building process with an in-depth analysis of all articles to identify characteristics, patterns, and categories of the corpus. Following, we defined three central concepts, defined as nodes in the taxonomy, to analyze (1) Environment, (2) Dialog modules, and (3) Architectures attributes found in the literature. We assume that these concerns enable us to identify the state-of-the-art healthcare conversational agents providing views from the technological, systemic and operational fields of these programs.

The first node, Environment, defines the contexts for using conversational agents in health, not focusing on the type of agent that acts in these contexts. The second node, **Dialog Modules**, defines the types of approaches applied in the articles selected and all models of communication used by agents in health care. The third node, Architectures, categorizes the techniques and systems used in the selected articles. Following each node, the taxonomy presents the corpus attributes related to each central concept. This detailing is presented later in the article, when we answer the specific questions.

GQ2 What is the state of the art related to conversational agents in health?

The selected articles that could pertain to the state of the art were analyzed by the number of citations and relevance of the studies from 2008 to 2018, being presented in Figure 14. They were classified according to H5-Index metric, from the highest to lowest score. We analyze the articles according to the number of citations they have received on the Google Scholar database. Moreover, we calculated the average of citations per year, considering the interval from the year of publication until 2018. This approach gives the reader an overview of the relevance of articles

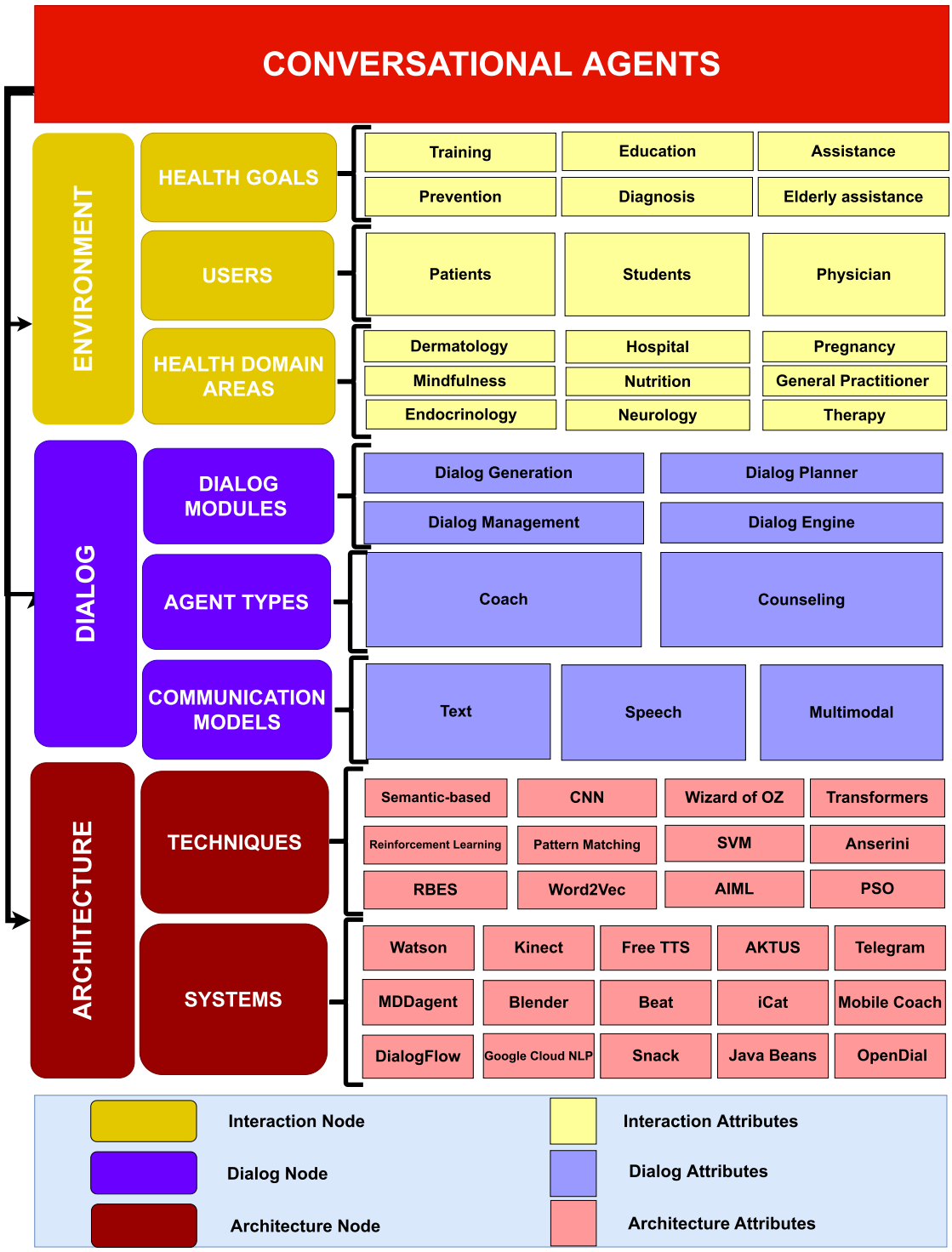
Table 3: Final corpus of articles published in journals.

Id	Article	Publisher	h5-Index	SJR	Quartile
A01	(ESTEVA et al., 2021)	Nature	414	19.469	Q1
A02	(TANAKA et al., 2017b)	IEEE	180	1.16	Q1
A03	(SEBASTIAN; RICHARDS, 2017)	Elsevier	110	1.55	Q1
A04	(D'ALFONSO et al., 2017)	Frontiers	84	1.04	Q1
A05	(HUDLICKA, 2013)	Elsevier	53	1.38	Q1
A06	(ZHANG; BICKMORE; PAASCHE-ORLOW, 2017)	Elsevier	53	1.38	Q1
A07	(BICKMORE; SCHULMAN; SIDNER, 2011)	Elsevier	50	1.03	Q1
A08	(SANKHAVARA, 2020)	Springer	44	0.145	Q1
A09	(KING et al., 2013)	Taylor-Francis	37	1	Q1
A10	(ALMARWI; GHURAB; AL-BALTAH, 2020)	Springer	35	1.03	Q1
A11	(TURUNEN et al., 2011)	Elsevier	33	0.54	Q2
A12	(BRESÓ et al., 2016)	Wiley	21	0.43	Q1
A13	(YASAVUR; LISETTI; RISHE, 2014)	Springer	20	0.36	Q2
A14	(RIZZO; KENNY; PARSONS, 2011)	JVRB	20	0.3	Q3
A15	(TANAKA et al., 2017a)	IEEE	17	0.48	Q2
A16	(TIELMAN et al., 2017)	IOSPress	17	0.28	Q3
A17	(HEERINK et al., 2010)	Springer	15	0.3	Q3
A18	(CHUAH et al., 2013)	MIT	13	0.23	Q3
A19	(KARPAGAM; SARADHA, 2014)	ICT	12	0.11	Q4
A20	(SHAKED, 2017)	IET	11	0.32	Q3
A21	(HIRANO et al., 2017)	SAGE	10	0.39	Q3
A22	(TRIVEDI et al., 2020)	Springer	9	0.23	Q3
A23	(EDWARDS et al., 2013)	Springer	9	0.23	Q3
A24	(WELLS et al., 2015)	PMC	7	0.21	Q3
A25	(CHUNG; CHO; PARK, 2021)	JMIR	34	0.1	Q3

Table 4: Final corpus of articles published in conferences.

Id	Selected Article	Publisher	h5-Index	CORE
A26	(YANG et al., 2019a)	ACL	157	A
A27	(GANHOTRA et al., 2020)	EMNLP	132	A
A28	(NIKITINA; CALLAIOLI; BAEZ, 2018)	ACM	74	C
A29	(MAGERKO et al., 2011)	AAAI	69	C
A30	(XIE et al., 2020)	ACM	61	A
A31	(KASINATHAN et al., 2017)	IEEE	34	-
A32	(PADAKI; DAI; CALLAN, 2020)	Springer	27	A
A33	(FADHIL et al., 2018)	ACM	22	-
A34	(LÓPEZ; EISMAN; CASTRO, 2008)	IEEE	18	B
A35	(RING; BICKMORE; PEDRELLI, 2016a)	MIT	17	A
A36	(ZHANG et al., 2014)	Springer	17	B
A37	(BICKMORE; PFEIFER; JACK, 2009)	ACM	17	C
A38	(BASKAR; LINDGREN, 2014)	Springer	16	C
A39	(JIN et al., 2017)	ACL	16	-
A40	(AMATO et al., 2017)	WAIAH	15	B
A41	(LISETTI et al., 2012)	FLAIRS	15	C
A42	(OCHS et al., 2017)	CASA	15	-
A43	(ALESANCO et al., 2017)	Springer	14	-
A44	(TOKUNAGA et al., 2016)	IEEE	12	C
A45	(HEERDEN; NTINGA; VILAKAZI, 2017)	IEEE	10	-
A46	(WARGNIER et al., 2016)	IEEE	10	-
A47	(LOKMAN; ZAIN, 2010)	Springer	10	-
A48	(RENTEA et al., 2012)	IEEE	9	-
A49	(KOWATSCH et al., 2017)	ETH	6	B
A50	(NI et al., 2017)	ISKE	5	B

Figure 13: Conversational Agent Taxonomy



Source: elaborated by the author

Figure 14: State of the Art articles considered in this survey with number of citations and average per year.

Source: elaborated by the author

considered in this survey.

The initial analyses sought the articles most cited in our sample. We identify two articles that have more citations than the others; they are (BICKMORE; PFEIFER; JACK, 2009) with 178 and (BICKMORE; SCHULMAN; SIDNER, 2011) with 77 citations based on Google Scholar.

In both articles, we identified high quality criteria, which means that in our assessment, these articles have all the attributes necessary to be considered of high-level quality in this review.

By analyzing the articles we observed that studies from 2008 and 2009, (the oldest studies present in the sample), have more citations than newer studies such as (D'ALFONSO et al., 2017), (TOKUNAGA et al., 2016), (HEERDEN; NTINGA; VILAKAZI, 2017),and (YASAVUR; LISETTI; RISHE, 2014), which do not have any citations to date. Moreover, other articles that have a significant number of citations in our sample are: ,(LISETTI et al., 2012),(HUDLICKA, 2013), (TURUNEN et al., 2011), and (CHUAH et al., 2013),with more than 20 citations each. We can observe that (BICKMORE; PFEIFER; JACK, 2009) and (BICKMORE; SCHULMAN; SIDNER, 2011) have a significant number of citations per year, being considered relevant studies. As relevant contributions of these articles we can point out the integration of ontologies to EHR (Eletronic Health Records), as well as the use of conversational agents in multimodal architecture for Health Literacy. We can also point out that very new articles, such as those published in the last years, tend to have fewer citations than older articles, regardless of their quality. So, they could impact the field in the near future. When using our research string without limiting the number of years, we analyzed other relevant articles in this field using Google Scholar.The article by (KING et al., 2013) presents a study focused on the role and efficacy of a culturally and linguistically adapted virtual advisor that provides tailored physical activity advice and support was tested in low-income older adults.The main results proving that these tools may reduce in reducing health disparities by ensuring e-health opportunities. Conversational agent sometimes is physically projected in the health field to show as a real person, mainly when used to train doctors or health professionals (CHUAH et al., 2013).

GQ3 What are the challenges related to conversational agents in health?

The last general question focuses on conversational agent challenges in the health context. In this sense, we sought to pose questions that were not answered in the corpus and could indicate future research directions. In this context, the issues presented sought to resolve aspects of various natures, such as method, meaning, technology and application. We summarized and clustered the results into five groups to facilitate interpretation, as shown in Table 5. Supporting this question, we discuss some challenges raised by selected articles related to conversational agents in health, focusing on the macro challenges and dividing them into a few groups.

The first group (G1) is related to concerns regarding the generation and understanding of dialog by agents as open questions proposed to be solved in future work, as well as new natural language processing algorithms for comprehension of a user's message. The second group (G2) is related to integration of conversational agents with other technologies. Additional challenges

Table 5: Challenges pertaining to conversational agents in health care.

Group	Challenges	Article Id
G1	Dialog Generation	A05, A13, A18, A34, A35, A38, A45, A46, A50.
G2	Integration with other technology	A25, A26, A36, A37, A40, A48.
G3	Elderly people	A09, A20, A28, A44.
G4	User experience	A02, A03, A04, A14, A19, A29, A31, A33, A41, A42, A43, A49.
G5	New approaches	A01, A06, A07,A08, A10, A11, A12, A15, A16, A17, A21, A22, A23, A24, A27, A30, A32, A39, A47.

are related to the adaptation and use of conversational agents for elderly people in our third group (G3). Challenges related to user experience regarding interactions and new interfaces are topics in group four (G4), referring to challenges in user context and user experience. Group five (G5) is related to new approaches involving methodologies and architectures.Technical challenges are quite common when discussing health agents. Concerns involving the generation and context of agents (LÓPEZ; EISMAN; CASTRO, 2008; HUDLICKA, 2013; JIN et al., 2017) as well as new natural language processing algorithms for comprehension of the user's message (WARGNIER et al., 2016; RING; BICKMORE; PEDRELLI, 2016a) are common themes. Natural language generation modules should be improved to work with ambiguities and varieties in the context of the conversation (LÓPEZ; EISMAN; CASTRO, 2008). Generating text and speech response in web environments can minimize patients' costs of access to agents (HUDLICKA, 2013). Improvements in voice recognition systems in hospital environments are part of (WARGNIER et al., 2016) work, as well as providing more elaborate responses. The classification and identification of depression topics in user messages are also part of future work in the health area (RING; BICKMORE; PEDRELLI, 2016a).

The second group is related to integration of agents with other technologies, such as electronic health records (EHRs) and medical documents, as well as integration with sensors and knowledge representation bases (BICKMORE; PFEIFER; JACK, 2009; RENTEA et al., 2012). Hospital records should serve as a basis for agent consultations, rather than being in separate stations. The use of EHR jointly to assistants could help in prevent diseases in the future, providing relevant information to the patient (BICKMORE; PFEIFER; JACK, 2009). Questions related to the adaptation and use of conversational agents for the elderly are very pertinent. In this scenario, the agents could act as reminders to assist elderly people with mental health problems (NIKITINA; CALLAIOLI; BAEZ, 2018). Furthermore, studies to improve the quality of care based on the reactions of elderly people to agents are of interest (TOKUNAGA et al., 2016). Multimodal architectures based on speech interactions for older people are another focus to future research. (SHAKED, 2017). Challenges related to user experience regarding interactions and new interfaces were discussed in several articles: (ALESANCO et al., 2017; KARPAGAM;

SARADHA, 2014; RIZZO; KENNY; PARSONS, 2011; KOWATSCH et al., 2017; LISETTI et al., 2012; MAGERKO et al., 2011; TANAKA et al., 2017b; KASINATHAN et al., 2017), which discussed improvements in communication techniques involving the user, using emotions in the dialog, and designing interfaces to make the conversation more realistic. 3D interfaces to the user interaction should be more present, offering multi-modality to the agent (KARPAGAM; SARADHA, 2014). Face-to-face interfaces will also be explored in future work to help in the obesity treatment (KOWATSCH et al., 2017). Moreover, new approaches to methodologies and architectures are considered as challenges. Modeling interactions between human-agent, exploring new hospital contexts (ZHANG; BICKMORE; PAASCHE-ORLOW, 2017), and new imitation methods (LOKMAN; ZAIN, 2010) are some future challenges noted in the corpus. Recommendation systems may serve as a tool to detect dementia in an easier way (TANAKA et al., 2017a).

SQ4 What are the main environment contexts where conversational agents in health act?

We investigated all article goals, domains, and contexts involving conversational agents in health. In this sense, Table 6 clarifies our findings.Training goals refers to coaching agents focused on enabling students to perform future tasks and consider new situations or assisting physicians to improve in their daily practices. The aim of the first group is to provide training on real situations that medical students or health professionals may face in their daily lives (HUDLICKA, 2013; RIZZO; KENNY; PARSONS, 2011).

In this sense, the agents can simulate patients in environments such as hospitals or facing emergencies where health professionals must make critical decisions (LÓPEZ; EISMAN; CASTRO, 2008; MAGERKO et al., 2011). Education agents are focused on teach health professionals and patients to understand health concepts sometimes referred as a tool for alphabetizing patients (BICKMORE; PFEIFER; JACK, 2009). Some tools are based on mobile devices, focusing on doubts and tips aimed at achieving better patient health (TIELMAN et al., 2017) or as a tool for teaching medical students in specific pathology's (CHUAH et al., 2013; JIN et al., 2017). Educators agents may deliver for moms breastfeeding counselors, providing an interpersonal continuity of care (ZHANG et al., 2014). The focus of the third group is on prevention, assisting in improving the relationships between patients and physicians (RENTEA et al., 2012). Agents can be useful to help patients in choosing the most proper disease prevention pathway (LOKMAN; ZAIN, 2010) or in suicide prevention or coping with depression (BRESÓ et al., 2016). Application for preventive care with mental health is shown in the work of (HIRANO et al., 2017). The conversational agent assistants have the characteristic of supporting physicians and patients in performing health-related daily tasks (ZHANG; BICKMORE; PAASCHE-ORLOW, 2017; RING; BICKMORE; PEDRELLI, 2016b). In the case of patients, they can assist in therapeutic and cognitive treatments (MCARTHUR et al., 2007; HEERDEN; NTINGA; VILAKAZI, 2017). Assistants could be tailored to mental health issues, helping in

Table 6: Environment interaction of Conversational agent.

Node	1º Level	2º Level	Article Id
Environment			
	Health Goals		
		Assistance	A04, A06, A07, A08, A11, A13, A19, A21, A24, A25, A26, A29, A34, A35, A44.
		Train	A01,A02, A05, A14, A22, A23, A27, A30, A28, A31, A33, A40, A41, A48.
		Elderly	A17, A20, A42, A45.
		Diagnosis	A15, A37, A46, A49.
		Education	A03, A16, A18, A32, A36, A38, A43.
		Prevention	A09, A10, A12, A39, A47, A50.
	Users		
		Patient	A01, A02, A04, A05, A06, A07, A08, A09, A10, A11, A12, A13, A15, A16, A17, A20, A21, A24,A25, A26, A27, A28, A31, A33, A35, A36, A37, A38, A41, A43, A44, A45, A46, A47, A48, A49.
		Physician	A14, A19, A23, A30, A32, A39, A40, A42, A50.
		Student	A03, A18, A29, A34.
	Health Domain Areas		
		Dermatology	A43, A14.
		Pregnancy	A23, A25
		Hospital	A01, A02, A08, A10, A21, A28,A29, A36, A42, A44.
		Therapy	A03, A04, A06, A17, A34, A40, A47.
		Neurology	A24, A31, A33, A50.
		Mindfulness	A05.
		Endocrinology	A09,A22, A26, A46.
		Nutrition	A07, A11, A48.
		G.Practitioner	A12, A13, A15, A16, A18, A19, A20, A27, A30, A32, A35, A37, A38, A39, A41, A45, A49.

recovery (D'ALFONSO et al., 2017). The fifth group focuses on diagnosis, in which the agent helps patients and physicians to predict diseases from symptoms or behaviors. These conversational agents could work in a clinical environment or on a patient's mobile device (ALESANCO et al., 2017; KASINATHAN et al., 2017). The combination of natural language processing capability with knowledge-driven could provide a diagnostic capability helping doctors to make decisions (NI et al., 2017). The health care field presents many studies related to assistants focused on elderly users (TURUNEN et al., 2011), and the sixth group focuses on this task. Some studies seek to discover a useful relationship between agents' conversational interfaces and older users (HEERINK et al., 2010), and along similar lines, different types of avatars and integration's of agents with elderly people have been researched(TOKUNAGA et al., 2016). Counselors for physical activity presented satisfactory results, increasing the health quality to the elderly people, seems a good alternative to being more explored in the future (KING et al., 2013).

We attempted to identify within which contexts conversational agents act in the health field. In this sense, we identified three types of context in which they can act: patients, physicians ,and students. Patient-centered applications have been developed, including agents focused on delivering diagnosis (HEERDEN; NTINGA; VILAKAZI, 2017) and in assisting patients in understanding diagnostic methods and results (BICKMORE; PFEIFER; JACK, 2009). Agents may focus on the treatment of diseases such as mental health, therapies, and skin tags via smartphones or other devices (ALESANCO et al., 2017; D'ALFONSO et al., 2017; LISETTI et al., 2012; NIKITINA; CALLAIOLI; BAEZ, 2018). Conversational agents can also act in prevention with patients if they have access to patients' health information or agents may offer assistance in overcoming physical inactivity and obesity (TURUNEN et al., 2011; RENTEA et al., 2012). Agents can also assist in health literacy by helping patients to understand basic medical concepts (BICKMORE; PFEIFER; JACK, 2009). New approaches to chatbots helping patients in choosing the most proper disease prevention pathway by asking for different information (starting from a general level up to specific pathways questions) and to support the related prevention check-up and the final diagnosis (AMATO et al., 2017). In the doctor context, the agent can act in health care institutions or in an ubiquitous manner through a mobile device. Agents may perform various tasks such as a support tool in the detection of patients with dementia or to control obesity. In addition to training professionals in the health environment, they can act as personal assistants to doctors (KOWATSCH et al., 2017; KARPAGAM; SARADHA, 2014; RIZZO; KENNY; PARSONS, 2011). Agents in the student context may work with tutorials and practices involving medical education (MAGERKO et al., 2011; LÓPEZ; EISMAN; CASTRO, 2008). Moreover, students can be trained by agents, for example by posing real situations to students as problems related to anorexia nervosa (SEBASTIAN; RICHARDS, 2017) or by participating in experiments conducted by agents (CHUAH et al., 2013).

We also listed the most representative domains where conversational agents act, which resulted in a considerable variability of domains in which this technology can be used. Agents

have been developed to interact in many health areas, helping in hospitals and emergency rooms (CHUAH et al., 2013; MAGERKO et al., 2011), as well as serving for example as an auxiliary to assist in auto-care for depression (RING; BICKMORE; PEDRELLI, 2016b)and nutrition (TURUNEN et al., 2011). Guidelines for prescription medications for the dermatology are delivered to health professionals by the conversational agent (ALESANCO et al., 2017).

Some researches were conducted in the field of obstetric care using conversational agents for obstetric support (CHUNG; CHO; PARK, 2021). A Q&A knowledge database-based chatbot (Dr. Joy) was developed and tested for perinatal women, focused on obstetric and mental health care, using a text-mining technique and contextual usability testing (UT). Breastfeeding promotion is an example of a healthy habit that necessitates the application of multifaceted longitudinal technology. Newborn intervention should begin with providing pregnant women with information on the benefits of breastfeeding, encouraging them to begin immediately, and ensuring that they have access to all of the information they will need. Although mothers are generally well-informed about breastfeeding shortly after birth, there are times when must be aware of common problems, as well as the option to contact a breastfeeding consultant (EDWARDS et al., 2013).

SQ5 What are the main dialogue components used by conversational agents in health?

The primary objective behind this question is to identify and classify all dialogue components used in our corpus. The main components of conversational agent dialogues found in our review are presented in Table 7.

Dialog agents types and structures of communication were classified in this section. Regarding dialog modules, the most representative in our sample was dialog management, which is responsible for managing the state of the dialog during interactions between an agent and individual (LÓPEZ; EISMAN; CASTRO, 2008). The dialog engine parses and computes the input for a reply to the final user (AMATO et al., 2017). The dialog generation work generating replies to users through machine learning algorithms or statistics approaches (JIN et al., 2017). The dialog planner has of the main feature uses as a flexible plan so that the dialog-based interaction can be dynamically conducted based on the knowledge that the system has acquired about the user (LISETTI et al., 2013).

Moreover, we have identified the main types of agents present in health dialogues: coaching agents and counseling agents. Coaching agents are used in general to train physicians or improve their handling of daily tasks, and in some cases, training patients to learn a new routine (MAGERKO et al., 2011). Coaching agents can provide brief interventions to change the behavior of an individual, or attempt to influence an individual to take better actions in the future (YAGHOUBZADEH et al., 2013). Counselors can be used as tool trainers that support improved health care of patients suffering mainly from emotional health problems or by improving physician and nurse working conditions (HUDLICKA, 2013). Counselor agents are socially engaged, building alliances with their patients, and encourage the patient to assist in

Table 7: Conversational agent dialogue modules

Node	1º Level	2º Level	Article Id
Dialog			
	Dialog Modules		
		Dialog Generation	A14, A28, A31, A38, A39, A42.
		Dialog Planner	A07, A12, A16, A41, A45, A47.
		Dialog Engine	A03, A04, A06,A08,A09 A10, A17, A18, A26, A27, A33, A34, A35, A36, A40, A49, A50.
		Dialog Management	A01,A02, A05, A11, A13, A15, A19, A20, A21, A22, A23, A25, A24, A29, A30, A32, A37, A43, A44, A46, A48.
	Agent Types		
		Counseling	A04, A05, A06, A07, A8, A09, A10, A12, A13, A14, A16, A17, A18, A19, A20, A22, A23, A24, A25, A26, A27, A29, A35, A36, A37, A39, A40, A41, A42, A43, A45, A47, A48, A49, A50.
		Coach	A01, A02, A03, A05, A11, A15, A21,A25, A28,A30, A31,A32, A33, A38, A44, A46.
	Communication models		
		Multimodal	A03, A05, A06, A09, A12, A014, A15, A16, A18, A19, A20, A24, A28, A29, A31, A36, A37, A39, A41, A42, A46.
		Speech	A13, A17, A34, A35.
		Text	A01, A02, A04, A07, A08, A10, A22, A23, A25, A26, A27, A30, A32, A11, A21, A33, A38, A40, A43, A44, A45, A47, A48, A49, A50.

improving their own health (JOHNSON; LABORE; CHIU, 2004). We present in Table 7 the classification.

Furthermore, we present the main communication models found. Interaction by text can work through text messages or web platforms, and can create conversations similar to interactions in social networks(ALESANCO et al., 2017). Consequently, many studies are seeking to improve this type of interaction, such as (KOWATSCH et al., 2017), which aims to design interfaces for text messages used to specifically support interactions between patients and physicians. Text dialog is helpful when combined with social networks, as it is familiar to most users (HEERDEN; NTINGA; VILAKAZI, 2017). Text techniques study were emphasized in (FADHIL et al., 2018) where were compared different dialogue styles and plain of text with approaches using emojis in conversations. Also, a new dialog architecture proposal was presented to have a dialogue with a human agent on health-related topics, where each component performs a set of tasks for the purpose to enable the agent to be enrolled in a dialogue (BASKAR; LINDGREN, 2014).

Speech interactions can help elderly people in many situations as a more natural form of interaction (RING; BICKMORE; PEDRELLI, 2016a). Some studies apply speech methods in robots to provide a better interface for interactions with users (HEERINK et al., 2010).In another study by (LÓPEZ; EISMAN; CASTRO, 2008), speech interactions were used to train students in real situations driven by virtual patients acting as normal patients manifesting different symptoms, encouraging students to think of possible correlated diseases. Multimodal dialogues may include text and speech interactions mixed with other techniques. In our samples, multimodal architectures were dominant. Some researchers such as (TANAKA et al., 2017b) argue that multimodal interactions may be a useful approach to assist people to overcome social difficulties, improving their narrative skills, conversation skills, and social skills. The work by (WARGNIER et al., 2016) aimed to create a multimodal agent to support elderly people at a hospital using a text-to-speech method, converting the text message to voice using a specific engine.

This type of approach is beneficial to elderly people or people who have problems reading text messages. There are also multimodal dialogues which use different types of interactions, such as text, speech, hand gestures, and facial gestures in combination. This can be seen in the study by (HUDLICKA, 2013), which aims to create a personal connection with a user through verbal and non-verbal interactions. Facial and gesture models are considered highly beneficial to establish a level of realism with patients by using virtual characters. The intent is to apply meditation techniques after this system is trained using a set of pedagogical strategies to support all user questions and behaviors. In some cases, agents based on multimodal interactions are intended to improve understanding when users are not familiar with a specific technology (BICKMORE; PFEIFER; JACK, 2009). This type of model can also assist in training individuals in medical skills through gestures and facial interactions (RIZZO; KENNY; PARSONS, 2011).

SQ6 In terms of architecture, what are the main systems and techniques used?

To define the architecture construct, we divided this node into two sub-nodes characterized by the types of systems, and techniques. In particular, in this taxonomy node, not all 40 articles were classified, since some articles did not present the employed techniques or systems. The taxonomy also yielded various techniques to process natural languages, such as statistical methods and machine learning methods. All points discussed it is presented in Table 8.

In our study, the literature on conversational agents in health identified various approaches and structures. As far as techniques are concerned, they have been addressed in the most diverse ways, such as the intent recognition, natural language comprehension, natural language generation and, within the context examined in this thesis, information retrieval.

Sentence-BERT has been used to research information retrieval systems and conversational agents. In (ESTEVA et al., 2021), the CO-Search, a semantic search engine designed to manage complex inquiries about COVID-19 using Siamese-BERT and TF-IDF as encoders for paragraphs embeddings to perform the task. In (YANG et al., 2019a) an end-to-end model for a question and answer system integrating SQUAD to Anserini toolkit was shown. The system uses a package developed by Anserini to deliver information back from the agent architecture. A better performance was shown by the comparison against the benchmark for this task. The unsupervised approach of this study achieved better results than in studies of similar models when comparing the correlation of embedding and probability of responses to queries. In the study of (YANG et al., 2019b), the BERT model with data augmentation strategy was used to fine-tune many different data sets vocabularies to achieve better performance results. The experimental results show significant improvements in effectiveness over previous approaches on English QA datasets, and we establish new baselines on two recent Chinese QA datasets. New data augmentation strategies dynamically annotate paragraphs as positive or negative instances to accompany training data, combined to BERT fine-tune. This research provides evidence that two English and two Chinese QA datasets can do well together (XIE et al., 2020). The BERT model is also used in conjunction with conversational agents to evaluate document prediction tasks, which involve a new set of public data (GANHOTRA et al., 2020). It has become increasingly common in the information retrieval field to research reformulation user queries and the model development and hybrid architectures based on this approach. The difficulty of locating appropriate documents for query expansion is well-known in the information retrieval field and was discussed in (SANKHAVARA, 2020), which presents a novel method for identifying appropriate documents for query expansion in biomedical document retrieval. The proposed approach requires minimal human intervention to identify relevant feedback documents and attempts to understand the relationship between query and answers in terms of document usefulness for query extension. In (ALMARWI; GHURAB; AL-BALTAH, 2020) is presented a hybrid approach to query expansion that combines statistical and semantic approaches. The study offers an effective weighting method based on particle swarm optimization (PSO) for selecting the ideal phrases for query expansion. BERT accuracy is significantly greater when dealing with

Table 8: Conversational agent Architecture.

Node	1º Level	2º Level	Article Id
Architectures			
	Techniques		
		Transformers	A01, A22, A23, A25
		PSO	A30
		Anserini	A32
		Wizard of Oz	A16, A17, A18, A20, A24, A33, A46.
		Reinforcement Learning	A13.
		RBES	A31.
		CNN	A39.
		Pattern Matching	A04, A12, A14, A21, A35, A37, A40, A42, A43, A45, A49.
		SVM	A15.
		Semantic-based	A05, A07,A08, A10, A11, A28, A34, A38, A47, A48.
		Word2Vec	A50.
		AIML	A19.
	Systems		
		Free TTS	A13, A19, A41.
		Beat	A06, A36.
		Pocket Sphynx	A35.
		Sonic	A14.
		Mobile Coach	A49.
		Kinect	A46.
		Dialog Flow	A43, A45.
		Watson	A40.
		Blender	A03.
		ACKTUS	A38.
		MDDagent	A02.
		Snack	A15.
		iCat	A17.
		Google C. NLP	A28.
		Java Beans	A29.
		Telegram	A33.
		OpenDial	A42.

lengthy natural language questions, demonstrating BERT's ability to extract valuable information from complex inquiries. In (PADAKI; DAI; CALLAN, 2020), query expansion is used to generate improved queries for BERT-based rankers and exhibited outstanding experimental performance for short and keyword questions. Ontologies can also help with information retrieval queries. In (TRIVEDI et al., 2020), we investigate the feasibility and accuracy of extracting a wide range of clinical concepts from free-text clinical charts using a query in a commercial natural language processing engine in a named entity recognition and normalization task.

Techniques such as convolutions neural networks (CNNs) have been applied in the health context to assist conversational agents in tasks such as identifying users' intent from their questions, interpreting their inquiries about the provided results (JIN et al., 2017). The identification of intentions was also the object of study in (FERNÁNDEZ-MARTINEZ et al., 2021), which sought to investigate the ability to detect multiple intentions using word-embeddings and recurrent neural networks.

The Wizard of Oz technique is an efficient way to examine user interaction with computers and allows rapid iterative development of dialog wording and logic. Some experiments have used this method for attention monitoring performance of a Louise agent and for interactive developments tests (WARGNIER et al., 2016). Approaches more complex, such as Markov chain and Reinforcement learning (RL), were used as a virtual counseling system, delivering brief alcohol health interventions by spoken dialogues interactions (YASAVUR; LISETTI; RISHE, 2014). Alternatively, Rule-Based Expert System model (RBES) were used by (KASINATHAN et al., 2017) to represent dialog patterns, receiving inputs from patients in the form of texts and extracting keywords to interpret meaning and to produce meaningful responses. Using inference, this approach provides a diagnosis of possible diseases based on the existing symptoms entered by users.

Moreover, almost all studies are inclined to use natural language processing using different types of models, such as bag-of-words. Especially in the health area, these techniques could be merged with the word embedding models, pre-training large dataset of medical documents. CNN models were used to pre-train the domain-specific word embedding from 5,958,529 short-text answers to the tagged dataset to calculate the class probability of the answers given (HASAN et al., 2016). Recently, in this field, some models have shown better results. The Bert model was used in a proposed architecture of a question-answer system providing an interaction of the physician with an extensive base of medicine via natural language. This model has presented excellent results (LIAO et al., 2020). Architecture uses an encoder-decoder model that takes the medical question-image pair as input and produces the answer as output. The encoder network consists of a pre-trained CNN model that extracts prominent features from a medical image and a pre-trained word embedding together with the LSTM to embed text data (ALLAOUZI; AHMED; BENAMROU, 2019). Additionally, medical documents can be represented by ontologies stored in OWL structures, permitting reasoning techniques to consult and extract implicit knowledge (BICKMORE; SCHULMAN; SIDNER, 2011).

Finally, the last sub-node includes the systems, which is related to frameworks and applications used to support the research of conversational agents in health. In this sense, we listed the most representative tools adopted by researchers in our corpus. Multimodal studies usually have a speech feature in their architectures mixed with other components. Pocket Sphinx is a speech recognition system used as a US-English acoustic model and dictionary to provide a grammar-based analysis tool. Facial systems such as Kinect are commonly used to classify expressions and assist in emotional analysis (WARGNIER et al., 2016). Text-to-speech is a recurrent task appearing in conversational agent studies that can also involve hand gestures, body posture shifts, gaze shifts, eyebrow movement, and head nods. In this sense, the BEAT system could help to interpret these gestures automatically, passing the translation to a text-to-speech tool (WELLS et al., 2015). APIs for text dialog, such as the Dialogflow system, could help in NLP and NLU tasks (ALESANCO et al., 2017).

As seen in this section, the work related to conversational health agents addresses a very distinct number of fields. Our architecture aims to work with health literacy, and therefore we look for areas that have disinformation as the most latent problem. Our choice was the field of pregnancy that needs a lot of reliable information (UGURLU; ORHAN, 2019) (POPOVA et al., 2019).

3.10 Final Remarks

This study discovered through a review of the health computing science literature that conversational agents applications in the health field are widely used, with numerous investigations conducted about this tool in recent years. The review found frustrations with the tool's relationship between expectation and performance, a lack of additional experiments with recent technologies in the conversational context, and some unexplored architectural gaps. We argue that more research on pregnant users is required to investigate and discover key aspects of the user-conversational agent relationship. Furthermore, we acknowledge the need for new studies and contributions in the context of conversational health agent architectures.

4 CONVERSATIONAL AGENT DEVELOPMENT

This chapter discusses the resources used to develop the conversational agent and the use cases for which the agent will be used.

4.1 Project Decision

This chapter will go over the resources used to create the conversational agent and the use cases for which the agent will be used. This section will go over the major decisions and constraints of the project. For the time being, the model only supports text chat interactions, but it will soon support voice chat. The contents cover 1,000 days period, including approximately 270 days of pregnancy and 730 days during the baby's first two years. As an implication of advising and answering questions, the contents cover breastfeeding, exams, nutrition, vaccinations, and exercises. The model aims to develop an architecture capable of offering appropriate answers from unstructured information in a user input (language understanding) and accessing structured and unstructured knowledge stores. A conversational agent will retrieve information using predefined responses, knowledge graphs, and transformers models to accomplish this. Additionally, this system is committed to understanding context and provides a framework for dealing with multiple intents.

4.2 Conversational Agent Framework

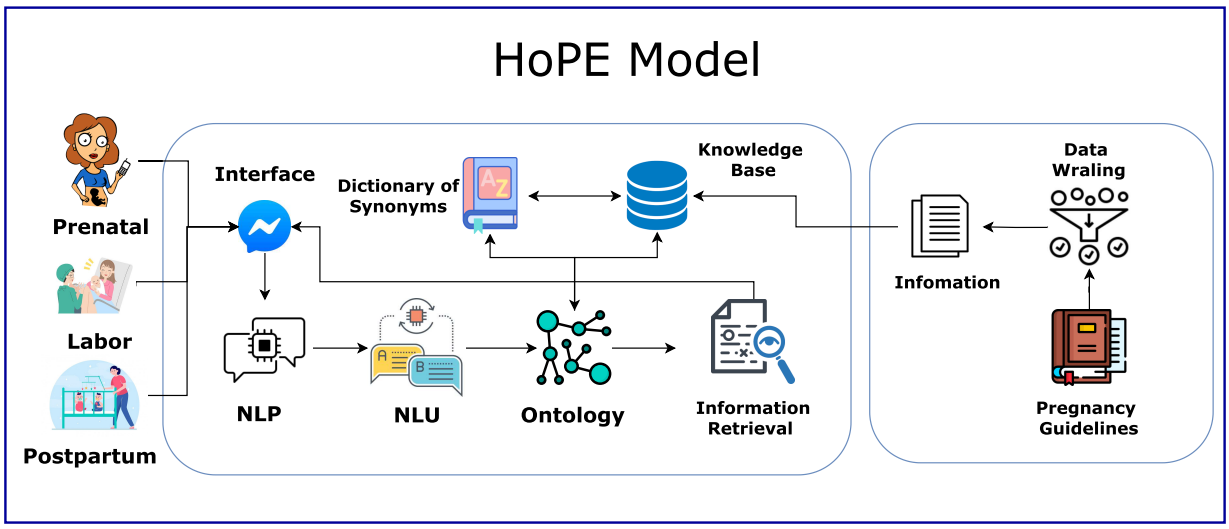
In this section, we presented the HoPE model architecture. HoPE is addressed to the needs of an audience that is constantly on the lookout for information. Pregnancy time can be critical for parents. The concept behind the framework is to provide access to official pregnancy guidelines. In addition, we intend to issue warnings during several critical periods. This proposal aims to combine semantic strategies to retrieve the most assertive information possible in response to pregnant women's questions.

The framework shown in Figure 15 is the structure that embraces all processes involving the HoPE model architecture. It operates on REST architecture and can be accessed by various interfaces. Their structure can be linked to chat systems such as Facebook messenger [1].This framework's components include the concepts of intention recognition, dialogue management, and information retrieval. The requests' output is typically *JSON* format, parsing to extract the required information.

The dialog structure proposed by HoPE uses a composite of predefined rules, NLP machine learning engines, and ontology-oriented dialogs. Rule-based strategy is responsible for basic input or output in our conversation agent. In addition, they are rigid structures that provide pragmatism in the conversation: greetings, goodbyes, initial explanations, agent feedback, and

[1] <https://www.messenger.com/>

Figure 15: HoPE general architecture



Source: elaborated by the author.

other items relevant to this type of structure.The use of buttons is one of the most used ways by rules-based chatbots, offering initial options to the user and proposing a continuity in the dialogs. And therefore they are also used (MELLADO-SILVA; FAÚNDEZ-UGALDE; LOBOS, 2020).

NLP machine learning engines are usually associated with predefined rules in platforms for developing chatbots. Its classic structure aims to use intents, entities, and context. An intent corresponds to an offline process in which the conversational agent is trained on example sentences related to that intent. This matching process is known as intention classification (BOONSTRA, 2021).

The output of this process is a score, in which the closest intent is retrieved. Intent classification can be supported by entities and by contexts. Entities aid in the correct identification of intent. They are defined with keywords of that intent and significantly help recognize the user's intent. Also, the conversational agent frequently relies on context to provide an effective response. Context is required to make the interaction feel more natural and understandable. The agent uses the intent configuration to maintain conversational coherence by establishing contextual inputs and outputs (BOONSTRA, 2021). Some of the challenges for developing conversational agents using only NLP engines include insufficient training, little data variability in training sets, difficulty with complex sentences, unforeseen contexts, and stagnation in the process of detecting user intent (PODGORNY; KHABURZANIYA; GEISLER, 2019)(ENGELMANN et al., 2021).

As a contribution, the HoPE model seeks to add components to improve the dialogue's conduct and precision. The dialogue manager module uses natural language understanding strategies and ontology to reach this aim. Tokenization, Part-Of-Speech, Normalization, and Stemming are the most common NLU strategies to process user input. These processes aid in

the capture of entities of interest that were previously unforeseen during the intent recognition stage.

The model proposed here aims to manage complex dialogs, and we do so by utilizing ontology. Its application to management tasks has been thoroughly investigated in (TEIXEIRA; MARAN; DRAGONI, 2021) (CHANG et al., 2019) (QUAMAR et al., 2020). The OntONeo ontology is composed of entities that correspond to domains in a pregnant woman's electronic health record. This structure incorporated predefined relationships defined by specialists and was also used to structure and store content from pregnancy guidelines in our model.

This process was also aided by using a terminology dictionary based on the content of these guidelines. These artifact aids in the ontology queries, acting as a refinement for searching for stored contents. The information retrieval module is the component in charge of responding to user queries. This module includes a neural network model trained on large datasets and capable of semi-supervised semantic searches. In this case, we use a Sentence-BERT model, which is cutting-edge for this type of task. We improved the model's capability for use in the HoPE model by increasing its understanding of data on the pregnant woman's health.

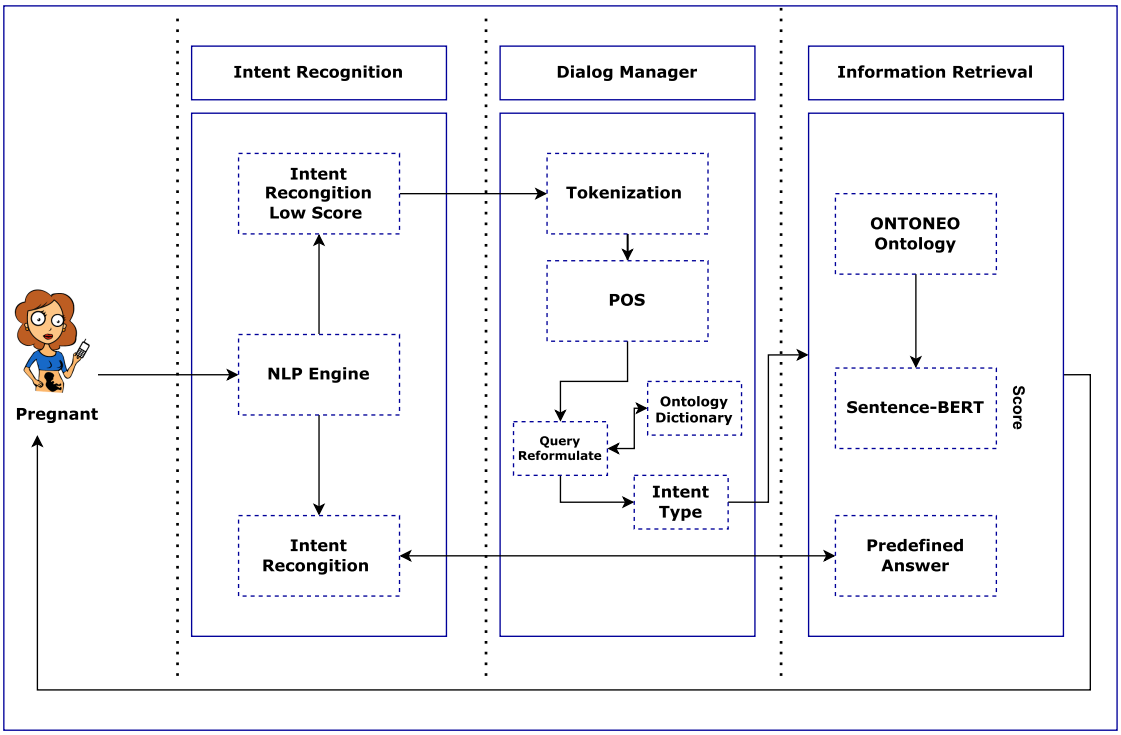
This study's data came from online health guidelines and protocols. The documents were compiled using data from the websites of the Brazilian government and health secretaries. Two gynecology professionals gathered and sent these materials. Ten of the documents were in pdf format for NLP processing, with the other two being digitized PDFs. The thousand days of pregnancy time was the focus of these materials. We determined that scientific articles and case studies were inadequate for our purposes. The whole process is presented in Figure 16

4.2.1 Intent Recognition

NLP engines are used to execute this step in the conversational agent's structure. User interactions are assigned a confidence rating based on user input (0-100%). The confidence is contingent upon the NLP model recovering a specific intent. Traditionally, chatbot systems have relied on a threshold to determine whether an intention is recognized. Classification can succeed or fail in this case due to two significant issues: precision (the agent rates the intention with high reliability but provides the incorrect answer) and recall (the agent does not recognize the intention with satisfactory reliability).

According to previous research (PODGORNY; KHABURZANIYA; GEISLER, 2019)(ENGELMANN et al., 2021), tool-based entities and intentions, in general, may be a proxy for difficulties with unmapped contexts, disambiguation, and decreased reliability when examples of distinct intentions overlap. The HoPE model proposes strategies for supplementing the conversational agent with a new NLU, disambiguation, and classification process to increase accuracy and recall coefficients. An overview of the HoPE dialog management module for the conversational agent architecture is provided in the next section.

Figure 16: Conversational Agent Architecture in Online Phase



Source: elaborated by the author

4.2.2 Dialog Manager

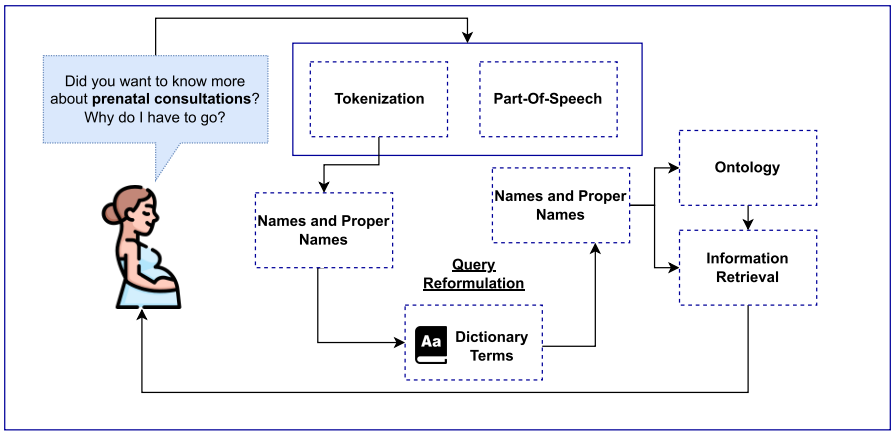
The dialog management module aims to orchestrate actions following the intention recognition phase. This module receives the previous module's intentions and entities. However, the intention recognition module may fail in some cases due to misunderstood intents or unknown entities. Another limitation is that the system can only understand sentences with 256 input characters maximum (IOVINE; NARDUCCI; SEMERARO, 2020).

Token extraction allows for query rewriting. Each entry in our dictionary comes with a synonym. If an entity already has a definition in the dictionary, that definition is kept. When new terms are introduced, we try to find (n...1) synonyms for the corresponding word. After this step, we check if any synonyms are present as an ontology term in the dictionary. If we find it, we replace it with a term from our ontology. Otherwise, the conversation is routed directly to the information retrieval module.

For instance, the word "physical exercise" exists in our corpus because it is frequently used in health guidelines. Therefore, we increased the variation of this term with synonyms such as gym, crossfit, and yoga. And to verify that the conversational agent recognizes the entity detected in the preceding module, it must first be validated against the entity dictionary.

If the term is not an entity in the corpus but is among the related synonyms, we replace it with the synonym. For instance, in the question "Can I do yoga before breastfeeding?" the dictionary of terms will rephrase the sentence to "Can I exercise before breastfeeding?". After this phase,

Figure 17: The dialog manager entity extraction and treatment process in HoPE architecture



we use the entities as input for ontology research. The strategy is to retrieve responses present in the Data Property of the ontology that contains these entities. Thus, a group of sentences is retrieved for the right domains, reducing the probability of false positives in the inference phase.

Finally, the retrieved sentences are vectorized and incorporated into the information retrieval models for the last phase of the HoPE model. If we do not find the correct entity, the sentence will go directly to the information retrieval module. A summary of the process is presented in Figure 17.

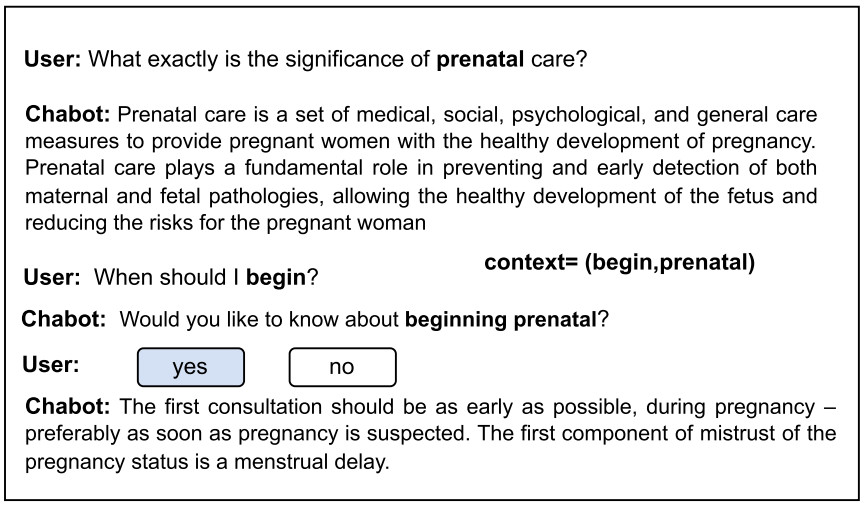
Some sentences did not have a key entity in their structure, as shown in Figure 18. Often, users provide imperative phrases that link to the preceding context. We applied a simple disambiguation strategy based on previous contexts within the framework proposed in this thesis to deal with this scenario. It is common for FAQ conversation agents to respond to users' questions about entities of interest (EOI).

All interactions are stored in a NoSQL database using the ID as a reference and the entities extracted in dialogues. The strategy initially identifies if the sentence does not have a defined subject. Then, we retrieve the preceding entity by utilizing the reasoning mechanism to discover related Classes and Data Properties. Finally, we provide options for confirmation regarding the subject of their interest related to that domain. This way, we can continue the conversation without breaking the flow and experience. If no prior context (EOI) exists for the question, we request that the user rephrase their sentence.

4.2.3 Information Retrieval

In this section, we present the information retrieval module. The architecture has a hybrid character, with more than one way to retrieve information. We use this model to capture and better understand the utterance meaning for questions with less reliability, seeking to retrieve the most assertive answer for users. The information retrieval system is divided into two steps

Figure 18: Disambiguation strategy using previous context.



Source: elaborated by the author

detailed in the following sections.

4.2.3.1 Intent compound detection

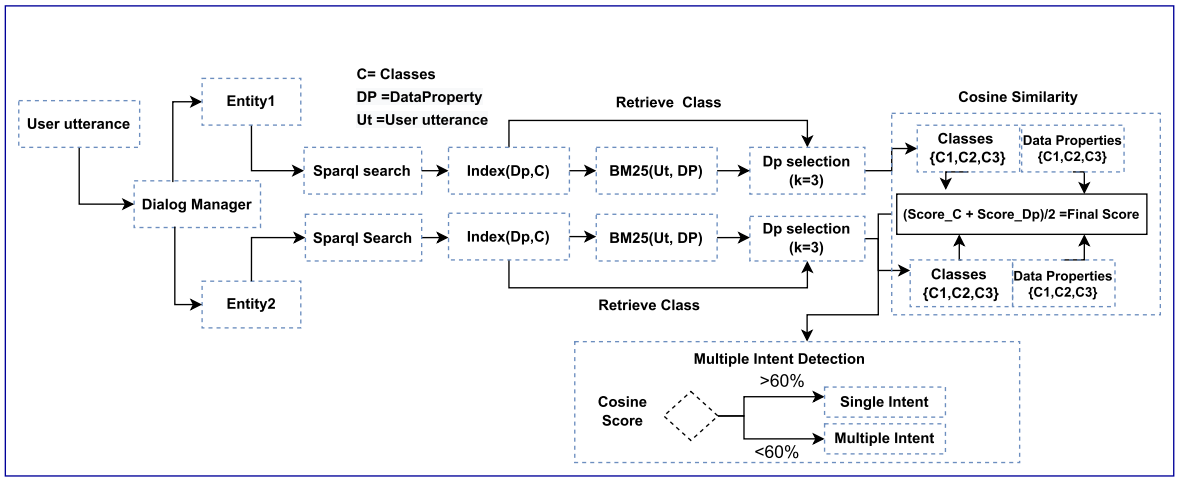
This section discusses how to detect compound intents through HoPE strategies. Frequently, certain user utterances contain one or more intents. The HoPE proposal seeks to use the relationship of domains to understand the sentence type. When the NLU process extracts more than one entity, it goes over with the dialog manager step. If only one entity is discovered, this module is not activated because it is assumed to be merely an intention detected.

A SPARQL query is run against the ontology to determine if this entity has any relationships. The objective of this module is to identify the relationship between two ontology classes. As with the dialog management stage, we guarantee that the classes are related, mapping previously extracted ontology terms. However, it is hard to determine whether the ontology contains any relationships without mapping the input entities, so the interaction goes to another module.

Due to key entities frequently pointing to the subject the user wishes to discuss, they are critical to understanding the type of relationship. The content extracted from health guidelines was stored as Data Properties annotations in the OntONeo ontology and retrieved through a keyword search. If an entity is in a sentence, it will be retrieved. The query returns a set $c=1...n$ of several possible answers, along with their domain classes, in the form of a triplet $i(a,c)$, where i represent the index, a denotes the retrieved Data Properties(responses), and c the respective classes(concept).This process is in Figure 19.

We reduced the number of possible answers using the BM25 Okapi lexical algorithm, which takes the user's sentence as input, removes stopwords, and re-ranking three responses that most closely approximate the user's input. We get the concepts C related to each response retrieved. Finally, we want to know how similar the classes derived from the input entities are. We used

Figure 19: HoPE modeling to identify multiple intentions in user interaction



Source: elaborated by the author

cosine similarity as a string comparison from concept names and Data Property annotation. The idea is to verify the similarity degree from domains and annotations referred to each entity.

The cosine similarity score of less than 60% indicates that the entities are from different domains and do not have a significant relationship, resulting in multiple intentions for our architecture. On the other hand, a score of more than 60% indicates a relationship between the terms, indicating a single intention classification.

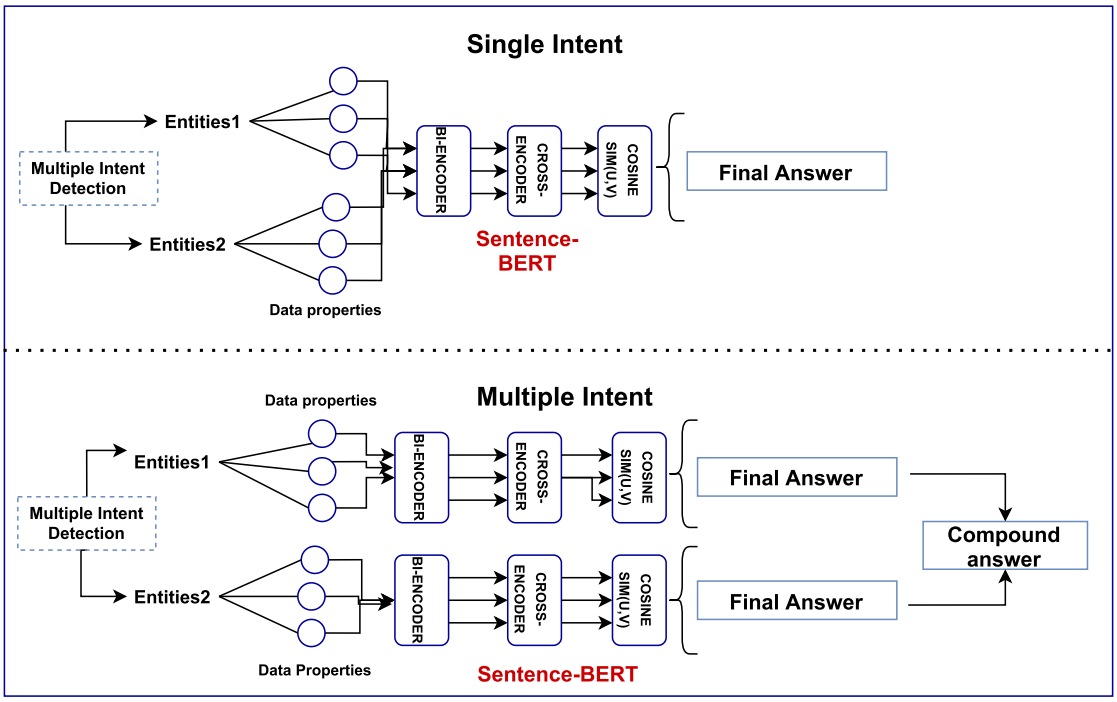
This step is skipped if the dialog management module recognizes only one entity or name, in which case the sentence is assumed to have a single intent. In the absence of an entity, the user is directed to the information retrieval module. In this case, the process of comprehending multiple intentions is deactivated.

4.2.3.2 Hybrid Information retrieval using Sentence-BERT

We present the recovery process for a user query in this section, shown in the process in Figure 20. If the previous module has identified a single intention, we join the retrieved sentences (Data Properties) in a list and index them in the pre-trained Sentence-Bert model. In the case of detecting two intentions, the model performs indexing sequentially for each set of sentences. The paragraph with the highest score is chosen to respond to user input. In the case of multiple intentions, the two retrieved sentences are concatenated and sent as sequential responses.

The retrieval process initiates in a bi-encoder network that receives user input and the set of possible responses from the ontology process. The model's output retrieves dense vectors from the documents closest to the user's input. However, bi-encoders do not have the best performance for this type of task, as they usually recover a lot of false positives. Therefore, we re-ranked the bi-encoder output using a cross-encoder model in which we scored the relevance of all candidates for the user's search query. Then, the sentence with the highest score is chosen

Figure 20: The HoPE model architecture for multiple intent detection and information retrieval



Source: elaborated by the author

to respond to user input.

This module can also be activated without ontology management when the entry sentence is "out-of-scope". Then, we respond to the user using the Sentence BERT model with pre-computed response embedding in clusters if no entity or name is found. In this case, we preload the representations of the paragraphs in indices and cluster them using the approximate nearest neighbor search (ANN). After recovery happens, we return the K=1 response with the highest score. In this way, the module works as a last attempt at information retrieval without going through the ontology module. The big difference here is that instead of using a supervised keyword search strategy in the ontology and returning the paragraphs referring to these terms, we use an unsupervised clustering strategy with groupings that the ANN strategy will perform.

The HoPE information retrieval module will present three scenarios: retrieve the answer with a relevant and correct score (greater than 65 percent), retrieve the answer with a relevant and incorrect score, or obtain an irrelevant score. In conversational agents, an irrelevant score is frequently defined as a fallback. A fallback value is less than a predefined threshold, indicating that we lack an adequate response to the question. This coefficient can be defined empirically via observational analyses of the experiments conducted and static data such as weighted averages and standard deviation.

4.3 Final Remarks

This chapter described the HoPE architecture for promoting health literacy among pregnant women via conversational agents. We have shown the functionalities and performance of the architecture's intention recognition, dialog management, and information retrieval modules. The findings and discussions related to the evaluations conducted using the HoPE architecture highlighted here are shown in the following section.

5 MATERIALS AND METHODS

This chapter discusses the proposed architecture evaluation and describes the process for conversational agent development. We divided it into two main sections: conversational agent model development and assessment. Firstly, we describe the whole process of HoPE model development, focused on steps that sustain the agent architecture.

We began by describing corpus processing, composed of knowledge derived from health guidelines. Additionally, we cover the ontology and dictionary procedures and the specifics of the Sentence-BERT training. The section on conversational agent evaluation was divided into two subsections: user evaluation and model evaluation. The user evaluation includes randomized studies with pregnant women and health professionals in 2020 and 2021. We seek to improve these two groups' understanding of conversational agents in health to assist pregnant women. The model's evaluation focused on experiments related to the HoPE model mechanisms and the proposed contributions. We seek to follow classical assessment approaches for information retrieval systems (SAMIMI; RAVANA, 2014). We use clinical studies and test collection as core assessments, among other methodologies. The experiments carried out are shown in Figure 21.

5.1 Materials

This section details the offline processes used to construct the HoPE model architecture. We discuss the collection and establishment of the knowledge base, the procedures for dictionary building, the ontology population, and the training of information retrieval models.

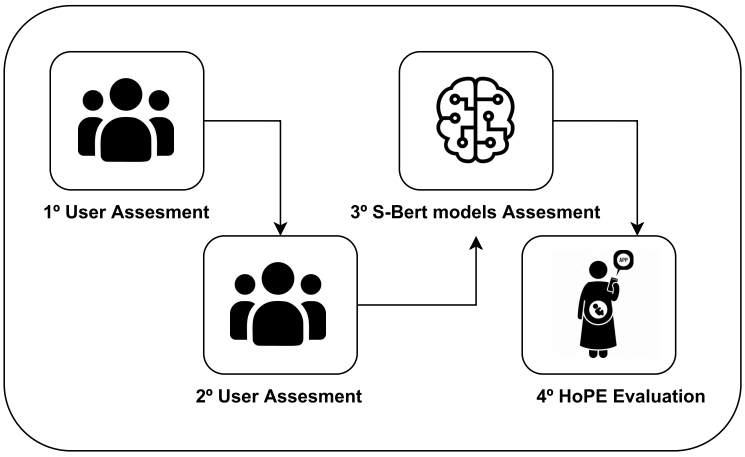
5.1.1 Corpus Construction

These documents were reviewed manually to ensure no line break adjustments or extraction failures occurred. Determined sentences retain the context of the preceding sentence, resulting in tokenized paragraphs. Uncommon sentences (with many spaces, special characters, and few words) were reviewed by humans. We used regular expressions to find this kind of sentence faster. The procedure resulted in the creation of a corpus containing 7.077 sentences. We used the content to build a knowledge base for conversational agents and fine-tune Sentence-BERT model training. Table 9 presents the distribution of the main topics present in the corpus.

The data organization is in a specific format for the evaluation phases. Figure 22 illustrates a comprehensive view of this process. We begin by organizing the domain's data in the following manner: (Q1, Q2, Score). The training process of Sentence-BERT networks with the data augmentation strategy does not need labeled data but rather organized data in pairs. When training Sentence-BERT networks, the data is structured in sentence pairs and not labeled.

We used unsupervised Sentence-BERT models to produce positive sentence pairs through

Figure 21: Flow of experiments carried out during 2019 and 2021.



Source: elaborated by the author.

Table 9: The percentages distribution by corpus topics

Topic	N
Gestational Symptom's	26%
Gestational Information	19 %
Nutrition	18 %
Illnesses	10 %
Vaccination	10 %
Gestational Risks	5%
Exams	4.5 %
Breastfeeding	3.5 %
Medication	3 %
Physical Exercises	1%

Source: elaborated by the author.

similarity. The objective of this model is to find similar pairs, which we empirically understood would be the minority.

We employ a semi-supervised annotation process (THAKUR et al., 2020)) to implement this organization. Using our dataset, we randomly select a sentence/paragraph from the set (X), retrieve the top 100 results using the cosine similarity distance, and then randomly select a sentence from the top 100 results (Y). The annotations (X, Y) are then applied using a score (Z) between 0 and 1. At the end of this process, we had 2.098 pairs with a similarity score of 1. We carry out more preprocessing steps such as duplicate pairs removal, human judgment, and class balancing. The removal of duplicated pairs was the removal of identical sentences (A, B) resulting from the pair generation procedure 1.500 pairs remained. We reviewed the pairs with a human judgment step to declare whether they were the same content or not. We identified about 325 paragraphs with similar content written in different ways. The authors and two medical researchers re-wrote 275 phrases to increase the positive examples. We focused on providing sentences with the same meaning but rewrote and of smaller size to adjust the model to deal with size asymmetries between short and long sentences. Previous studies (ALFEO; CIMINO; VAGLINI, 2021)(ZHANG et al., 2021)(KIM; YOO; LEE, 2021) pointed out that for better evaluation results, balancing examples was an important feature. Finally, we had 600 positive and 900 negative random pairs. This dataset is sufficient for fine-tuning the Sentence-BERT neural network(REIMERS; GUREVYCH, 2019) (ALFEO; CIMINO; VAGLINI, 2021).

5.1.2 NLU module

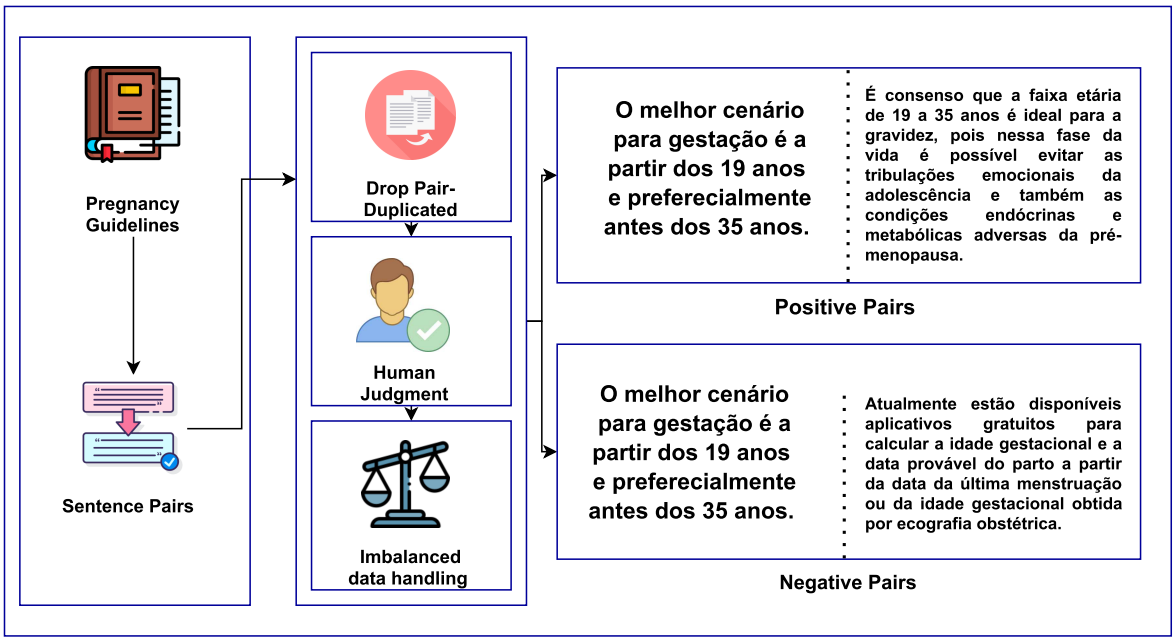
Training intents and entities are the first steps in the offline process (YAN; LIU; CHEN, 2011). The intention recognition module detects the user's intention, which is required for conversational agents to be successful. Classifiers frequently use labels to categorize data from various datasets. The training dataset is filtered to remove the incorrectly classified data, requiring retraining. Figure 23 explains how we structured the system of intents.

Greeting intents initiate or terminate a session with the user. In this system, actions are the type of intent responsible for identifying questions or desires of users. Complaint-type intentions address the user's unhappiness with the system's dialogue or information retrieval. In these cases, the chatbot provides feedback and excuses. We train a total of seven to ten utterances for each intention. Additionally, we have established at least one entity to assist with proper recognition. Entities populated in NLP Engine are intended to assist in recognizing intentions.

5.1.3 Dictionary of Terms

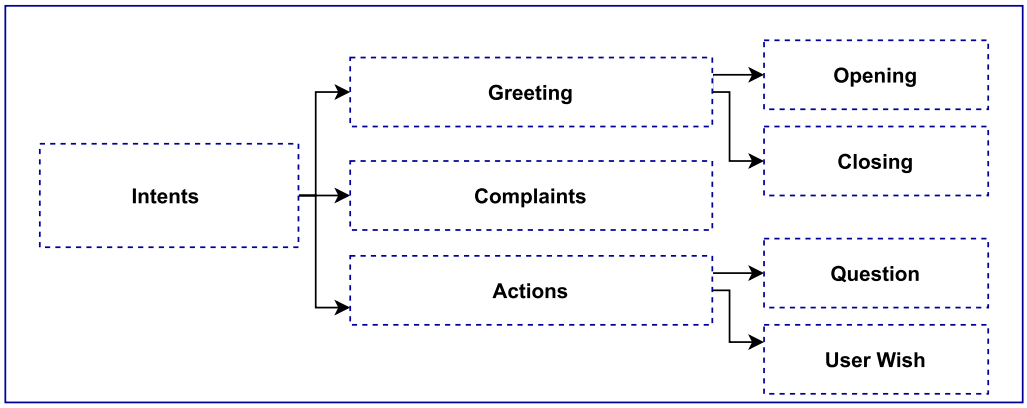
We created a query reformulation strategy through an dictionary of terms. We collected entities from each paragraph corpus and searched synonyms in the WordNetPT ontology for each

Figure 22: An example of positive and negative pairs present in our base. Positive pairs (Pregnancy is best starting at the 1 year and ending before the age of 35 years | The best scenario for pregnancy is from the age of 19 and preferably before the age of 35). Negative pairs (Pregnancy is best starting at 19 years and ending before the age of 35 years | Currently, free apps are available to calculate the gestational age and the probable date of delivery from the date of the last menstrual period or the gestational age obtained by obstetric ultrasound).



Source: elaborated by the author.

Figure 23: Predefined set of intents and actions for chatbot structure



Source: elaborated by the author.

term [1].In addition, the authors included synonyms not found in ontology (e.g., specific medical terms). Before being added to the dictionary, the synonyms are normalized and stemmed. The proposed strategy replaces the user's entities with existing dictionary terms. As we add words already known to the trained vocabulary, the probability of correct information retrieval increases.

5.1.4 Ontology Procedures

As a dialogue manager role, we use an ontology as a structure for the chatbot. The OntOneo is based on electronic medical records of pregnant women, and it contains concepts and relationships that are similar to the content extracted in this study. The offline process involved three steps: data property addition, relationship data properties creation, and individual relationship creation. First, the text passages are extracted and grouped by content similarity.

We use a manual process to separate this data. Then we add this data to the ontology structure as a data property. Each sentence is stored as a text *rdf:comment*, and a label to identify the text content. For instance: *rdf:comment* Avoid eating fried foods and bacon every day rdf:label (Can I eat bacon during pregnancy?).

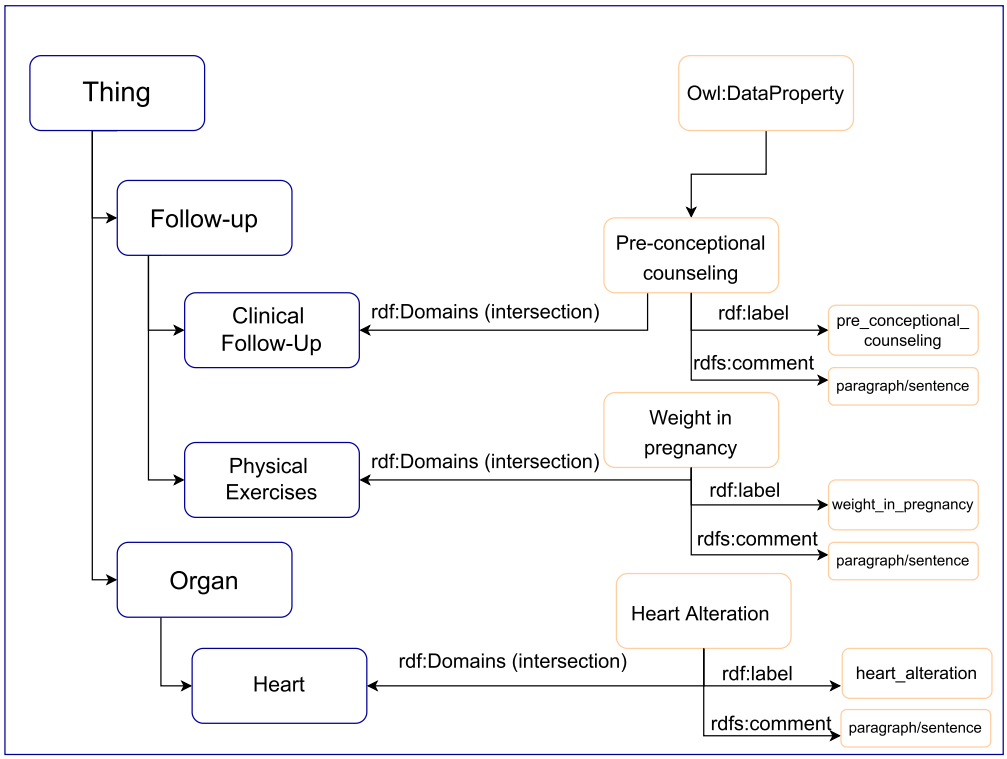
Each text was assigned to *(1..n)* ontology classes. Classes are concepts that represent different domains within the ontology. We related classes associated with each paragraph mapping by health professionals, who indicated which classes would relate to the sentences. The goal was to establish a connection between ontology entities (classes) and data properties. We shown this connection in Figure 24. We limit the number of classes per data property to four to maintain a balanced class count. The next step was to enrich these classes. We extract verbs from the responses and add them as ontology instances to the same classes with the data properties as before. Before adding these terms, we apply the normalization and stemming process to the data. As a result, we construct a network of links between classes (entities), instances, and data properties. We used Cellfie plugin [2] from the Protegé toolkit, to add new instances and data properties to the ontology.

5.1.5 Fine-Tune Procedures

In the HoPE architecture proposed in this work, the Sentence Bert or simply SBERT model is used to train an NLU system based on data from health pregnancy guidelines. This system aids in information retrieval by providing a method to compute dense vectors using state-of-the-art transformer networks. We use the traditional fine-tuning strategy and the data augmentation strategy proposed (REIMERS; GUREVYCH, 2019)for our experiments. The data augmentation strategy is the most appropriate due to the scarcity of labeled data.

[1]<https://github.com/recognai/spacy-wordnet>
[2]<https://github.com/protegeproject/cellfie-plugin>

Figure 24: Data properties and classes relationship



Source: elaborated by the author.

The data augmentation strategy is represented in 1, 2 3, algorithms. Training a cross-encoder layer on a benchmark dataset, labeling domain data with the cross-encoder, and training a bi-encoder on the labeled domain data are the two phases of data augmentation. The first training phase begins with a cross-encoder trained over a benchmark dataset. This dataset is also known as a "gold dataset". We chose the ASSIN2 benchmark used for Sentence Text Similarity (STS) tasks. The main idea is to use knowledge transfer to train a cross-encoder model. The ASSIN2 dataset is widely used in Brazil to evaluate supervised STS systems. There were 6,500 records in the training and validation data sets and 500 pairs of Brazilian Portuguese sentences annotated for inference and similarity tasks. The semantic similarity values ranged from 1 to 5, with entailment and none as the inference classes. The test dataset comprises approximately 3,000 pairs of sentences containing the same linguistic phenomena and notes as the training data. All data were manually recorded.

The second step is to label our dataset, referred to as the "silver dataset." As suggested by the strategy, we must normalize the pairs previously mentioned (between 0 and 1) for binary classification tasks. In this case, we use the newly trained cross-encoder in a benchmark to determine whether our data needs to be labeled. This is accomplished using the pre-trained SBERT model. The bi-encoder fine-tuning is responsible for conducting semantic research, which entails comprehending research content via lexical correspondence, context, and synonyms.

As a final result, we have re-ranked models and fine-tuned retrieval based on pregnancy guidelines. We used 256 Max Length for the tokenization layer and a MEAN-pooling strategy

Algorithm 1 Cross-Encoder training over STSb dataset

Input $BERT, STSb$
Output Cross-Encoder

procedure CROSS-ENCODER LOAD MODEL
    CrossEncoder(BERT)
procedure SPLIT BENCHMARK
    for x,y,z STSb.split(): **do**
        $x \leftarrow$ *Train*
        $y \leftarrow$ *Dev*
        $z \leftarrow$ *Test*
procedure CREATE DATA LOADER AND EVALUATION OBJECTS
    TrainDataloader(x)
    CECorrelationEvaluator(y)
procedure CROSS-ENCODER TRAINING
    Cross-Encoder.fit((x))

Algorithm 2 Cross-Encoder for labeling Pairs

Input $CrossEncoderPath, Train, PairDataset$ **Output** Cross-Encoder

procedure CROSS-ENCODER LOAD MODEL
    CrossEncoder(CrossEncoderPath)
procedure LABELING DATASET
    for x,y,z STSb.split(): **do**
        $x \leftarrow$ *Train*
    Cross-Encoder.predict(x)

Algorithm 3 Bi-Encoder Training

Input $LabeledDataset, SBERT$ **Output** Bi-Encoder

procedure BI-ENCODER LOAD MODEL
    qqTrain(LabeledDataset)
procedure CREATE DATA LOADER AND EVALUATION OBJECTS
    TrainDataloader(qqTrain)
    MultipleNegativesRankingLoss(S-BERT)
procedure BI-ENCODER TRAINING
    Bi-Encoder.fit(([TrainDataloader, MultipleNegativesRankingLoss]))

Table 10: Specific hyper-parameters for model training

Training Parameters						
	BERTimbau		BERT-Mult.		Paraphrase	
Parameters	Cross-Enc.	Bi-Enc.	Cross-Enc.	Bi-Enc.	Cross-Enc.	Bi-Enc.
Batch Size	16	12	16	8	16	12
Learning rate	2e-5	2e-5	2e-5	2e-5	2e-1	2e-1
Epochs	8	10	8	5	5	1

for all training performed. In addition, we run 1000 evaluation steps and use the ADAM optimizer for all models. These parameters are frequently used in studies that perform training in SentenceBERT networks (REIMERS; GUREVYCH, 2019).

Furthermore, we set two distinct loss functions: for cross-encoders TripletLoss and bi-encoders MultipleNegativesRankingLoss. The triplet loss algorithm tunes the network given an anchor sentence a, a positive sentence p, and a negative sentence n, such that the distance between a and p is less than the distance between a and n. MultipleNegativesRankingLoss was chosen as the bi-encoder function loss because it is widely considered to be the optimal function loss for training embeddings for retrieval setups containing positive pairs (e.g. (query, relevant doc)) (HENDERSON et al., 2017). The hyper-parameters that were customized for each specific model are listed in Table 10.

Batch sizes are determined based on the model's performance in the execution environment to maximize efficiency. The number of epochs used to train the cross-encoder, and bi-encoder layers were selected according to the model's performance. We modified this hyper-parameter exhaustively until finding the optimal coefficient without over-fitting. The learning rate for the Paraphrase model was decreased because it performed better at this rate.

5.2 Evaluation methods

This section discusses human user clinical studies and the conversational agent framework architecture evaluation.

5.2.1 User's evaluation

This section presents two clinical studies with pregnant women, health professionals, and physicians. The first study aimed to assess the usability perception of a chatbot trained with nutrition data (from pregnancy guidelines) from the perspective of pregnant women and health professionals. We aimed to contribute to these two groups' current perceptions, seeking to understand their opinions about the tool. The second experiment assesses users' and physicians' perceptions of a conversational agent trained in the whole content previously presented in Table

9. We contribute to the health area through the UTAUT2 adapted model application in the pregnant women context. Unlike the other study, this one aims to confront the data from qualitative and quantitative perceptions using a mixed-method analysis. We used the DialogFlow tool as an NLP engine for user interaction in clinical studies.

The approval for both projects was obtained from the Ethics Committee in Brazil through the number (3599678) and approved by the Municipal Health Department of Passo Fundo, where the experiments happened. We have defined some chatbot use cases for these experiments. They are as follows:

UC.1 **Greetings and Farewells**

This action can be started during the conversation but is often used at the beginning or end. It is not necessary to start or finish a conversation with a greeting or a farewell.

UC.2 **Concerns about prenatal and postnatal care in general**

In this scenario, the user inquires about a particular subject and answers from the conversational agent. For inquiries that are not within the scope of our content, the response may be correct, incorrect, or unanswered (in which case, we ask the individual to rephrase the question).

UC.3 **Acknowledgment or confirmation**

Typically, the acknowledgment and confirmation scenarios respond to the user's inquiry. In both instances, there are motives. Confirmation questions (yes, no, got it, did not get it) can be answered using the preceding context. As an appreciation token, we identified the key terms and expressed gratitude to the user for their participation.

UC.4 **Phrases without entities**

In the absence of entities, the conversational agent promotes continuation options through the previous context-tokens. In a strategy to persist in the conversation with the user, we provide some options related to the entity. If no previously-stored entities are discovered, the agent requests a rewrite of the phrase.

UC.5 **Complaints**

Finally, there is the scenario of a complaint. This type of situation usually occurs due to incorrectly answered or unanswered interactions. Again, we apologize to the user and save the tokens and requested sentences for future interactions.

[2]http:/platformabrasil.saude.gov.br

5.2.2 1º User Experiment- Development and Validation of Conversational Agent to pregnancy nutritional education

We have chosen to conduct clinical experiments to investigate the learning capacity of pregnant women based on the information retrieval by the chatbot. This method aimed to test the effectiveness of conversation, user satisfaction, and system usability (COHN; CHEN; YU, 2019). We evaluated the experimental design in a pilot study with a small number of participants to validate the experiment's design on a larger scale. The pilot study recommended changing the full-scale experiment based on these findings to ensure accuracy. A pilot study can provide valuable information while also highlighting design differences. We used a semi-structured quantitative survey to assess the user experience (EDIRISOORIYA; MAHAKALANDA; YAPA, 2019).

5.2.2.1 Participation and Evaluation

The non-probabilistic convenience sampling was used to select pregnant participants and health professionals, with chosen participants based on their availability rather than statistical criteria. We also used the snowball strategy for selecting physicians for this study, in which we asked each participant to recommend additional known candidates. Pregnant women were sampled in health centers with a face-to-face phase to explain the experiment, instructions to access the prototype via the internet, and clear up any doubts.

The study also analyzed the different perceptions of pregnant women and health professionals with conversational agents, considering the literacy levels. The authors translated and adjusted the questionnaire for this study. The scales proposed by (JACK et al., 2015)) are the basis for the proposed questionnaire. The instrument aims to determine the effectiveness of the conversation, user satisfaction, and system usability site (COHN; CHEN; YU, 2019). We are shown in Table 11.

5.2.2.2 Chatbot Design

The content brought by the chatbot sought to solve the doubts related to the nutrition of pregnant women in prenatal care. We called BotMaria. A doctor who worked in the public health network and was an assistant in this study selected this content. The content selection occurs according to the perceived importance and frequency in their work environment. In the DialogFlow NLP engine, we train 35 intents with five to seven sentences for training and two to three entities per intent. We use context mechanisms to guide the button-driven dialog. The main topics addressed were menu suggestions, prohibited foods, drinks, supplementation, and awareness for baby feeding in the first days. An example of this flow is in Figure 25.

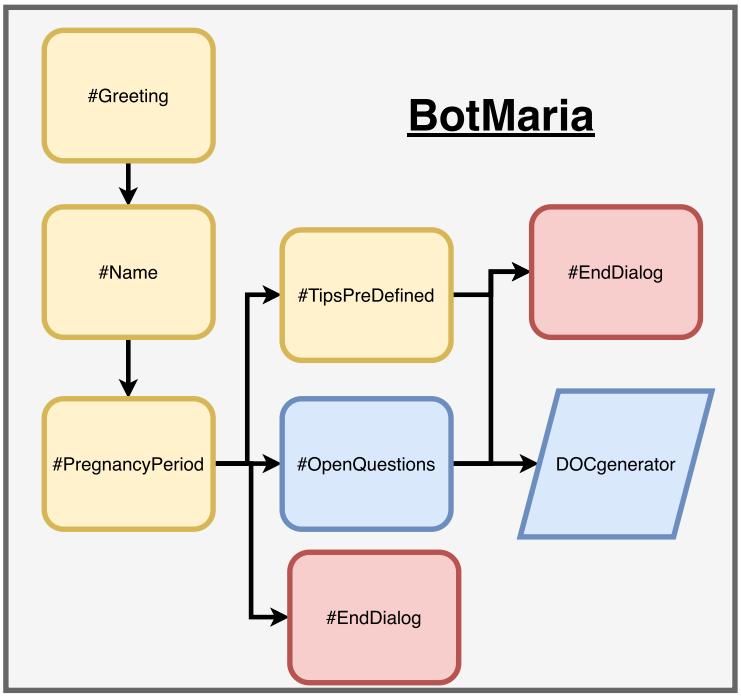
Table 11: Data collection instrument for pregnant women and health professionals

Items
Age
(18 - 24)
(24 - 30)
(30 - 41)
(more than 41)
Education
(Elementary School)
(High school)
(University education)
(Postgraduate studies)
Medical Specialty
(Obstetrician)
(Family Doctor)
(Nutritionist)
Items
I felt satisfied using the assistant
I felt comfortable with the assistant
It is easy to interact with the assistant
I liked the content of the answer
The assistant showed empathy
The assistant teach me a new knowledge
I believe that this robot can be used to assist pregnant women
I would recommend this robot to a pregnant patient.
Does the content presented affect decision-making ?
The content provided by the robot is reliable
Is the language suitable for all social levels?

5.2.2.3 Intervention

We assessed the effectiveness of the experimental design in a pilot study with a small number of participants to validate the experiment design on a larger scale. Questionnaires were explained to pregnant women during pre-consultation at the health center. Later, it was agreed with the pregnant women to interact and answer the questionnaires in their homes. The same processes were valid for health professionals. From November 2019 to February 2020, we conducted this experiment. We compose our sample with (N=35). The first is composed of pregnant women between 18 and 40 years old with mobile and internet access. The second group,

Figure 25: Chatbot Maria for interaction design



Source: elaborated by the author

comprised of nutritionists, obstetricians, and family doctors, was recruited through snowball sampling, in which study participants recruit future subjects from among their acquaintances.

The first study stage was conducted in person or over the phone to obtain the pregnant woman's consent and report the research terms. We explain the study design and send the questionnaire via messenger service or email after receiving verbal consent over the phone or in person. Both groups go through this process. The chatbot is on the GestanteHelp page on Facebook Messenger [3].Both groups should test BotMaria even if their dialogue is more directed at the pregnant women group.

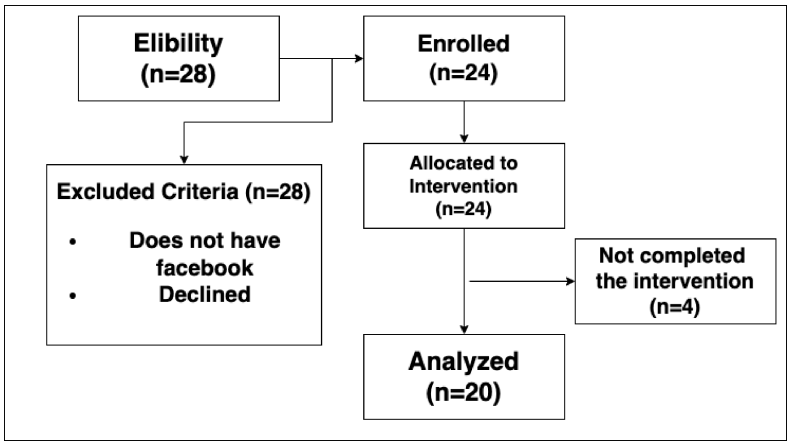
The first group had its questionnaire composed of two demographic categories variables six ordinal variables related to the level of perception of the individual about the agent. The questionnaire for health professionals involved two demographic categories variables and five ordinal variables to the perception level of each professional about the agent. We used a five-point Likert scale for ordinal variables, with one indicating strong disagreement and five indicating agreement with the question statement. Two physicians reviewed both questionnaires to ensure that the questions were consistent and clear.

5.2.2.4 Data Analysis

The study employed descriptive analysis to compute the mean and standard deviation for each variable and a 2-samples t-test hypothesis method to determine whether the group's per-

[3]<https://www.messenger.com/>

Figure 26: CONSORT: checklist template for the clinical trial



ceptions differed statistically. According to (WINTER; DODOU, 2010),the analysis of two distinct groups through a five-point Likert can be performed consistently by the 2-samples t-test for samples above 30 individuals. For this evaluation we considering the alpha p-value = 0.05. If the p-value is less than 0.05, we reject the null hypothesis, that is, that the perceptions between the groups are equal. The analysis occurs in the Jupyter Notebook tool [4].

5.2.3 2º User Experiment- Mixed perception over chatbot aimed at pregnant education

In this evaluation, we propose a pilot study to assess the use of a chatbot tool for prenatal and postnatal monitoring. We conducted a qualitative survey of physicians and a quantitative survey of pregnant women, analyzing their responses by the parallel convergence method and comparing them to look for confirmations, distortions, and new insights. Validations were proposed for usability, trust, hedonism, intention to use, ease of use, anxiety, and medical influence. We believe that the research can provide new insights into the context of chatbots for this topic.

5.2.3.1 Participants and Evaluation

We conducted a randomized study using the Consolidating Standards of Reporting studies (CONSORT) guideline. The CONSORT is a statement that includes ten recommendations for conducting clinical studies. The CONSORT criteria are well-known as a gold standard for this type of research. Therefore, the checklist given in (HECKSTEDEN et al., 2018) was used as a template to gauge overall compliance to assess the quality of each study. The CONSORT recommendations applied are shown in Figure 26.

The study began in October and ended in late November 2021. Due to the inability to meet the study's eligibility requirements, 24 of the 28 participants chosen were excluded. In addition, four more questionnaires were discarded due to inadequacy.

[4]<https://jupyter.org>

N	Question
1	What is your opinion about chatbot topics?
2	What is your opinion about the language type used?
3	In your opinion, is the information correct and reliable ?
4	Do you consider it easy to interact with this tool?
5	Do you have any suggestions for future improvements?

Table 12: Questionnaire developed for qualitative assessment with physicians

This study used qualitative and quantitative instruments to collect information about the chatbot from physicians and patients. One of our objectives was to validate perceptions based on age, and thus both questionnaires included an age variable. The study divided pregnant women and physicians into two groups. Pregnant women in the prenatal or postnatal stages, as well as physicians' from maternal care units, are recruited. The pregnant woman needed to have a Facebook account to participate in the studies. The health professionals were not excluded from the study. We enrolled a total of seven physicians and thirteen pregnant women. Participants interacted with the chatbot for seven days before answering the questionnaire. After the period, they were reminded to use the tool and complete the questionnaire.

We opted for a reduced sample because our objective with this first study was to capture perceptions related to the two groups without failing to observe contributions and reported failures about the chatbot. Numerous participants could contribute to neglecting some improvements and biasing our mixed analysis in case of many outliers.

The study looked at a pilot version of a text-based chatbot designed to help pregnant women. We conduct structured and individual interviews with physicians to obtain information about the chatbot's content, language type, facility, and trust. Also, we ask for recommendations for improving the chatbot's performance for the group of physicians'. The questionnaire is shown in Table 12.

Pregnant women were applied to the quantitative methodology. Both methodologies aimed to collect opinions and perceptions about the chatbot. To assess health literacy in pregnant women, we used the UTAUT2 scale of (VENKATESH; THONG; XU, 2012), which has also been used in other studies (CHANG et al., 2021)(MOKMIN; IBRAHIM, 2021) (DUARTE; PINHO, 2019).We summarized this scale to reduce the questionnaire's response time, redesigning several items and adding others. We removed several items, leaving only those most relevant to this study. Table 13 depicts the final questionnaire.

The data was collected through an online form, a link that participants could access and evaluate via email, and the chatbot structure via a button. In addition, we used a structured questionnaire sent via WhatsApp [5] for qualitative data collection, receiving responses from physicians' asynchronously via text and audio.

[5]<https://www.whatsapp.com>

N	Performance Expectancy
Q1	I find the knowledge is helpful and the chatbot enables me to understand health better.
Q2	Using chatbots improves the knowledge of my health.
	Effort Expectancy
Q3	It is easy for me to improve my health knowledge when using the chatbot.
Q4	I considered it easy to improve health knowledge from a chatbot.
	Attitude Toward Improving Health
Q5	Improving health via chatbot is fun
Q6	Using a chatbot makes the conversation more interesting.
	Medical Influence
Q7	I believe that the use of a chatbot can support my doctor.
Q8	I believe that my doctor would approve the chatbot's use.
	Facilitating Conditions
Q9	I have the resources necessary to use the chatbot.
Q10	I know what is necessary to communicate with the chatbot
	Self-Efficacy
Q11	I could use a chatbot if I have much time
Q12	I could use a chatbot if I find helpful information
	Anxiety
Q13	I feel anxious about getting information via chatbot.
Q14	Using chatbot is intimidating for me
	Trust
Q15	I think that I trust in this technology
Q16	I think that this technology is designed for my needs

Table 13: Questionnaire for pregnant women Intervention

5.2.3.2 Measures and Hypothesis

In this section, we describe the original and modified constructs that support UTAUT2 and the hypotheses that can be derived from them. The studies listed below serve as the foundation for our model (ALMAHRI; BELL; MERHI, 2020)(VENKATESH; THONG; XU, 2012), (CHANG et al., 2021)(MOKMIN; IBRAHIM, 2021) (DUARTE; PINHO, 2019), (DECMAN, 2020)introducing a new change in the construct from social and medical influence.

Performance Expectancy (PE) is defined as "the degree to which an individual believes that employing the system will assist him or her in achieving advances in job/task performance." Hypothesis 1: We believe that PE positively affects pregnant women's behavioral intention to use chatbots.

Effort Expectancy (EE):is described as "the degree of ease connected with the use of the system." Hypothesis 2: Effort Expectancy has a beneficial effect on pregnant women's behavioral intention to use chatbots.

Facilitating Condition (FC): is defined as "the degree to which the individual believes that an organizational and technical infrastructure exists to support the system's use". Hypothesis 3:

Pregnant women's behavioral intention to use chatbots is positively influenced by the Facilitating Conditions.

Medical Influence (SI):is defined as "the degree to which an individual perceives that their physician believes they should use the new system". Hypothesis 4: Medical influence positively affects pregnant women's behavioral intention to use chatbots.

Attitude Toward Improving Health (AH): refers to "the fun or pleasure derived from using technology". Hypothesis 5: Attitude Toward Improving Health positively affects pregnant women's behavioral intention to use chatbots.

Self Efficacy (SE):refers to "people's ability to self-serve with the tool". Hypothesis 6: Self Efficacy positively affects pregnant women's behavioral intention to use chatbots.

Anxiety (AX): is defined as "uncertainty about the behavior and the seriousness of the potential consequences of the behavior." Hypothesis 7: Anxiety increases pregnant women's behavioral intention to use chatbots.

Trust (TR): individuals must trust technology in educational settings, even more so if responses are to be collected accurately and anonymously". Hypothesis 8: Trust positively affects pregnant women's behavioral intention to use chatbots.

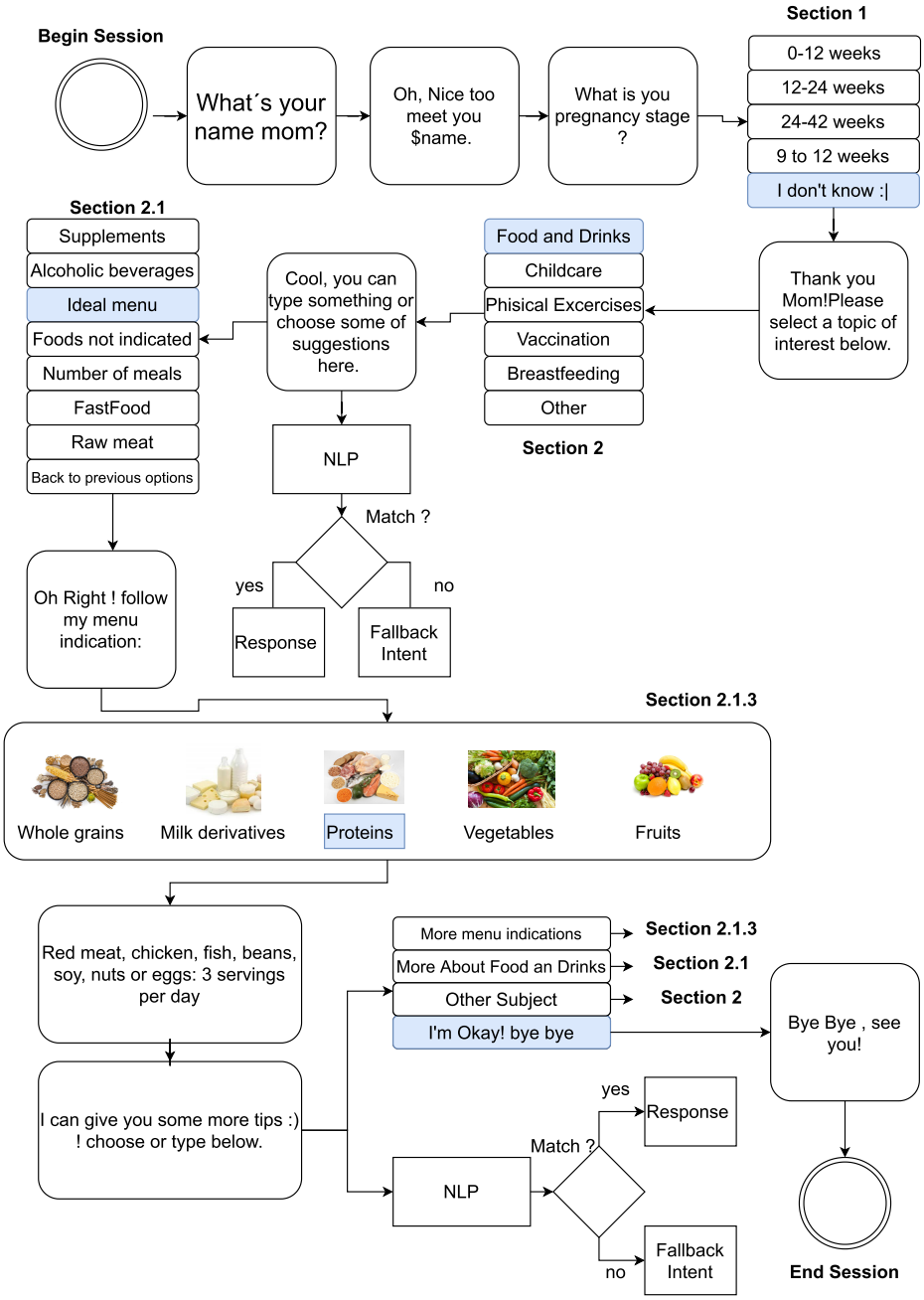
5.2.3.3 Chatbot Design

The chatbot development used the DialogFlow tool Figure 27 illustrates the dialogue flow. A researcher oversaw and assisted in the conversation's flow development and content selection. The topics were chosen based on their importance. Whether or not the user was pregnant, the design was built on both circumstances. At the end of each flow, there was an option to return to other flows and continue the interaction or end the conversation by providing feedback. The dialogs used rule-based and NLP strategies. We train the Dialogflow machine learning engine on ten training phrases plus three entities to support the intention classification for navigation/interaction via questions. If a phrase was misunderstood, a fallback was activated to redirect the user to another flow or repeat the query.

5.2.3.4 Data Analysis

We propose a convergent parallel analysis based on quantitative and qualitative data. This design aims to provide equal weight to analysis when looking for patterns or contradictions in the data by comparing and contrasting the outcomes of both groups. We used mean and standard deviation to identify which components were most and least essential to respondents for quantitative evaluation of pregnant women. We transcribed the doctors' interviews to increase understanding of the outcomes, reviewing and summarizing significant points generated to record essential concepts from the conversations. The viewpoints of physicians' were expressed using literal citations.

Figure 27: Chatbot Maria for interaction design



Source: elaborated by the author

Table 14: Pre-trained BERT models were used in this thesis for fine-tuning and experiments.

N	Model	Citation	Corpus
1	BERTImbau [6]	(FILHO et al., 2018)	brWaC
2	Paraphrase-Multilingual [7]	(REIMERS; GUREVYCH, 2019)	Microsoft-Multilingual
3	BERT-Multilingual [8]	(DEVLIN et al., 2018)	Wikipedia

5.2.4 Model evaluation

This section describes two experiments conducted for information retrieval model evaluation. The first experiment assessed Sentence-BERT models that had been adjusted to health-related data. We evaluated through validation and test processes to determine which model should be integrated into the HoPE architecture. The second experiment assessed our model's ability to retrieve information. The sections that follow will discuss the experiments and evaluation techniques used.

5.2.5 3º User Experiment- Pre-trained Portuguese Sentence-BERT models for retrieval pregnancy information

We present some evaluations of Portuguese pre-trained models fitted with in-domain data to determine which information retrieval model is optimal for the HoPE architecture. We conducted two assessments: the first evaluated the quality of embeddings generated, and the second assessed the models' performance on our validation and test data using similarity metrics. To facilitate the comparison, we chose three major models for this evaluation. They are summarized in Figure 14.

The BERTimbau model is pre-trained on data from the brWaC corpus composed of 2.7 billion tokens annotated with tagging and parsing information. The number of web pages contributing to this set is 120,000 (FILHO et al., 2018). Another model used is the Sentence-BERT paraphrase Multilingual. It is pre-trained on MAchine Reading COmprehension (MS-Marco) corpus proposed in (NGUYEN et al., 2016), which offers a large dataset extracted from web documents using the most advanced version of the Bing search engine. About 100,000 queries are present in this dataset, used for information retrieval tasks. For our study, we used its translated version into Portuguese.

The last model used was the traditional BERT-Multilingual, trained on millions of Wikipedia articles and translated into Portuguese (DEVLIN et al., 2019). Among the main differences between these models, the BERTimabau model was developed and uses a pre-trained dataset

constituted in Brazilian Portuguese, unlike the other two translated models.

We divided the data into training, validation, and test sets for this experiment. We allocated 80% for the training data, 10% for validation, and 10% for testing. We also present the metrics used to evaluate models. The Spearman and Pearson correlation coefficients were used to analyze the cross-encoders training phase, evaluating the embedding similarity. These coefficients have been extensively used in previous works to accomplish this task (REIMERS; GUREVYCH, 2019)(LI et al., 2020).

5.2.5.1 Evaluation Metric

We present the metrics to evaluate the Sentence-BERT models presented in this chapter. First, the Spearman and Pearson correlation coefficients analyzed the cross-encoders training phase, assessing the embedding similarity. These coefficients have been extensively used in previous works to accomplish this task (REIMERS; GUREVYCH, 2019)(LI et al., 2020).

Spearman's correlation coefficient is a rank-based alternative to Pearson's correlation coefficient that works with non-normally distributed and non-linear variables. Its use is not restricted to continuous data analysis; it can also be used for ordinal attribute analyses (CROUX; DEHON, 2010).

$$Sc = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

The Pearson correlation coefficient is the most frequently used metric for determining linear correlations between two normally distributed variables; it is occasionally abbreviated as the "correlation coefficient." Pearson coefficients are frequently estimated using a Least-Squares fit, with 1 indicating a perfect positive relationship, -1 defining a perfect negative relationship, and 0 denoting no relationship between variables (BENESTY et al., 2009).

$$Pc = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

For the Bi-Encoder layer evaluation, some distances are used. Here we present some of those. Based on Pythagoras' theorem, the Euclidean distance is a distance metric between two points or vectors in a two- or multidimensional (Euclidean) space. The distance is obtained by squaring the sum of the squared pair-wise distances in each dimension (LI; ZENG, 2018).

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Another evaluation metric is cosine similarity. The cosine similarity of two n-dimensional sample vectors determines their direction, regardless of their magnitude. It is calculated by taking the dot product of two numeric vectors and normalizing the result by the vector length product, with output values close to 1 indicating a high degree of similarity (LI; ZENG, 2018).

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

The Manhattan distance is used to calculate the distance between two real-valued vectors or points. It is calculated as the absolute difference between their Cartesian coordinates. For example, the Manhattan Distance can be defined for a plane that contains a data point p1 with coordinates (x1 y1) and its nearest neighbor p2 with coordinates (x1 y1) (x2, y2) (GRECHE et al., 2017).

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

The last metric used for evaluation in this experiment was the dot product, equal to the product sum of the horizontal components and the vertical components products (KILMER; MARSHALL; SENGER, 2020).

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

For this experiment, we apply stratified k-fold cross-validation (FUSHIKI, 2011) with k five (the class distribution remains identical for each fold) in our corpus to allow generalizing of our results. To assess the speed performance of HoPE architecture, we use two distinct measures: speed for queries requiring online vs. offline encoding and inference speed to get the answer to the user.

5.2.6 4º User Experiment- HoPE architecture evaluation

The evaluation of the HoPE architecture is covered in this section. To assess the effectiveness of the HoPE model, we conducted three experiments: test collection, inference speed evaluation, and multiple intents assessment. The model precision and recall were assessed in the Test Collection experiment. We used sixty question/answer pairs as gold pairs. As previously stated in (SAMIMI; RAVANA, 2014), this experiment comprises three components: a document corpus, which is a collection of large-scale documents; topics, which are collections of search queries; and relevance assessments, which involve human advisors.

The corpus for this experiment contains 3000 texts for research. For search collection, we use a batch of 60 queries. Some questions were derived from FAQs found on internet websites, while others were derived from the history of pregnant women's interactions in our clinical studies. We attempted to combine questions that contained short and long sentences, sentences with minor orthographic errors, and sentences that were ambiguous (one or two words). For the collection of sentences, two gynecology physicians performed the judgment phase to identify pairs (consultation/retrieved response). Pairs (query/correct retrieved answer) and (query/no answer) (1.0) were labeled. Three of these pairs are shown in Table 15.

Table 15: Three examples of Golden pairs for retrieval information evaluation

N	Questions	Possible Answer
1	Can I drink alcohol?	Do not consume alcohol
2	What foods should I avoid during pregnancy? can I have a hamburger	Eat a small meal every three hours, - Avoid fried foods, coffee, black tea, companion tea, fatty and spicy food.
3	Birth plan?	The birth plan is a document prepared by the pregnant woman about her preferences, desires, and expectations about childbirth and birth, including some procedures of the professionals.

The inference speed, alternatively referred to as inference time, is a statistic used to quantify the time required to compare the history of a conversation or an input text to millions of candidate responses (TAHAMI; GHAJAR; SHAKERY, 2020). In this evaluation, we calculate the time for the encoding process of sentences to the model and the inference speed for information retrieval. Within the HoPE architecture, two encoding methods in the model have been used: pre-computed embeddings and online embedding generation.

The embedding representation of 247 paragraphs/sentences was generated offline, serialized, and indexed by ANN. Approximate Nearest Neighbor (ANN) search is a relevant strategy that preprocesses a set A of N vectors so that given a query vector b, an (approximately) closest vector can be found efficiently (RUBINSTEIN, 2018). Online embeddings are generated during the ontology ranking stage from selected paragraphs/sentences. The Google Collaborative [9] environment, which provides open-source GPUs and CPUs, is used for speed test experiments.

To evaluate multiple intent detection, we create a set of ten sentences and evaluate our system's ability to classify whether a sentence contains multiple intents or not correctly. The percentage hit rate is used as a metric to determine whether sentences have more than one intention. We categorize sentences as high and low complexity to identify possible system failures.

5.2.6.1 Evaluation Metric

This section will discuss the metrics used to evaluate the previously mentioned experiments. For HoPE model retrieval evaluation we applied two major metrics: confustion matrix and mean reciprocal rank. QA systems use mean reciprocal rank (MRR) as the measure to facilitate the evaluation of a system with this focus, which is defined as the inverse of the rank of the retrieved result. The higher it is, the better, with MRR=1.0 being the best case (when the result is at rank 1) (SHAH; CROFT, 2004).

We also apply a confusion matrix (DENG et al., 2016), a machine learning construct that stores information about a classification system's actual and expected classifications. The struc-

[9] <https://colab.research.google.com/?utm_source=scs-index>

Figure 28: Confusion Matrix for Information Retrieval assessment

Machine Classification	Confusion Matrix	Human Classification		
		Relevant	**Irrelevant**	**Total**
	Retrieved	Hits	Noise	***Hits + Fallback***
	Not Retrieved	Fallback	Reject	***Noise+Reject***
	Total	***Hits + Fallback***	***Noise+Reject***	***Hits+Noise+ Fallback+Reject***
	Table Description	**Recall**	***Hits / (Hits + Fallback) * 100%***	
		Precision	***Hits / (Noise+Reject) * 100%***	
		Accuracy	***Hits+ Rejected / (Hits+Noise+ Fallback+Reject) * 100%***	

Source: **Based in Turney, 2000**

ture is bi-dimensional with one dimension representing the object's current class and another representing the class predicted by the classifier. We measure the conversational agent's ability to retrieve or discard responses relevant to the end user using this metric. We presented this metric in Figure 28 below:

The items that make up the confusion matrix are described as:(1) Hit classified as relevant by human and system; (2) Noise, classified as irrelevant for the user but relevant for the system; (3) Fallback, classified as crucial for the user but irrelevant to the system. (4) Reject, classified as irrelevant for humans and the system. Through the matrix, we obtain indicators for evaluating the model.

Precision and recall are frequently combined in the F-measure of efficiency to provide a unified metric for a system (RAGAB et al., 2021). The F1-measure performance metric is the most used for text classification. It is known to be more informative and valuable than classification due to the widespread phenomenon of class imbalance in text classification (GUO et al., 2020).

The F1-Measure is a special measure with an equal weighting of recall and precision in the information retrieval context. It has a maximum value of 1 and a minimum value of 0.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

We employed two measures to assess inference speed: corpus size and processing type. Initially, we evaluated the HoPE architecture encoding using the entire corpus **C** = 3000 that

had previously been loaded pre-computed embedding and a smaller corpus **C** = 6 from ontology retrieval. Also, we seek to compare the inference performance for all models, when predicting the final response to the user.

Finally, the last evaluation aimed to test the classification system for multiple intentions. We use a simple hit rate metric to evaluate this system, shown in the equation below. The evaluation seeks to determine whether or not the model can match when a sentence has multiple intentions.

$$HitRate = \frac{hits}{hits + misses}$$

5.3 Final Remarks

In this chapter, we described all offline processes that support the architecture proposed in this dissertation, such as the construction of the corpus, the procedures for structuring and enriching data in ontologies and dictionaries, and, finally, the algorithms for training the models used in our experiments. Also, we present the methodologies, research instruments, and evaluation metrics that we apply in the proposed experiments. The next chapter describes how the HoPE architecture for conversational agents works, its main components, and its performance in online environments.

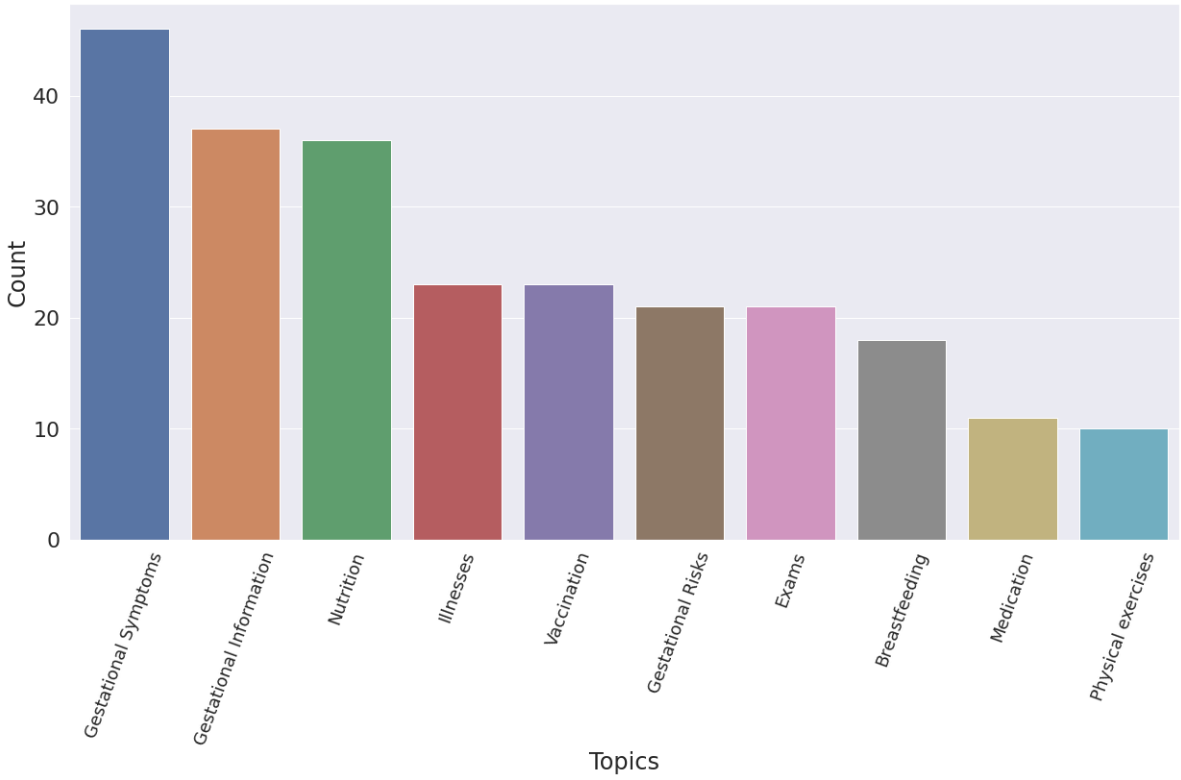
6 RESULTS AND DISCUSSION

This section contains the findings and discussion of the evaluations. In all experiments, we used data from pregnancy health guidelines. We began the assessment with an exploratory analysis of the corpus documents, followed by an ablative study of Sentence-BERT models to find the best performances to incorporate the HoPE architecture. Finally, we evaluated the information retrieval ability of the HoPE model, discussing aspects highlighted in the experiment, improvements, and limitations.

6.1 Corpus Evaluation

In this section, we explore the pregnancy guidelines content. We found several topics related to thousand days. We chose a smaller sample size for the evaluation steps because it contained more critical information about the post and prenatal periods. We selected this sample based on an examination of two assistant physicians. Two hundred forty-seven sentences were separated and then classified by topic. Figure 29 depicts the distribution of this sample.

Figure 29: Topics that appear more frequently in our documents



Source: elaborated by authors.

Ten topics were listed for the sample used as a knowledge base in the experiments. The most frequently discussed topics were gestational symptoms, gestational information, and nutrition.

Medication and physical activity had the lowest representation in our sample.

Prenatal symptoms are addressed in Gestational Symptoms, whereas general pregnancy information, such as frequently asked questions and curiosities, is covered in Gestational Information topics. Nutritional advice for pregnant women's eating habits and feeding recommendations for newborns.

Pregnancy alteration addresses aspects that change during pregnancy in the pregnant woman's body. For example, illness is a topic that covers information about the most common diseases that occur during pregnancy, whereas vaccination covers immunization information for pregnant women and newborns. Gestational Risks refer to data about preventive behaviors during pregnancy, while exams contain information on what to do during pregnancy monitoring. The final topics are breastfeeding, with incentive content and good practices—medication, which addresses medicines and cosmetics allowed during the prenatal period. And physical exercise, with content, focused on the pregnant woman's weight and physical exercises.

We also analyzed the frequency for each topic to understand the most frequent terms among the domains as shown in Figure 30. According to our corpus's analysis, the most frequently mentioned gestational symptoms included childbirth plans, gestational factors, age doubts, and menstruation. The majority of gestational information focuses on consultations, gestational age, and pregnancy phases. Nutritional problems are related to information about foods. The topic of illnesses brought up more frequent subjects such as syphilis and prematurely, whereas vaccines brought up more frequent alerts about influenza and contraindications, syndromes, examination techniques, and diagnoses. Breastfeeding discussed the aspects of breast milk intake, how to do it, and the benefits, considering medication discussed terms related to prevention and physical exercise discussed weight training and perineal exercises.

6.2 Clinical Study 1- Development and Validation of Conversational Agent to pregnancy nutritional education

In this section, we present the results of our first assessment. We divided our analyses into a) descriptive analysis, b) group's perceptions analysis, and c) most significant variables analysis. The study used instruments such as email and Whatsapp tools [1] to send the questionnaire links and the Facebook messenger to interact with the final user, as shown in Figure 31.

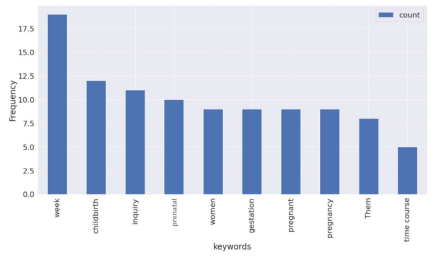
We observed that the pregnant women group who received the questionnaire via Whatsapp messenger answered 100% of the questionnaires (27 of 27 people). On the other hand, only 30% (8 of 25 people) answered the questionnaire among those who received it by email.

Fifty individuals were invited to participate in this study; ten tested the agent but did not complete the questionnaire, and five did not have Facebook accounts. To begin, we examined the groups' baseline demographic characteristics. For groups one and two, we have the variables

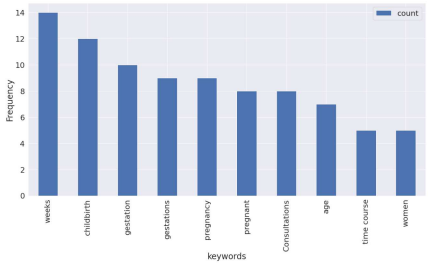
Figure 30: Most frequent keywords for each topic.

[1] <https://www.whatsapp.com>

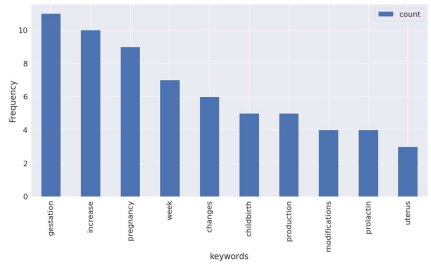
Figure 30: Most frequent keywords for each topic



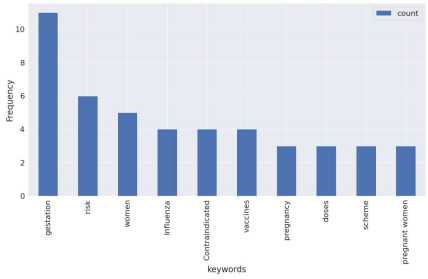
(a) Gestational Symptoms



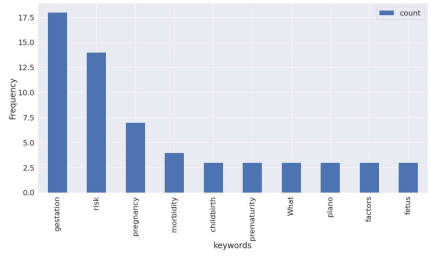
(b) Gestational Information



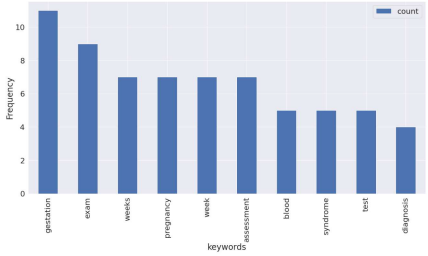
(c) Nutrition



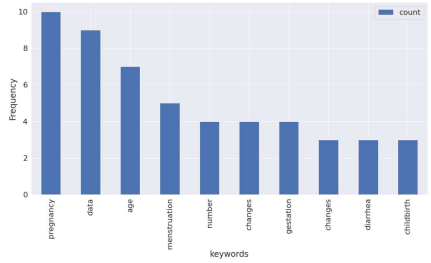
(d) Illnesses



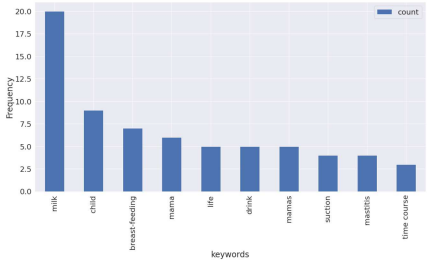
(e) Vaccination



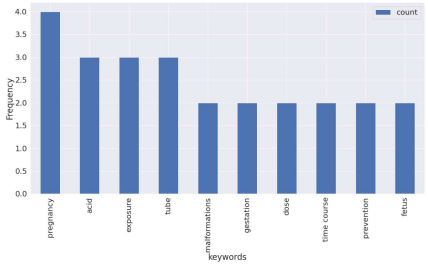
(f) Gestational Risks



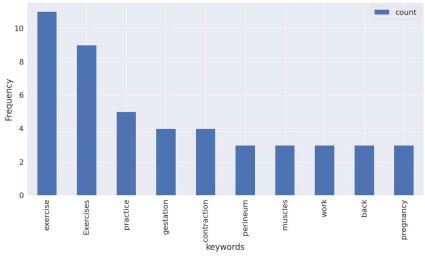
(g) Exams



(h) Breastfeeding



(i) Medication



(j) Physical Exercises

Source: elaborated by authors.

Figure 31: Example of real user interactions in the experiment

Source: elaborated by the author

Regarding the general group, the age range of the participants was between 30 and 41 years old (Mean = 51%). For the group of pregnant women, most of the participants were also in this age range (Mean = 36%), while health professionals were all over the age of 30. As for the educational level of pregnant women, most respondents have higher education (Mean = 36%).

As for pregnant women, most of them have a higher education degree, although the group is quite heterogeneous. With the advancement of the internet and technology, women with higher education are three times more likely to seek advice with new technologies than women with less education (SAYAKHOT; CAROLAN-OLAH, 2016). Furthermore, the literacy level correlates to the information absorption and possible prevention of disease during pregnancy (MURAKAMI et al., 2009). As far as health professionals are concerned, the group majority consisted of an obstetrician who is professional most often in the prenatal care of pregnant women. The results presented by the two groups show the conversational agent and the model presented approval rates for the continued evolution of the study.

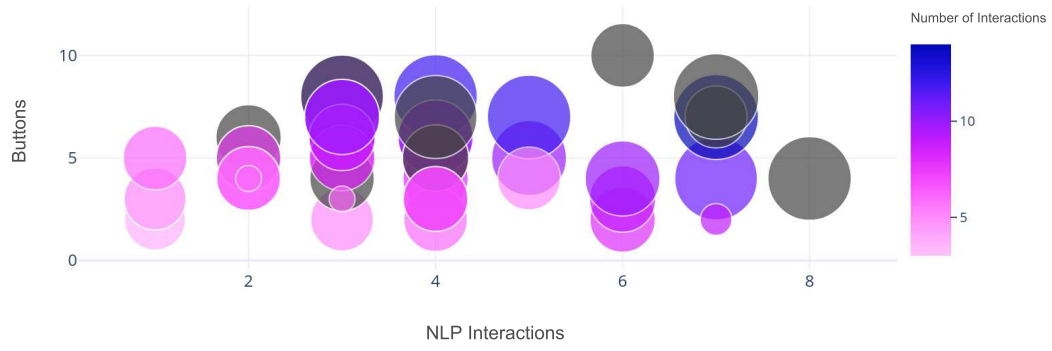
Figure 32 the number of interactions that triggered the model implemented using text interactions (axis x) and the number of interactions by buttons (axis y) that pregnant (purple) and health professionals (black) used for search information. The bar next to the chart depicts the interactions of pregnant women on a color scale; the higher the number of interactions, the darker the shade of purple. The NLP model answered 75 open-ended questions, with 74% responding and 26% not responding. Unanswered questions were analyzed in the DialogFlow tool's fallback module and considered outside the scope of our training.

By default, the conversational agent architecture does not respond to questions that are not within the scope of the training. Open questions for the user provide the experiment with a

Table 16: Descriptive analysis of the groups participating in the sample

Items	**Pregnant (n=25)**	**Professionals (n=10)**
Age		
(18 - 24)	32	-
(24 - 30)	32	-
(30 - 41)	36	100
(more than 41)	-	-
Education		
(Elementary School)	24	
(High school)	28	
(University education)	36	-
(Postgraduate studies)	12	-
Medical Specialty -Open Question		
(Obstetrician)		50
(Family Doctor)		20
(Nutritionist)		30

Figure 32: Analyzes of responses generated by the model versus predefined responses. The size of the circles indicates the intensity of the interaction



Source: elaborated by authors.

Table 17: Average and Standard Deviation analysis for each questionnaire item

Items	Mean (n=10)	SD (n=10)
I felt satisfied using the assistant	4.40	0.70
I felt comfortable with the assistant	4.44	0.71
It is easy to interact with the assistant	4.28	0.89
I liked the content of the answer	4.52	0.71
The assistant showed empathy	4.36	1.07
The assistant teach me a new knowledge	4.56	0.76
I believe that this robot can be used to assist pregnant women	4.7	0.67
I would recommend this robot to a pregnant patient.	4.5	0.52
Does the content present affect decision-making?	4.4	0.69
The content provided by the robot is reliable.	4.6	0.51
Is the language suitable for all social levels?	4.6	0.51

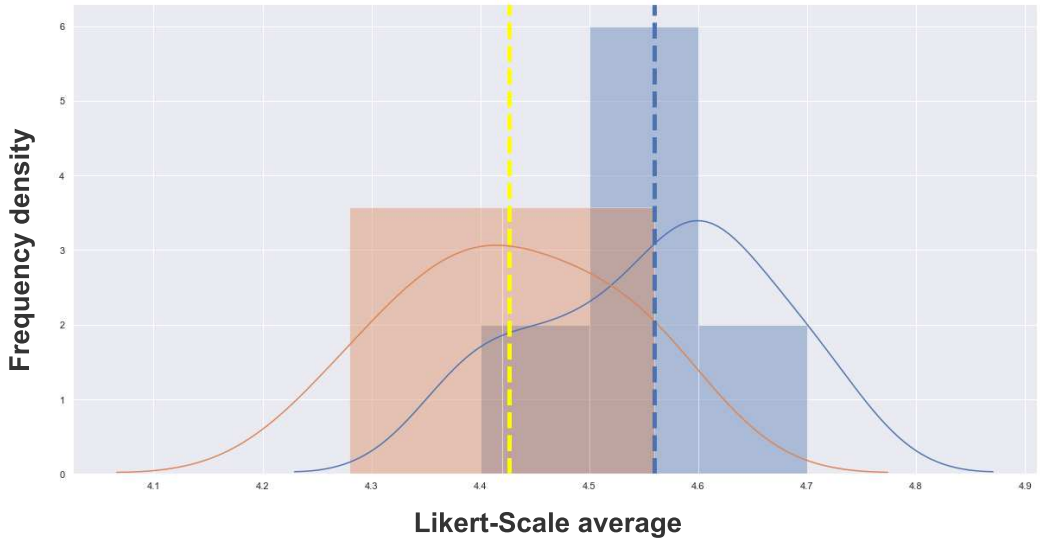
sense of understanding and a greater nature of the dialogue, encourage the user to engage, and make the conversation less boring (SHEVAT, 2017). In future research, we plan to assess the model's performance in terms of answers provided in this experiment.

Afterward, we started our analysis of each variable using an average and standard deviation. In the first group, the variable ("The assistant taught me new knowledge") was reported as the most significant for the group (Mean = 4.8, SD = 0.78). In contrast, the least important for this group was the variable ("It is easy to interact with the assistant ") (Mean = 4.28, SD = 0.79). In the second group, the variable ("I believe this robot can be used to aid pregnant women") was the one with the highest agreement among health professionals (Mean = 4.7, SD = 0.67). In contrast, the variable ("The content affects the decision-making? ") (Mean = 4.4, SD = 0.69) the least significant.

Then, we compared the groups using a 2-sample test-t for different sample sizes. The results did not reject the null hypothesis, with no statistically significant difference between the two groups (P-value = 0.713). The results show that for this sample of health professionals and pregnant women, perceptions of nutritional information through chatbots do not have significant differences. Figure 33 shows the difference between the group's average and the average concentration of participants' answers. We could observe that both groups have a good perception of the experience of chatbot use.

In group one perception formed by pregnant women, the conversational agent taught them new knowledge, which is the most significant variable in the survey. At first view, the conversational agent had a good relationship with the pregnant women through accessible language, delivering the required information (GARDINER et al., 2013),(FADHIL; WANG; REITERER, 2019).

Figure 33: Frequency density of answers plus average of responses per group plus difference between means. The bars represent the average concentration, where the orange is the health professionals and blue represents pregnant women. The density lines demonstrate the variability within the groups and the dotted lines demarcate the difference between the means of the two groups



Source: elaborated by authors.

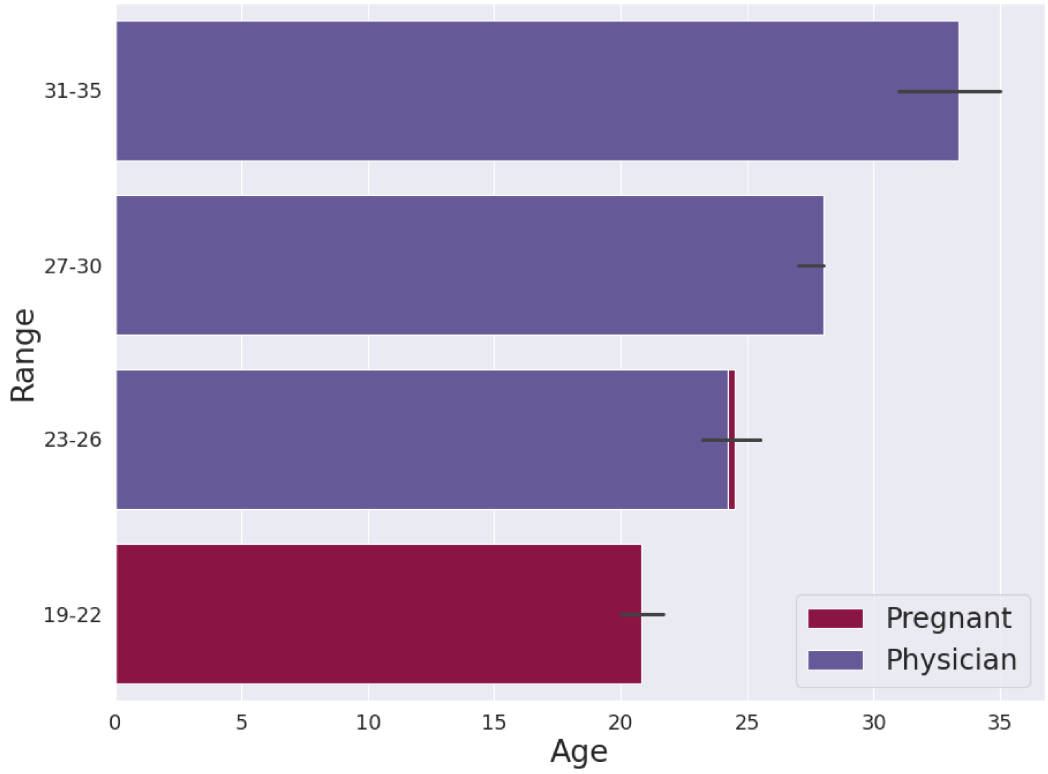
The experiment results also showed a good perception of health professionals about the conversational agent. These results are corroborated by other studies (MAARUP et al., 2019) (SARWAR et al., 2019). Despite this, the specific relationship of doctors with these technologies still needs research, with greater depth and breadth, as highlighted in(MAARUP et al., 2019). Although the number of people interviewed was lower than in another group, the results were superior to those presented in a health care provider experiment, which concluded that the agents were not yet prepared to act as assistants during the pregnancy period.

Health professionals recognize that agents can also influence pregnant women's decision-making, even though this result was less expressive than the other variables in the study. This conclusion is supported by the reliable information sources used in this study, as the NLP model consumes data from the state government's guidelines. Most users promoted positive feedback related to their experiences with the conversational agent. However, some people have expressed frustration or annoyance due to unanswered questions. This situation happened in other studies of conversational agents and pregnant women (SHAWAR; ATWELL, 2009).

6.3 Clinical Study 2 - Mixed perception over chatbot aimed at pregnant education

The average age of participants in this second study was less than 30 years. Pregnant women with an average age of 23 years and a low income answered the quantitative questionnaire. While the physicians' group was made up of newly trained physicians, residents, and specialists, all of whom were 28 years old. Figure 34 depicts the age distribution for the two groups 32.

Figure 34: Age distribution for physician's and pregnant in chatbot assessment



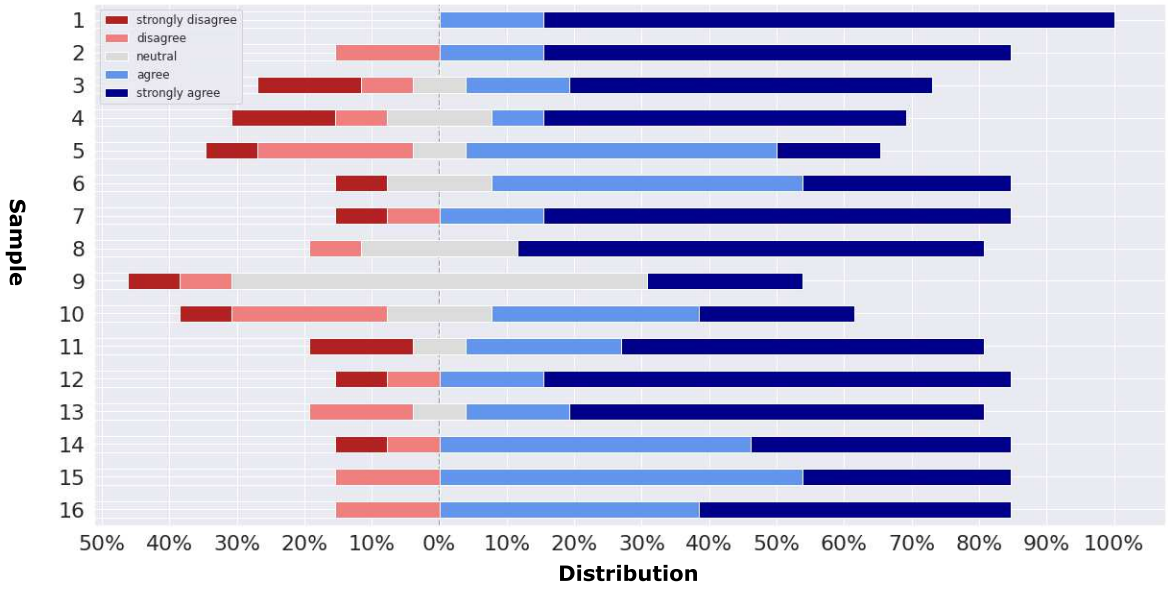
Source: elaborated by authors.

Pregnant women were assessed using a questionnaire with a five-point Likert scale. The constructs are rated from 1 to 5 on a scale of 1 to 5. (1 = strongly disagree, 5 = strongly agree). Incompatibility with a pregnant woman's mobile phone was reported during the experiments, but it was solved using the desktop application. Figure 35 describes the distribution of responses using the Likert scale. The pregnant majority of respondents have positive attitudes toward the chatbot, with more answers closed to 5 than 1.

Based on the data from the Likert scale, we grouped the items by construct to see which one stood out the most. Figure 36 presents the results from the experiments carried out with patients. The bars in blue color indicate the mean, while the error bars (black) indicate the standard deviation for each questionnaire item. We found that, in general terms, the results obtained support the use of the chatbot by pregnant users. The construct that had the highest average in this experiment was the Expectation of performance ([Mean 4.61][SD 0.74]), while the construct that obtained the lowest average was the Facilitating Conditions construct ([Mean 3.30],[SD 1.24]).

A qualitative questionnaire was used to discover the physicians' perceptions. The goal was to determine physicians' attitudes toward suggestions and address any issues. Among those present were general practitioners, gynecologists, and medical students. All professionals and students agreed on the tools used. Table 18 depicts the primary physician's point of view.

Figure 35: Pregnant women's response distribution on a five-point Likert scale.



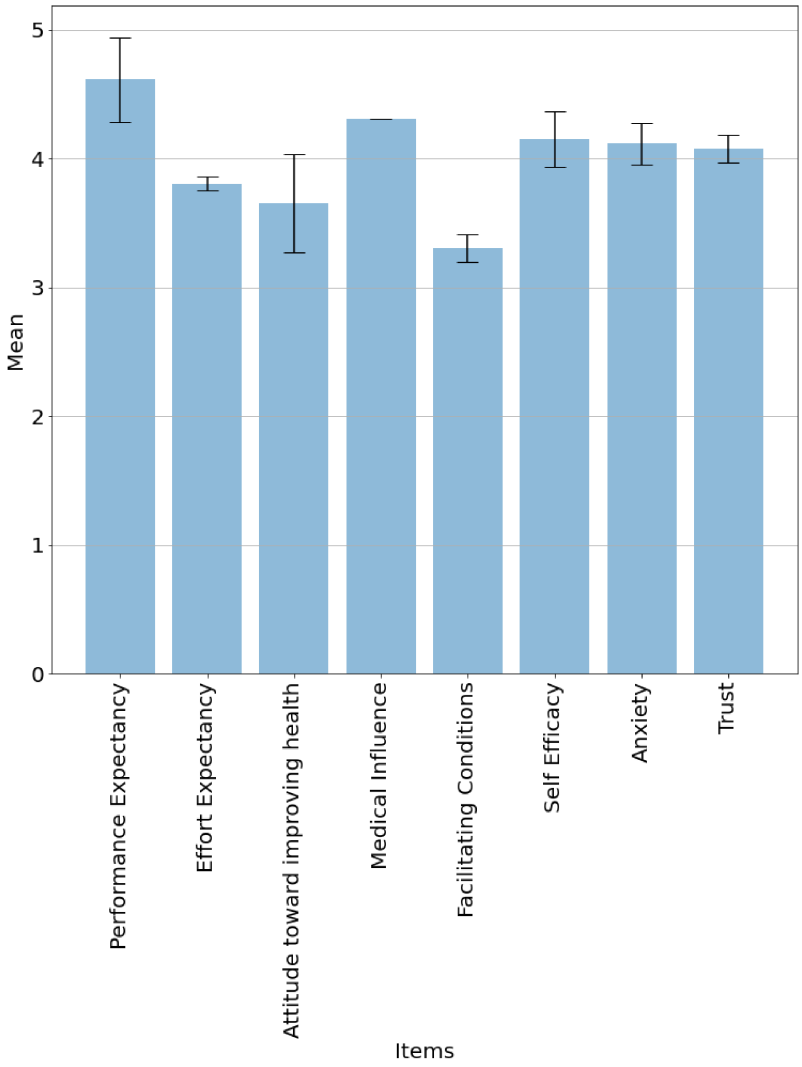
Source: elaborated by authors.

Table 18: Physician's perception about chatbot evaluation

General consideration about the chatbot
1. "A great tool to help pregnant women and resolve doubts"
2. "Quick and simple mechanism of resolving issues"
3. "I thought it was complete and important for this period"
Language Perceptions
4. "Adequate and effective "
5. "Pretty easy to understand "
6. "Very good! Sympathy and simplicity promote proximity to the reader "
Content Perceptions
7. "Not in the information, but in the functioning of the chat. I got stuck, without the chat moving forward... after a while it worked normally"
8. "Clear and lay language, easy to interpret and understand. I believe that topics can be divided by themes, not just diseases 1, diseases 2..."
Future directions for chatbot
9. "I suggest a better approach in the questions, so the person can better choose what they want to know, what is their doubt"
10. "Perhaps some notification/message recommending finding a doctor based on more serious situations/that may pose a risk to the pregnant woman depending on the topic of choice or doubt"
11. "Topics related to covid"
12. "Sexual relationship during pregnancy
13. "Approach a user profile in the father role. He also has doubts that he should participate in prenatal care, and the program could also involve him"

Source: elaborated by authors.

Figure 36: Mean and Standard Deviation of pregnant women perception



Source: elaborated by authors.

The first hypothesis tested in this experiment with pregnant women was that the expectation of use would positively influence the intention to use the chatbot, which was validated. In addition, the (CHUNG; CHO; PARK, 2021) study, which included obstetric and mental health care content for perinatal women and their families, found that high-quality contents provide users with relevant value, increasing their interest in the tool.

Physicians' assessment that the tool brought complete and crucial content. Still, the chatbot was considered easy interaction by professionals. In this sense, we ask about the language used, since one of the main focuses of this chatbot is its use in health units with low-income people (VAIRA et al., 2018). According to all physicians', the language used was clear and didactic.

We assessed this behavior in pregnant women using the Effort Expectancy and Facilitating Conditions constructs. These constructs stood out as the worst among the items evaluated, despite having a more positive than negative perception on average. Effort Expectancy aspects related to understanding the content and obtained an average [3.8].

Facilitating conditions envisaged as a hypothesis that pregnant easily use the chatbot since they would have the necessary resources. However, we received two informal feedbacks from pregnant women who complained that the chatbot interactions were too slow and had incorrect button functionality.

In addition, the chatbot's performance was not tested on different mobile phones, which may have interfered with the user's experience with their device. The approach had reservations in the question presented and the separation of topics. Healthcare professionals have reported that button captions have become small, and navigation could be improved by reducing the number of buttons.

Although text dialogues are available, the chatbot was built on the rule-based concept, encouraging users to interact by clicking rather than using text dialogues. As a result, less than a quarter of all respondents used NLP mechanisms. This finding and results collected in another study (MUGOYE; OKOYO; MCOYOWO, 2019),suggested interaction by text, ensuring a satisfactory level of accuracy for responses, leading to engagement and availability of information that the increase the use of the tool.

Additionally, we examined physicians' attitudes toward chatbots to assist pregnant women. The study was carried out with pregnant women belonging to the Unified Health System in Brazil (SUS). The SUS pregnant's (FONSECA et al., 2012) are quite diverse, making simplified strategies and popular language the best alternatives for this tool. According to professionals, this technology is prepared to assist pregnant women due to its simple and quick nature. The majority of women participating in the study strongly believe that their doctors would approve of using the chatbot. However, there is no signed agreement on the aspects aimed at a feeling of leisure or fun when interacting with the chatbot.

We evaluated the Self-Efficacy construct, which refers to the infrastructure and knowledge requirements for interacting with the chatbot. The results show that pregnant women are not entirely comfortable with this tool in terms of Self-Efficacy. Despite the low average age, which may indicate a higher level of familiarity with this technology, the item ('Using the chatbot is intimidating for me') reach (avg=4.00), indicating possible discomfort or a lack of habit with this type of technology. The average for agreement on facilitating conditions (avg= 3.07) was the lowest, indicating that not all pregnant women believe they have the necessary conditions (internet, cell phone, technical knowledge) to use the chatbot.

On the other hand, pregnant women agree that they trust the chatbot (avg = 4.17) and want to learn more from it (avg = 4.11). One of the most difficult challenges this technology can overcome is providing information without instilling anxiety (MAEDA et al., 2020).). These tools are helpful for professionals, but they may also include new features. A variety of benefits for the chatbot have been proposed, including improved patient education and treatment adherence compliance. One of the future suggestions was a search-based system for finding a doctor in an emergency. There are already studies of generalist chatbots acting as physicians' counselors (JAMEEL; ANWAR; KHAN, 2021). Covid disease was omitted from our materials and was

Table 19: Fine-Tune cross-encoder models vs literature models

Models	Spearman
Sentence-BERTimbau Augmented	**90.55**
Sentence-BERT Multilingual Augmented	**90.33**
Sentence-BERT Multilingual	**89.21**
Sentence-BERTimbau	**83.97**
Not Trained on STS	
Avg. GloVe embeddings	58.02
InferSent	46.35
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	79.23
Trained on STS	
BERT-STSb-base	84.30
SBERT-STSb-base	84.67
SRoBERTa-STSb-base	84.92
BERT-STSb-large	85.64
SBERT-STSb-large	84.45
SRoBERTa-STSb-large	85.02

suggested to improve the content. Some studies have addressed this topic involving pregnant women and the prompt support of information for this disease (WANG et al., 2020a)(BAHJA; ABUHWAILA; BAHJA, 2020). Finally, the father's role during the Thousand Days was mentioned briefly but insufficiently in a few topics, necessitating further expansion of this content. According to doctors, pregnant women have numerous doubts about this subject, and sexual activities are not included.

6.4 Model's evaluation - Pre-trained Portuguese SBERT model's applied for retrieval pregnancy information

This section presents the findings of an information retrieval models evaluation that has been fine-tuned for use in health guidelines. First, we assess the embedding generation and validation processes. In the first assessment, we measured the models embedding the performance for finetuning cross-encoder. Next, we measured the Spearman rank correlation between the cosine similarity of the sentence embedding and pairs labeled. It has been used to measure semantic textual similarity in other studies (LI et al., 2020) (CARLSSON et al., 2021)(YIN et al., 2020). We show the performance in Figure 19, highlighting the templates used in our article in bold. The data not highlighted are from other studies in the literature that also performed the fine-tuning process with benchmark data. We put these results into the table to provide an overview of the performance of cross-encoder networks for this type of task.

The strategy of using pre-trained data has already been used in some studies with BERT networks in the health area (WANG et al., 2020c) (DAI et al., 2019). The models trained using

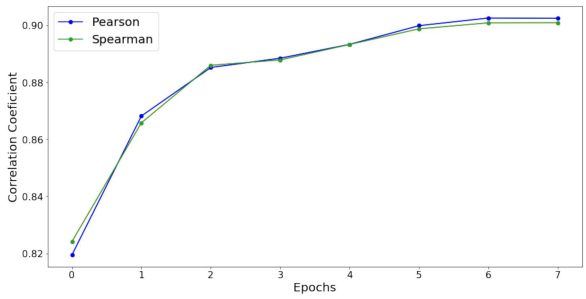
the Data Augmentation strategy have a slightly higher coefficient than models trained using the traditional bi-encoder architecture. The BERTimbau Augmented model, with a Spearman Coefficient (SC) of 90.55, performed best in this assessment than the BERT Multilingual Augmented model, which had a coefficient of 90.33. Models without data augmentation, such as the BERTimbau model fine-tuned in-domain, had a coefficient of 89.21, while the BERTimbau reached 83.97. For each epoch, we also calculated the Pearson coefficient. The loss and cosine similarity were calculated, resulting in the final output shown in Figure 37.

Each model's hyper-parameters were adjusted. At this point, we confirmed that the BERTimbau model has the best performance, having been trained in 8 epochs, whereas the BERTMultilingual model over-fitting when trained with this number. For fine-tuning the cross-encoder, we chose the ASSIN2 benchmark. The diversity of the vocabulary and the quality of the annotations were relevant factors in the dataset selection. Multilingual models in the traditional BERT framework were fitted to data from ASSIN2, and someone else obtained great results (RODRIGUES; COUTO; RODRIGUES, 2019). Pre-trained BERT models had already demonstrated the performance in the ASSIN2 (CABEZUDO et al., 2019) data classification task, surpassing the state-of-art.

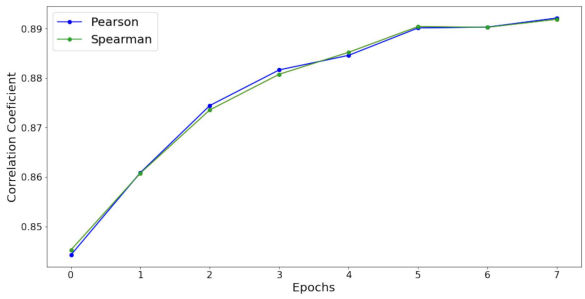
Spearman's and Pearson correlation coefficient for augmented models was superior to nonaugmented models. The results showed a significant difference between the Pearson coefficient for augmented and non-augmented models. Unaugmented data did not receive a fine-tuning of the benchmark data, reinforcing our belief that using a previous set can generate good results, improving the quality of embedding. This behavior has also been corroborated by (CHOI et al., 2021). For the Spearman coefficient, models trained exclusively on in-domain data demonstrated a correlation coefficient greater than 80%, demonstrating that in-domain data also has high-quality embedding.

This experiment also evaluated the models by putting them through their paces in our data domain. We used all previous models and a new one for this test: Paraphrase MultilingualMiniLM. We chose this model because it performed great in semantic search experiments (WANG et al., 2020b). To evaluate the validation and test sets, we use a set of similarity metrics, including the Dot-Product measure, Manhattan and Euclidean distances, and Cosine Similarity. We labeled our data using a cross-encoder BERTimbau Augmented as the model base for this evaluation. We chose this model because it performed the best in previous experiments for Spearman and Pearson coefficients. We present the results in Figure 38.

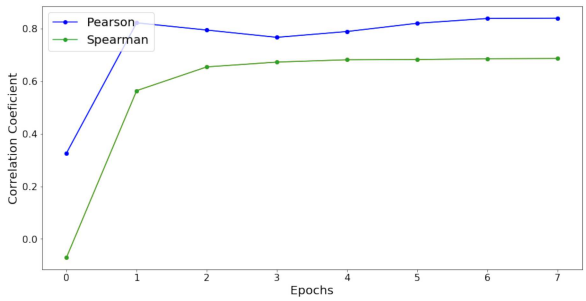
The labeling method used was binary scores, with 0 for distant sentence pairs and 1 for close sentence pairs. In this sense, we benefit from the labeling data transferring knowledge strategy from the STS benchmark, as the dataset was proposed for Natural Language Inference tasks. For bi-encoder training, we added the Paraphrase-Multilingual model. This model is native to Siamese networks, so fine-tuning is faster than in the other two models, and it also received fewer epochs. We used a lower learning rate for this model. It performed well in this experiment, with the best overall average among the similarity metrics for both the test and



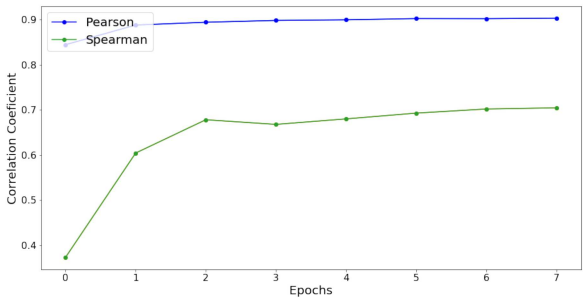
(a) BERTimbau Augmented



(b) BERT-Multilingual Augmented

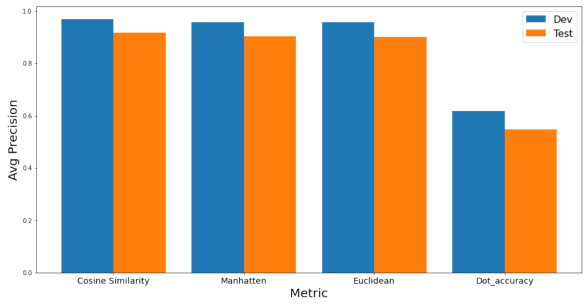


(c) BERTimbau

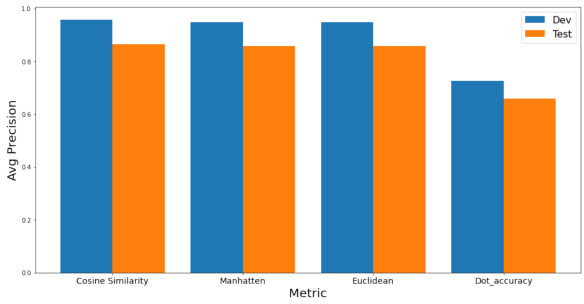


(d) BERT-Multilingual

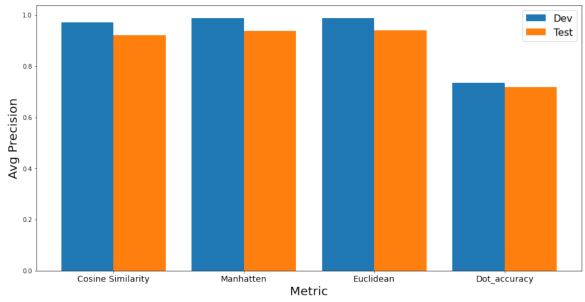
Figure 37: Evaluation of cross-encoder models



(a) BERTimbau Augmented



(b) BERT-Multilingual



(c) Paraphrase-Multilingual-MiniLM

Figure 38: Evaluation of bi-encoder models

validation sets. Similar results related to lower learning rates were observed in (KHATTAB; ZAHARIA, 2020).

For cosine similarity, one of the main metrics for this assessment, the best average was for the BERTimbau and BERT-Multilingual models. Compared to the obtained results, the Paraphrase-Multilingual model proved to be more consistent for this metric, with a bi-encoder classification reaching avg = 0.95. As a result of this research, we determined that the cross-encoder trained in BERTimbau Augmented and the bi-encoder trained in Paraphrase-Multilingual were the best combinations to use in subsequent experiments with the HoPE model architecture.

The two models' combination yields favorable results for response and re-rank recovery systems. The (THAKUR et al., 2020) study combines a cross-encoder with the ability to perform full attention over the input pair with bi-encoders, which map each input independently to a dense vector space. The cross-encoder strategy does not assume the similarity scoring function between the input and the candidate label. Instead, the input concatenation and a candidate serve as a new input to a non-linear function that scores its match based on the desired dependencies. In (HUMEAU et al., 2019) Poly-encoder caching candidates for a given label, thereby determining a shorter inference time, while the cross-encoder extracts information.

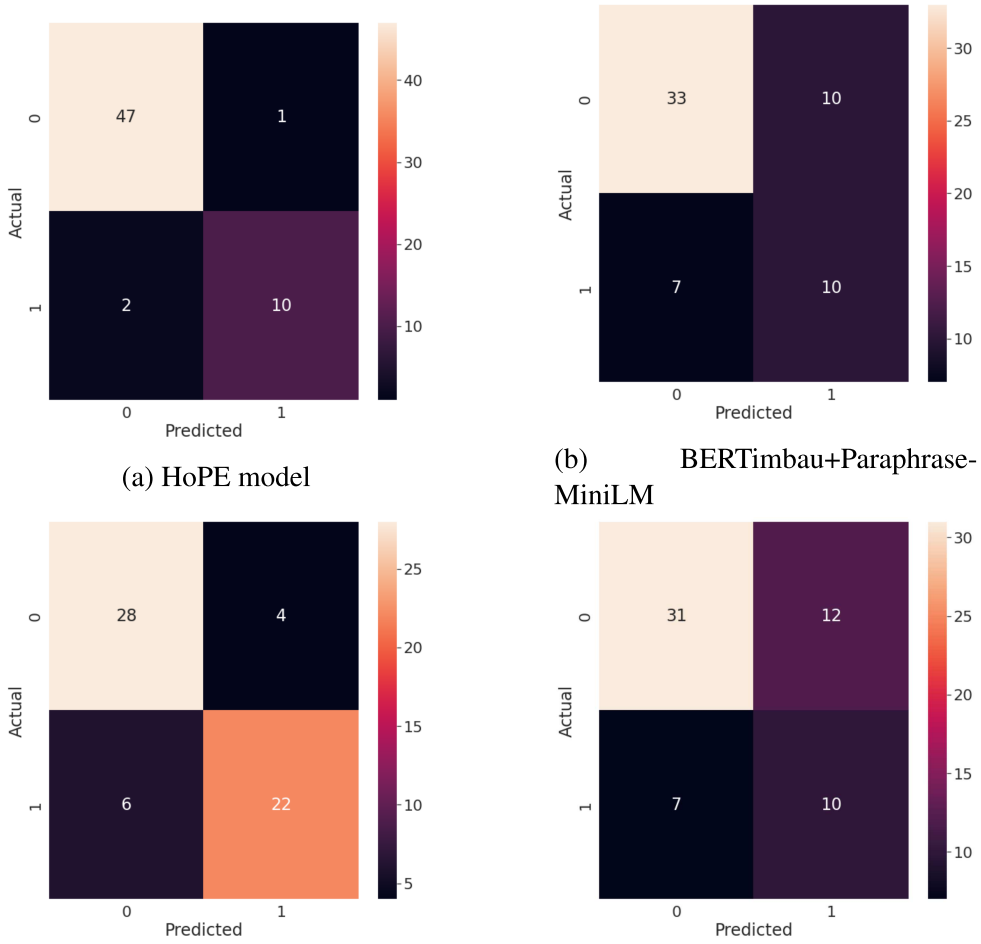
6.5 HoPE architecture evaluation

This section evaluates the HoPE architecture to see how well it performs common tasks for conversational agents and information retrieval systems. The first evaluation sought to determine the accuracy and recall rate of the HoPE model for a test set of sixty phrases. Next, we use three additional models: Sentence-BERTimbau-Paraphrase-Multilingual, which is already included in the HoPE architecture but needs to be tested separately with new input data to verify its performance, and the BM25 Okapi model, which has been used successfully in several studies of information retrieval, and the Paraphrase-Multilingual model, which has achieved great results for semantic search in other studies (HUERTAS-GARCÍA et al., 2021). In Figure 6.5, we show the confusion matrix of the four models.

Through the confusion matrix, we obtained the precision, recall, and F1-Score coefficients that facilitate the interpretation to measure the performance of each model on the data set. The HoPE model presented the best results from the F1-Measure analysis (0.896). With slightly worse performance, the Sentence-BERTimbau-Paraphrase-Multilingual reach an F1-Score of (0.816). The BM25 lexical model obtained a coefficient (0.747) while the paraphrase model reached (0.728). The results can be seen in Figure 20.

The HoPE architecture comprises two main modules: dialog management formed by OntONeo ontology populated with preaching content extracted from pregnancy guidelines and a recently trained bi-cross model on (SBERTimbau + Paraphrase-Multilingual). At first, the evaluation aimed to show the confusion matrix produced for the set of tests with user queries. However, the confusion matrix revealed that the HoPE model offered great precision for user

Figure 39: Confusion Matrix for HoPE and other information retrieval models



(a) HoPE model

(b) BERTimbau+Paraphrase-MiniLM

(c) BM25

(d) Paraphrase-MiniLM

Source: elaborated by authors.

Table 20: Model's accuracy evaluation using K-Folds

Models	Precision	Recall	F1-Score	Std	Max	Min	MRR
HoPE	**0.906**	**0.968**	**0.896**	**1.338**	**0.914**	**0.881**	**0.808**
SBERTimbau+ Paraphrase-Multil.	0.849	0.909	0.816	0.980	0.827	0.803	0.753
BM25 Okapi	0.801	0.847	0.747	1.162	0.764	0.734	0.549
SBERT Paraphrase-Multil.	0.785	0.821	0.728	0.952	0.741	0.716	0.746

queries. Moreover, the model excelled at sorting hits and fallbacks, having a low percentage of false positives and false negatives. The most common classification errors for all tested models were spelling or applied to short sentences.

The HoPE model had better MRR indicators than the other models in the evaluations, with a gain of 5% above the second-best model. This evaluation measured the predictive ability of the model to select the correct answer.

The ambiguity or lack of context associated with short words may cause poor prediction, as BERT models rely on attention mechanisms to identify relationships between the words contained in the sentence (VASWANI et al., 2017). Sentences with rare terms in the corpus also had lower scores (e.g., cereal). The vocabulary application with a greater sentence variability for fine-tuned and pre-processing strategies should improve the co-efficient predictions. Other assumptions raised in other discussions (TAHAMI; GHAJAR; SHAKERY, 2020) (SINGH; SCARTON; BONTCHEVA, 2021). Evaluate the combination of pre-trained bi-encoder models with a cross-encoder layer for this specific task of retrieval information.

All BERT models had similar performance for false negatives. The BM25 model, however, predicted few model hits and an expressive number of fallbacks. The results are understandable, as this is a lexical model not fitted to our data. For instance, because the correct answer was not found in the corpus, the sentence "A newly vaccinated woman must wait before initiating a new pregnancy?" was classified as a fallback in our dataset. On the other hand, models recovered vaccine-related phrases that were not classified as correct responses. We considered the results reasonable because the models were trained on a smaller set of in-domain data. As with (HAN; EISENSTEIN, 2019), future studies may benefit from various examples for more accurate classification.

The (BERTimbau+Paraphrase) model appeared to be a good fit for the presented dataset. The primary distinction between this model and the HoPE architecture is that our architecture employs online paragraph encoding and pre-filtering based on ontology. This type of clustering facilitates classification by eliminating ambiguity in sentences with limited context and reducing the dimensionality of possible answers. The use of ontologies in chatbots already has known positive effects on natural language generation and information retrieval tasks (AVILA et al., 2019)(NAZIR et al., 2019). However, this model combined with convolution networks for conversational agents is still little explored (SENESE et al., 2020) (YOO; JEONG, 2020). The F1-Score coefficient shows a difference from the HoPE architecture compared to models only pre-trained. The results indicate that an approach based on ontology or another clustering strategy combined with information retrieval models can produce satisfactory results. However, we believe that the other models can provide excellent results if trained in a larger set of sentence pairs.

In addition, we assessed the system's ability to detect multiple intentions. We tested this with a new dataset of 10 sentences labeled as single or multiple. The sentences were organized into three levels: low and high complexity. Despite having a 90% hit rate on the dataset, the

Table 21: Evaluation of Single and Multi-intents sentences

N	Complexity	Phrase	Ground T.	Predicted
1	low	What if I have any bleeding during my pregnancy?	Single	Single
2	low	Can I exercise when pregnant?	Single	Single
3	low	What Prenatal Vitamins Should I Take?	Single	Single
4	low	What should I eat when pregnant and What should I drink?	Multiple	Multiple
5	low	What other pre-natal vitamins or supplements do I need?	Single	Single
6	low	Why should I be taking folic acid tablets?	Single	Single
7	low	How much caffeine can I have during pregnancy?	Single	Single
8	high	What about smoking and pregnancy?	Single	Single
9	high	How much exercise should I do? Gym it is a option?	Multiple	Single
10	high	Why do I need a midwife? When should I arrange to see a midwife?	Multiple	Multiple

system failed for one sentence. Therefore, short sentences and ontology-known terms were used in the low-complexity questions. Longer phrases and uncommon terms were used in highly complex phrases. Table 21 shows the results.

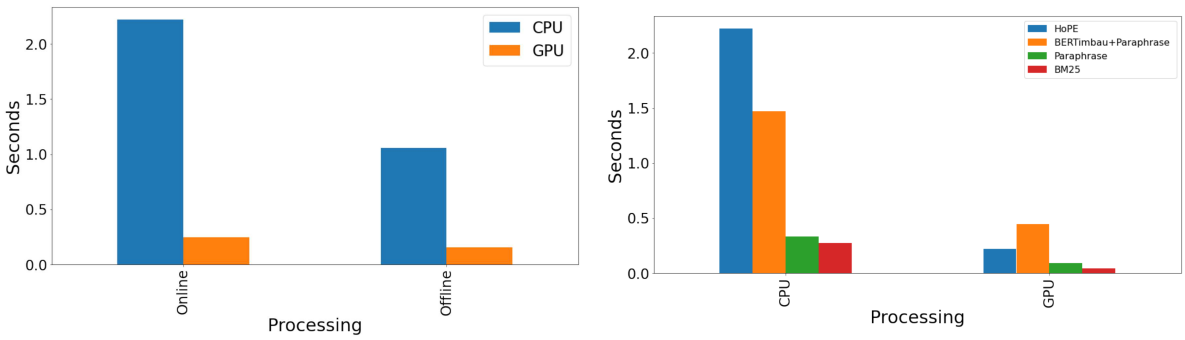
The challenge for the multiple classification intentions is recurrent in some studies involving chatbots. In (ALTINOK, 2018) the classification process is similar to our proposal, using ontology entities to identify the type of question. In this study, the dialogue is driven by a banking ontology developed from named entities, proper names, and verbs for anaphora resolution and disambiguations. The use of pre-defined slots is also a good solution and has shown results (RYCHALSKA; GLABSKA; WROBLEWSKA, 2018).

This HoPE module is also ontology-driven, and it uses the similarity between domains retrieved for each entity to determine whether they have the same user intent or not. We used the BM25 lexical system for its ability to make concise matches via keywords and for being agile in terms of retrieval.

The last experiment was to evaluate the inference speed for each model. For this experiment, we test five phrases from the previous section. The experiments were carried out using a GPU Tesla P100, Intel(R) Xeon(R) CPU @ 2.20GHz, RAM 12.69 GB. To ensure a fair comparison, we applied the same phrases for each experiment to ensure a fair comparison, resetting the kernel after each model inference.

We began by evaluating the HoPE model's encoding process. Sentences containing recognized entities are encoded online. If no entity is found, the pre-loaded and indexed response

Figure 40: HoPE performance for Inference and Encoding time



(a) Encoding Performance

(b) Inference Performance

Source: elaborated by authors.

paragraph embeddings of ANN are used. To understand the time difference between the two methods, we compare them. The results are shown in Figure 6.5.Once compared to a CPU environment, the pre-computed embedding method outperformed the online method by an average of two seconds. When using a GPU environment, the difference retracts.

The results show that the HoPE architecture takes longer to process than other models (CPU = 2.223), with an average difference of two seconds for information retrieval on CPU-powered systems. However, the system improves significantly with GPU (GPU = 0.222) and has a time almost identical to the BERTimbau-Augmented model when used alone.

Loading pre-computed examples seemed to be faster than encoding small datasets online. The HoPE architecture uses both forms within its operation. It was found that even if it has a high volume of indexed paragraphs, the preload approach through some indexing systems such as ANN can provide twice the speed in CPUs and GPUs. This behavior can be observed in other systems when the pre-computing of embedding is up to 10 times faster than online computing (YOON et al., 2020).

However, online encoding improves assertiveness by reducing the amount of content indexed in the prediction model. The encoding of more likely paragraphs for the input query makes it a valid alternative for conversational agents. Also, the GPU system brings a speed gain to the system three times greater. Additionally, we compare the inference times of the models. This type of analysis is critical for conversational agents, as response time is a key factor in end-user engagement (KANKARIA et al., 2021).

Bi-encoder models achieved the best performance, with less than one second query response times. We conducted this experiment using the online encoding of the HoPE model. It performed up to three times slower than other models, with a one-second delay when a GPU was used. Despite the poor timing performance, the HoPE architecture was within a 5-second threshold for inference, considered a benchmark for conversational agents. Despite the poor time performance, the HoPE architecture was within a five-second threshold for a reply, considered a benchmark for conversation agents (INAMDAR; SHIVANAND, 2019).

6.6 Final Remarks

This chapter discusses the experiments proposed in this dissertation. Throughout the clinical studies, groups of pregnant women, health professionals, and doctors had similar and positive perceptions of agents in health. Among the limitations, the low number of users does not allow us to generalize the behavior of users in the two clinical studies. In addition, we restricted to the southern region of Brazil due to ethics committee approval only covering this region. Both studies aimed at work in cost care centers, with the majority of the respondent population being low-income. Thus, although we can better understand the behavior of this population stratum, we do not cover the other income brackets and the impact of conversational agents on this group.

Furthermore, the model's evaluation shows the BERT state-of-art architecture applied in conversational agents context, reaching relevant results over the data augmentation strategy for fine-tuning pregnancy guidelines data. Also, their performance can be enhanced when combined with modeling involving ontologies, as we saw in the HoPE evaluation. We limited the analyses to Sentence-BERT architectures and restricted results to this universe. The experiments tested a few hyperparameters (epochs and learning rates) which could be considered a limitation. More hyperparameters could be tested in future works. We restrict the assessment of the HoPE architecture's assertive capacity to the confusion matrix because this is the most commonly used metric in the context of conversational agents. Other parameters, such as Mean squared Error (MSE) or Mean Reciprocal Rank (MRR), could be also relevant (JEONG et al., 2020). Medical assistants created the test sets based on historical data from pregnant women's questions. More recent or real-time questions were not possible because the experiment format did not allow for this type of evaluation. The intent type recognition module demonstrated its ability to identify composite intents with distant domains in an entity-relationship tree. Its limitations at this stage are determinate recognition of only one or two intentions in the same sentence, making it impossible to identify more intents. Another limitation pointed out is the low number of test phrases used to evaluate this module.

7 CONCLUSION

Qualified and reliable information is a relevant factor for the gestational period. One of the factors that impact the population is the high incidence of errors in decision-making by pregnant women or family members during the pre and postnatal period due to incomplete or incorrect information. These factors make several researchers seek new strategies and solutions to this problem. The conversational agent's development to deliver secure information to pregnant women is a typical target for health computing researchers. The validation with pregnant patients and health professionals happens through many studies, but further investigations are still needed. Another factor for this study type is the contexts where they are applied; after all, climate, economics, and political situations can directly influence the perceptions of participating users.

How the conversational agent interacts and handle with humans are also considered key factors for the success of this type of strategy. For these topics, many researchers have developed new architectures for conversational agents that are friendly, engaging, and capable of finding assertive information for users' questions. While this is happening, there are technological advances in the artificial intelligence field in modeling and techniques that can be valuable in new architectures aimed at pregnant women. As previously stated, the current thesis is justified because it aims to conduct a validation and development study of architecture with the goal of better understanding the perceptions of pregnant women and health professionals in Brazil regarding the use of this tool as a vehicle for the support of reliable information, as well as the evaluation of a new architectural model to achieve greater assertiveness and connection with pregnant women.

The clinical investigation carried out in this thesis involved the conduction of two studies carried out in southern Brazil, with the participation of physicians, health professionals, and pregnant women. We present different methodologies for the two studies, aiming to complement them. The investigation related to the proposed architecture for conversational agents in the gestational period aimed to evaluate state-of-the-art models, sentence disambiguation resolution, and information retrieval by the HoPE architecture. For this reason, the architecture development used a combination of natural language processing techniques, accompanied by a dialog manager that uses the OntONeo ontology in its structure to resolve disambiguations and identify multiple intentions, in addition to a module for information retrieval, via Sentence-BERT networks adjusted to health guidelines data extracted for this dissertation.

Some limitations exist in this thesis investigation. The first is that the clinical studies carried out had a lower number of participants, which makes clear the inability to generalize the results to all territories in Brazil. Although the validation studies used novel methods, they were only tested after seven days of interaction with the final user. The orthographic treatment of the sentences was not carried out, which caused a break in the user's expectation regarding the performance of the conversation agent, as it sometimes failed to identify the intention correctly.

There were also limitations related to the development and evaluation of the HoPE architecture. Retrieval of images and tables present and health guidelines were not considered in this study. The architecture was not tested in a clinical study, only in a testing environment, from sets of questions/answers formulated by physicians and derived from a historical basis of previous experiments with pregnant women.

To reach target objectives, we started the study focused on extracting and curating the content that would serve as the basis for all evaluations. A gynecologist selected the data. We collected, cleaned, and structured the data into paragraphs and sentences. This content served as the foundation for all evaluations used in the dissertation.

We trained the chatbot with the extracted data containing nutritional information for the first clinical study. These data were fed to an NLP platform and used in experiments with pregnant women and health professionals. The study happened in a public institution with low-income pregnant women. The results showed that the perception of pregnant women and health professionals about the usability, learning, satisfaction, and clarity of the content is the same. In addition, there was a stronger desire for buttons to be used for user interaction rather than text. The most relevant variable in this experiment for pregnant women was the content learned, confirming that this is a pertinent construct for people with lower socioeconomic status. According to health professionals such as nutritionists, family doctors, and obstetricians, the chatbot demonstrated assisting pregnant women.

The second clinical study had the same chatbot structure used in the previous experiment, only with the content addition, bringing a more general aspect. However, the participants, the methodology, and the form of evaluation were different. A new pregnant women sample and a sample of doctors were selected. A mixed-method was applied to assess quantitative perceptions of pregnant women and compare them with qualitative insights from the physicians' responses. Both groups corroborated the conversational agent education ability, clarity, and completeness of the information. For pregnant women, it was clear that their doctors would agree with the use of the tool. However, it was evident that some infrastructure failures and the presentation of some contents could be improved. Everyone understood it as a crucial tool for the gestational period for doctors. However, they brought specific improvements such as the insertion of new content and corroborating the pregnant women, better distribution, and presentation of some topics.

The proposed HoPE architecture for conversational agents was developed and evaluated after clinical studies. First, we look at how well Sentence-BERT models could fine-tune data from health guidelines. We found that the models trained using data augmentation strategies performed better, and the augmented BERTimbau model trained in Brazilian Portuguese vocabulary was the best.

In the second part of the evaluation, we focused on incorporating the winning model in the previous stage into the HoPE architecture. The HoPE architecture had an ontology system populated with the conversational agent responses, functioning as a dialogue driver before the

information retrieval task, performed by the model adjusted in the first evaluation. The information retrieval model could also come into play in terms of the user's sentence were not part of the ontology's contents, and thus, the content was retrieved directly by the module. The evaluations involved the type of intention system (single or multiple), which obtained results above 90% of accuracy, the performance of HoPE for information retrieval that we metrified by the F1-Score and obtained a satisfactory result of 0.89, and the inference speed test where the architecture results were regular, taking about 3 seconds for information retrieval.

The contributions generated by this thesis resulted in the publication of two studies (MONTENEGRO; COSTA; RIGHI, 2019)(MONTENEGRO et al., 2018), the first focused on the systematic review of conversational agents in health, promoting different insights into conversational agents and the second provided the first experiments related to the use of information retrieval and semantic techniques for delivering reliable information to postpartum pregnant women. Recently, the study nominated as "The HoPE model architecture: a novel approach to pregnancy information retrieval based on conversational agents" derived from the experiments performed in this thesis was approved to be published by the Journal of Healthcare Informatics and Research (JHIR) [1]. The proposed HoPE architecture for conversational agents emerges in this publication. It comprises a hybrid architecture that employs Transformers models and ontologies tuned to pregnancy guidelines data. The main results demonstrate the ability to improve the assertiveness of conversational agents and the possibility of integrating with NLP engines as a supported architecture for understanding user input sentences. We still have two articles under review related to experiments with pregnant women, doctors, and health professionals that contribute to discussions about usability, satisfaction, health literacy, engagement, approval, anxiety, improvements, and adjustments for conversational agents in the context of pregnancy. Additionally, the study demonstrates the contrasts and similarities between the group's perspectives and the extent to which the topics under discussion affect each individual. In summary, this thesis addressed aspects aimed at the acceptance of conversational agents in the gestational context and the development and validation of an architecture aimed at providing users with reliable information about the thousand days of gestation. The HoPE architecture model aims for assertive information delivery and the best user experience in conversations with conversational agents. For a better end-user experience, we believe that HoPE could integrate conversational agent frameworks and NLP engines. We have several proposals for future work from this dissertation. We intend to cover previously unexplored areas in this dissertation's toward clinical studies to larger populations with diverse demographic strata. Furthermore, apply a study with multimodal interactions for conversational agents. The goal is to analyze voice and text interactions and use spell correction features presented in many tools that support voice NLP. Also, we would like to test the HoPE architecture in a production environment, with humans interacting. In this way, we will be able to detect gaps for improvements and failures so that we can move forward with the possibility of using this architecture in real-life environments for preg-

[1] <https://www.springer.com/journal/41666>

nant women. We would also like to evaluate a disambiguation system proposed in this thesis for the user interactions that are not questionable, but affirmative or imperative. We also intend to apply a complementary study on the multi-intention identification module, to test it for other sentence structures.

REFERENCES

ABDUL-KADER, Sameera A; WOODS, John. Survey on chatbot design techniques in speech conversation systems. **International Journal of Advanced Computer Science and Applications**, Citeseer, v. 6, n. 7, p. 72–80, 2015.

ABRO, Waheed Ahmed et al. Multi-turn intent determination and slot filling with neural networks and regular expressions. **Knowledge-Based Systems**, Elsevier, v. 208, p. 106428, 2020.

ADIKARI, Achini et al. Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare. **Future Generation Computer Systems**, Elsevier, v. 126, p. 318–329, 2022.

ALAMBO, Amanuel et al. Covid-19 and mental health/substance use disorders on reddit: A longitudinal study. In: SPRINGER. **International Conference on Pattern Recognition**. [S.l.], 2021. p. 20–27.

ALESANCO, Álvaro et al. Bots in messaging platforms, a new paradigm in healthcare delivery: application to custom prescription in dermatology. In: **EMBEC & NBC 2017**. [S.l.]: Springer, 2017. p. 185–188.

ALFEO, Antonio L; CIMINO, Mario GCA; VAGLINI, Gigliola. Technological troubleshooting based on sentence embedding with deep transformers. **Journal of Intelligent Manufacturing**, Springer, p. 1–12, 2021.

ALLAOUZI, Imane; AHMED, Mohamed Ben; BENAMROU, Badr. An encoder-decoder model for visual question answering in the medical domain. In: **CLEF (Working Notes)**. [S.l.: s.n.], 2019.

ALMAHRI, Fatima Amer Jid; BELL, David; MERHI, Mohamad. Understanding student acceptance and use of chatbots in the united kingdom universities: a structural equation modelling approach. In: IEEE. **2020 6th International Conference on Information Management (ICIM)**. [S.l.], 2020. p. 284–288.

ALMARWI, Hiba; GHURAB, Mossa; AL-BALTAH, Ibrahim. A hybrid semantic query expansion approach for arabic information retrieval. **Journal of Big Data**, SpringerOpen, v. 7, n. 1, p. 1–19, 2020.

ALOMARI, Ayham et al. Deep reinforcement and transfer learning for abstractive text summarization: A review. **Computer Speech & Language**, Elsevier, p. 101276, 2021.

ALTI, Adel; LAOUAMER, Lamri. Agent-based autonomic semantic context-aware platform for smart health monitoring and disease detection. **The Computer Journal**, 2021.

ALTINOK, Duygu. An ontology-based dialogue management system for banking and finance dialogue systems. **arXiv preprint arXiv:1804.04838**, 2018.

AMATO, Flora et al. Chatbots meet ehealth: automatizing healthcare. v. 14, p. 381–388, 2017.

AMITH, Muhammad et al. Early usability assessment of a conversational agent for hpv vaccination. **Studies in health technology and informatics**, NIH Public Access, v. 257, p. 17, 2019.

AMITH, Muhammad; ROBERTS, Kirk; TAO, Cui. Conceiving an application ontology to model patient human papillomavirus vaccine counseling for dialogue management. **BMC bioinformatics**, BioMed Central, v. 20, n. 21, p. 1–16, 2019.

ARAMAKI, Eiji et al. Vocabulary size in speech may be an early indicator of cognitive impairment. **PloS one**, Public Library of Science, v. 11, n. 5, p. 13, 2016.

AVILA, Caio Viktor S et al. Medibot: An ontology based chatbot for portuguese speakers drug's users. 2019.

BAHJA, Mohammed; ABUHWAILA, Nour; BAHJA, Julia. An antenatal care awareness prototype chatbot application using a user-centric design approach. In: SPRINGER. **International Conference on Human-Computer Interaction**. [S.l.], 2020. p. 20–31.

BAKOUAN, Mamadou et al. A chatbot for automatic processing of learner concerns in an online learning platform. **International Journal of Advanced Computer Science and Applications**, v. 9, n. 5, p. 168–176, 2018.

BANSAL, Himanshu; KHAN, Rizwan. A review paper on human computer interaction. **International Journals of Advanced Research in Computer Science and Software Engineering**, v. 8, p. 53–56, 2018.

BASKAR, Jayalakshmi; LINDGREN, Helena. Cognitive architecture of an agent for human-agent dialogues. In: SPRINGER. **International Conference on Practical Applications of Agents and Multi-Agent Systems**. [S.l.], 2014. p. 89–100.

BEAUDRY, Jeremy et al. Getting ready for adult healthcare: Designing a chatbot to coach adolescents with special health needs through the transitions of care. **Journal of pediatric nursing**, Elsevier, v. 49, p. 85–91, 2019.

BENESTY, Jacob et al. Pearson correlation coefficient. In: **Noise reduction in speech processing**. [S.l.]: Springer, 2009. p. 1–4.

BICKMORE, Timothy; GIORGINO, Toni. Health dialog systems for patients and consumers. **Journal of biomedical informatics**, Elsevier, v. 39, n. 5, p. 556–571, 2006.

BICKMORE, Timothy; SCHULMAN, Daniel; YIN, Langxuan. Engagement vs. deceit: Virtual humans with human autobiographies. In: SPRINGER. **International Workshop on Intelligent Virtual Agents**. [S.l.], 2009. p. 6–19.

BICKMORE, Timothy et al. Promotion of preconception care among adolescents and young adults by conversational agent. **Journal of Adolescent Health**, Elsevier, v. 67, n. 2, p. S45–S51, 2020.

BICKMORE, Timothy W et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. **Journal of health communication**, Taylor & Francis, v. 15, n. S2, p. 197–210, 2010.

BICKMORE, Timothy W; PFEIFER, Laura M; JACK, Brian W. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: ACM. **Proceedings of the SIGCHI conference on human factors in computing systems**. [S.l.], 2009. p. 1265–1274.

BICKMORE, Timothy W; SCHULMAN, Daniel; SIDNER, Candace L. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. **Journal of biomedical informatics**, Elsevier, v. 44, n. 2, p. 183–197, 2011.

BICKMORE, Timothy W et al. Improving access to online health information with conversational agents: a randomized controlled experiment. **Journal of medical Internet research**, JMIR Publications Inc., v. 18, n. 1, 2016.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: " O'Reilly Media, Inc.", 2009.

BJELKE, Maria et al. Using the internet as a source of information during pregnancy—a descriptive cross-sectional study in sweden. **Midwifery**, Elsevier, v. 40, p. 187–191, 2016.

BOONSTRA, Lee. Getting started with dialogflow essentials. In: **The Definitive Guide to Conversational AI with Dialogflow and Google Cloud**. [S.l.]: Springer, 2021. p. 29–57.

BORAH, Bhriguraj et al. Survey of textbased chatbot in perspective of recent technologies. In: SPRINGER. **International Conference on Computational Intelligence, Communications, and Business Analytics**. [S.l.], 2018. p. 84–96.

BOUDJELLAL, Nada et al. Abioner: a bert-based model for arabic biomedical named-entity recognition. **Complexity**, Hindawi, v. 2021, 2021.

BR, Shambhavi; KUMAR, Ramakanth. Kannada part-of-speech tagging with probabilistic classifiers. **international journal of computer applications**, Citeseer, v. 48, n. 17, p. 26–30, 2012.

BRESÓ, Adrián et al. Usability and acceptability assessment of an empathic virtual agent to prevent major depression. **Expert Systems**, Wiley Online Library, v. 33, n. 4, p. 297–312, 2016.

CABEZUDO, Marco Antonio Sobrevilla et al. Nilc at assin 2: Exploring multilingual approaches. In: **ASSIN@ STIL**. [S.l.: s.n.], 2019. p. 49–58.

CARLSSON, Fredrik et al. Semantic re-tuning with contrastive tension. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2021.

CHANG, Wei-Cheng et al. Taming pretrained transformers for extreme multi-label text classification. **arXiv preprint arXiv:1905.02331**, 2019.

CHANG, Yung-Tzung et al. Extending the utility of utaut2 for hospital patients' adoption of medical apps: Moderating effects of e-health literacy. **Mobile Information Systems**, Hindawi, v. 2021, 2021.

CHOI, Hyunjin et al. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In: IEEE. **2020 25th International Conference on Pattern Recognition (ICPR)**. [S.l.], 2021. p. 5482–5487.

CHUAH, Joon Hao et al. Exploring agent physicality and social presence for medical team training. **Presence: Teleoperators and Virtual Environments**, MIT Press, v. 22, n. 2, p. 141–170, 2013.

CHUNG, Kyungmi; CHO, Hee Young; PARK, Jin Young. A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study. **JMIR Medical Informatics**, JMIR Publications Inc., Toronto, Canada, v. 9, n. 3, p. e18607, 2021.

CIAMPI, Mario et al. Some lessons learned using health data literature for smart information retrieval. In: **Proceedings of the 35th Annual ACM Symposium on Applied Computing**. [S.l.: s.n.], 2020. p. 931–934.

COHEN, Kevin Bretonnel; DEMNER-FUSHMAN, Dina. **Biomedical natural language processing**. [S.l.]: John Benjamins Publishing Company, 2014.

COHN, Michelle; CHEN, Chun-Yen; YU, Zhou. A large-scale user study of an alexa prize chatbot: Effect of tts dynamism on perceived quality of social dialog. In: **Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue**. [S.l.: s.n.], 2019. p. 293–306.

CONSORTIUM, World Wide Web et al. Rdf 1.1 concepts and abstract syntax. World Wide Web Consortium, 2014.

CRISS, Shaniece et al. The role of health information sources in decision-making among hispanic mothers during their children's first 1000 days of life. **Maternal and child health journal**, Springer, v. 19, n. 11, p. 2536–2543, 2015.

CROUX, Christophe; DEHON, Catherine. Influence functions of the spearman and kendall correlation measures. **Statistical methods & applications**, Springer, v. 19, n. 4, p. 497–515, 2010.

DAI, Zhenjin et al. Named entity recognition using bert bilstm crf for chinese electronic health records. In: IEEE. **2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)**. [S.l.], 2019. p. 1–5.

D'ALFONSO, Simon et al. Artificial intelligence-assisted online social therapy for youth mental health. **Frontiers in psychology**, Frontiers, v. 8, p. 796, 2017.

DECMAN, Mitja. Factors that increase active participation by higher education students, and predict the acceptance and use of classroom response systems. **International Journal of Higher Education**, ERIC, v. 9, n. 4, p. 84–98, 2020.

DENG, Xinyang et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. **Information Sciences**, Elsevier, v. 340, p. 250–261, 2016.

DEVLIEGER, Roland. Recruiting in the interpregnancy period. In: **Preconceptional Origins of Child Health Outcomes Workshop**. [S.l.: s.n.], 2021. p. 14.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

____. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>.

DIVYA, S et al. A self-diagnosis medical chatbot using artificial intelligence. **Journal of Web Development and Web Designing**, v. 3, n. 1, p. 1–7, 2018.

DOLIANITI, Foteini et al. Chatbots in healthcare curricula: The case of a conversational virtual patient. In: SPRINGER. **International Conference on Brain Function Assessment in Learning**. [S.l.], 2020. p. 137–147.

DUARTE, Paulo; PINHO, José Carlos. A mixed methods utaut2-based approach to assess mobile health adoption. **Journal of Business Research**, Elsevier, v. 102, p. 140–150, 2019.

EDIRISOORIYA, Maninda; MAHAKALANDA, Indra; YAPA, Tolusha. Generalised framework for automated conversational agent design via qfd. In: IEEE. **2019 Moratuwa Engineering Research Conference (MERCon)**. [S.l.], 2019. p. 297–302.

EDWARDS, Roger A et al. Use of an interactive computer agent to support breastfeeding. **Maternal and child health journal**, Springer, v. 17, n. 10, p. 1961–1968, 2013.

EGEDE, Joy et al. Designing an adaptive embodied conversational agent for health literacy: a user study. In: **Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents**. [S.l.: s.n.], 2021. p. 112–119.

EMYGDIO, Jeanne Louize; ALMEIDA, Maurício Barcellos. Representações formais do conhecimento aplicadas à interoperabilidade semântica de terminologias clínicas. **Múltiplos Olhares em Ciência da Informação**, v. 9, n. 2, 2019.

ENGELMANN, Débora et al. Dial4jaca–a communication interface between multi-agent systems and chatbots. In: SPRINGER. **International Conference on Practical Applications of Agents and Multi-Agent Systems**. [S.l.], 2021. p. 77–88.

ESTEVA, Andre et al. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. **NPJ digital medicine**, Nature Publishing Group, v. 4, n. 1, p. 1–9, 2021.

FADHIL, Ahmed; GABRIELLI, Silvia. Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In: ACM. **Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare**. [S.l.], 2017. p. 261–265.

FADHIL, Ahmed; SCHIAVO, Gianluca; WANG, Yunlong. Coachai: A conversational agent assisted health coaching platform. **arXiv preprint arXiv:1904.11961**, 2019.

FADHIL, Ahmed et al. The effect of emojis when interacting with conversational interface assisted health coaching system. In: ACM. **Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare**. [S.l.], 2018. p. 378–383.

FADHIL, Ahmed; WANG, Yunlong; REITERER, Harald. Assistive conversational agent for health coaching: a validation study. **Methods of information in medicine**, Georg Thieme Verlag KG, v. 58, n. 01, p. 009–023, 2019.

FERNANDA, Farinelli. **ONTONEO**. 2018. Disponível em: <http://bioportal.bioontology.org/ontologies/ONTONEO>.

FERNÁNDEZ-MARTINEZ, Fernando et al. An approach to intent detection and classification based on attentive recurrent neural networks. **Proc. IberSPEECH**, p. 46–50, 2021.

FILHO, Jorge A Wagner et al. The brwac corpus: A new open resource for brazilian portuguese. In: **Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)**. [S.l.: s.n.], 2018.

FLORIDI, Luciano; CHIRIATTI, Massimo. Gpt-3: Its nature, scope, limits, and consequences. **Minds and Machines**, Springer, v. 30, n. 4, p. 681–694, 2020.

FONSECA, Cátia Regina Branco da et al. Risk factors for low birth weight in botucatu city, sp state, brazil: a study conducted in the public health system from 2004 to 2008. **BMC research notes**, Springer, v. 5, n. 1, p. 1–9, 2012.

FRASER, Kathleen C; MELTZER, Jed A; RUDZICZ, Frank. Linguistic features identify alzheimer's disease in narrative speech. **Journal of Alzheimer's Disease**, IOS Press, v. 49, n. 2, p. 407–422, 2016.

FUSHIKI, Tadayoshi. Estimation of prediction error by using k-fold cross-validation. **Statistics and Computing**, Springer, v. 21, n. 2, p. 137–146, 2011.

GALVAO, Adjamir M et al. Persona-aiml: An architecture developing chatterbots with personality. In: IEEE COMPUTER SOCIETY. **Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3**. [S.l.], 2004. p. 1266–1267.

GANESH, Surya et al. Statistical transliteration for cross language information retrieval using hmm alignment model and crf. In: **Proceedings of the 2nd Workshop on Cross Lingual Information Access**. [S.l.: s.n.], 2008.

GANHOTRA, Jatin et al. Conversational document prediction to assist customer care agents. **arXiv preprint arXiv:2010.02305**, 2020.

GARDINER, Paula et al. Reaching women through health information technology: the gabby preconception care system. **American Journal of Health Promotion**, SAGE Publications Sage CA: Los Angeles, CA, v. 27, n. 3_suppl, p. eS11–eS20, 2013.

GOEL, Parth; GANATRA, Amit. A survey on chatbot: Futuristic conversational agent for user interaction. In: IEEE. **2021 3rd International Conference on Signal Processing and Communication (ICPSC)**. [S.l.], 2021. p. 736–740.

GOEURIOT, Lorraine et al. Medical information retrieval: introduction to the special issue. **Information Retrieval Journal**, Springer, v. 19, n. 1-2, p. 1–5, 2016.

GOLDENTHAL, Steven B et al. Assessing the feasibility of a chatbot after ureteroscopy. **mHealth**, AME Publications, v. 5, 2019.

GOOGLE. **Google Scholar Metrics**. Google, 2019. 1 p. https://scholar.google.com/intl/en/scholar/metrics.htmlmetrics. Disponível em: <https://scholar.google.com/intl/en/scholar/metrics.html#metrics>. Acesso em: 20 jan. 2019.

GRECHE, Latifa et al. Comparison between euclidean and manhattan distance measure for facial expressions classification. In: IEEE. **2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)**. [S.l.], 2017. p. 1–4.

GUO, Jiafeng et al. A deep look into neural ranking models for information retrieval. **Information Processing & Management**, Elsevier, v. 57, n. 6, p. 102067, 2020.

HAN, Xiaochuang; EISENSTEIN, Jacob. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. **arXiv preprint arXiv:1904.02817**, 2019.

HASAN, Mehedi et al. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. **Journal of biomedical informatics**, Elsevier, v. 62, p. 21–31, 2016.

HECKSTEDEN, Anne et al. How to construct, conduct and analyze an exercise training study? **Frontiers in physiology**, Frontiers, v. 9, p. 1007, 2018.

HEERDEN, Alastair van; NTINGA, Xolani; VILAKAZI, Khanya. The potential of conversational agents to provide a rapid hiv counseling and testing services. In: IEEE. **the Frontiers and Advances in Data Science (FADS), 2017 International Conference on**. [S.l.], 2017. p. 80–85.

HEERINK, Marcel et al. Relating conversational expressiveness to social presence and acceptance of an assistive social robot. **Virtual reality**, Springer, v. 14, n. 1, p. 77–84, 2010.

HENDERSON, Matthew et al. Efficient natural language response suggestion for smart reply. **arXiv preprint arXiv:1705.00652**, 2017.

HERBERT, David; KANG, Byeong Ho. Intelligent conversation system using multiple classification ripple down rules and conversational context. **Expert Systems with Applications**, Elsevier, 2018.

HERSH, William; HERSH; WESTON. **Information retrieval: A biomedical and health perspective**. [S.l.]: Springer, 2020.

HIRANO, Mari et al. Designing behavioral self-regulation application for preventive personal mental healthcare. **Health psychology open**, SAGE Publications Sage UK: London, England, v. 4, n. 1, p. 2055102917707185, 2017.

HU, Qian et al. Collaborative data relabeling for robust and diverse voice apps recommendation in intelligent personal assistants. In: **Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI**. [S.l.: s.n.], 2021. p. 113–119.

HUDLICKA, Eva. Virtual training and coaching of health behavior: Example from mindfulness meditation training. **Patient education and counseling**, Elsevier, v. 92, n. 2, p. 160–166, 2013.

HUERTAS-GARCÍA, Álvaro et al. Countering misinformation through semantic-aware multilingual models. In: SPRINGER. **International Conference on Intelligent Data Engineering and Automated Learning**. [S.l.], 2021. p. 312–323.

HUMEAU, Samuel et al. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. **arXiv preprint arXiv:1905.01969**, 2019.

HUSSAIN, Shafquat; ATHULA, Ginige. Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In: IEEE. **2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)**. [S.l.], 2018. p. 698–703.

IFTENE, Adrian; VANDERDONCKT, Jean. Moocbuddy: a chatbot for personalized learning with moocs. In: **RoCHI–International Conference on Human-Computer Interaction**. [S.l.]: Public Library of Science, 2016. p. 91.

INAMDAR, Vaishnavi Ajay; SHIVANAND, RD. Development of college enquiry chatbot using snatchbot. **DEVELOPMENT**, v. 6, n. 07, 2019.

IOVINE, Andrea; NARDUCCI, Fedelucio; SEMERARO, Giovanni. Conversational recommender systems and natural language:: A study through the converse framework. **Decision Support Systems**, Elsevier, v. 131, p. 113250, 2020.

JACK, Brian et al. Reducing preconception risks among african american women with conversational agent technology. **The Journal of the American Board of Family Medicine**, Am Board Family Med, v. 28, n. 4, p. 441–451, 2015.

JACK, Brian W et al. Improving the health of young african american women in the preconception period using health information technology: a randomised controlled trial. **The Lancet Digital Health**, Elsevier, v. 2, n. 9, p. e475–e485, 2020.

JAIN, Aditya; KULKARNI, Gandhar; SHAH, Vraj. Natural language processing. **Int. J. Comput. Sci. Eng**, v. 6, n. 1, 2018.

JAMEEL, Umar; ANWAR, Aqib; KHAN, Hashim. Doctor recommendation chatbot: A research study: Doctor recommendation chatbot. **Journal of Applied Artificial Intelligence**, v. 2, n. 1, 2021.

JANG, Sooah et al. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. **International Journal of Medical Informatics**, Elsevier, v. 150, p. 104440, 2021.

JEONG, Chanwoo et al. A context-aware citation recommendation model with bert and graph convolutional networks. **Scientometrics**, Springer, v. 124, n. 3, p. 1907–1922, 2020.

JIN, Lifeng et al. Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system. In: **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**. [S.l.: s.n.], 2017. p. 11–21.

JOHNSON, W; LABORE, Catherine; CHIU, Yuan-Chun. A pedagogical agent for psychosocial intervention on a handheld computer. In: **AAAI Fall Symposium on Dialogue Systems for Health Communication**. [S.l.: s.n.], 2004. p. 22–24.

JÚNIOR, André Eduardo da Silva et al. Tendência do estado nutricional de gestantes adolescentes beneficiárias do programa de transferência condicionada de renda brasileiro bolsa família no período 2008-2018. **Ciência & Saúde Coletiva**, SciELO Brasil, v. 26, p. 2613–2624, 2021.

KADRI, Youssef; NIE, Jian-Yun. Effective stemming for arabic information retrieval. In: **Proceedings of the challenge of arabic for NLP/MT conference, Londres, Royaume-Uni**. [S.l.: s.n.], 2006. p. 68–74.

KANKARIA, Romit Vinod et al. Raah. ai: An interactive chatbot for stress relief using deep learning and natural language processing. In: IEEE. **2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. [S.l.], 2021. p. 1–7.

KARPAGAM, K; SARADHA, A. An intelligent conversation agent for health care domain. **ICTACT Journal on Soft Computing**, v. 4, n. 3, 2014.

KASINATHAN, Vinothini et al. Intelligent healthcare chatterbot (hecia): Case study of medical center in malaysia. In: IEEE. **Open Systems (ICOS), 2017 IEEE Conference on**. [S.l.], 2017. p. 32–37.

KHATTAB, Omar; ZAHARIA, Matei. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: **Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval**. [S.l.: s.n.], 2020. p. 39–48.

KILMER, Shelby; MARSHALL, Caleb; SENGER, Steven. Dot product chains. **arXiv preprint arXiv:2006.11467**, 2020.

KIM, Taeuk; YOO, Kang Min; LEE, Sang-goo. Self-guided contrastive learning for bert sentence representations. **arXiv preprint arXiv:2106.07345**, 2021.

KING, Abby C et al. Employing virtual advisors in preventive care for underserved communities: results from the compass study. **Journal of health communication**, Taylor & Francis, v. 18, n. 12, p. 1449–1464, 2013.

KOWATSCH, Tobias et al. Text-based healthcare chatbots supporting patient and health professional teams: Preliminary results of a randomized controlled trial on childhood obesity. In: ETH ZURICH. **Persuasive Embodied Agents for Behavior Change (PEACH2017)**. [S.l.], 2017.

LAGAN, Briege M; SINCLAIR, Marlene; KERNOHAN, W George. Internet use in pregnancy informs women's decision making: a web-based survey. **Birth**, Wiley Online Library, v. 37, n. 2, p. 106–115, 2010.

LARKEY, Leah S; BALLESTEROS, Lisa; CONNELL, Margaret E. Improving stemming for arabic information retrieval: light stemming and co-occurrence analysis. In: **Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.: s.n.], 2002. p. 275–282.

LEE, Jinhyuk et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, Oxford University Press, v. 36, n. 4, p. 1234–1240, 2020.

LI, Bohan et al. On the sentence embeddings from pre-trained language models. **arXiv preprint arXiv:2011.05864**, 2020.

LI, Deqing; ZENG, Wenyi. Distance measure of pythagorean fuzzy sets. **International journal of intelligent systems**, Wiley Online Library, v. 33, n. 2, p. 348–361, 2018.

LIAO, Zhibin et al. Medical data inquiry using a question answering model. In: IEEE. **2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)**. [S.l.], 2020. p. 1490–1493.

LISETTI, Christine et al. I can help you change! an empathic virtual agent delivers behavior change health interventions. **ACM Transactions on Management Information Systems (TMIS)**, ACM, v. 4, n. 4, p. 19, 2013.

LISETTI, Christine L et al. Building an on-demand avatar-based health intervention for behavior change. In: **FLAIRS Conference**. [S.l.: s.n.], 2012.

LIU, Xiaodong et al. The microsoft toolkit of multi-task deep neural networks for natural language understanding. **arXiv preprint arXiv:2002.07972**, 2020.

LOKMAN, Abbas Saliimi; ZAIN, Jasni Mohamad. Chatbot enhanced algorithms: A case study on implementation in bahasa malaysia human language. In: SPRINGER. **International Conference on Networked Digital Technologies**. [S.l.], 2010. p. 31–44.

LÓPEZ, Víctor; EISMAN, Eduardo M; CASTRO, Juan Luis. A tool for training primary health care medical students: The virtual simulated patient. In: IEEE. **Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on**. [S.l.], 2008. v. 2, p. 194–201.

LV, Yuanhua; ZHAI, ChengXiang. When documents are very long, bm25 fails! In: **Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval**. [S.l.: s.n.], 2011. p. 1103–1104.

MAARUP, Mercedes et al. Radical technological innovation and perception: A non-physician practitioners' perspective. **ICH ITA**, p. 45, 2019.

MAEDA, Eri et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. **Reproductive BioMedicine Online**, Elsevier, v. 41, n. 6, p. 1133–1143, 2020.

MAGERKO, Brian et al. Dr. vicky: A virtual coach for learning brief negotiated interview techniques for treating emergency room patients. In: **AAAI Spring Symposium: AI and Health Communication**. [S.l.: s.n.], 2011.

MALIZIA, Alessio et al. Sema4a: An ontology for emergency notification systems accessibility. **Expert systems with applications**, Elsevier, v. 37, n. 4, p. 3380–3391, 2010.

MASS, Yosi; ROITMAN, Haggai. Ad-hoc document retrieval using weak-supervision with bert and gpt2. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2020. p. 4191–4197.

MCARTHUR, Stephen DJ et al. Multi-agent systems for power engineering applications—part i: Concepts, approaches, and technical challenges. **IEEE Transactions on Power systems**, IEEE, v. 22, n. 4, p. 1743–1752, 2007.

MELLADO-SILVA, Rafael; FAÚNDEZ-UGALDE, Antonio; LOBOS, María Blanco. Learning tax regulations through rules-based chatbots using decision trees: a case study at the time of covid-19. In: IEEE. **2020 39th International Conference of the Chilean Computer Science Society (SCCC)**. [S.l.], 2020. p. 1–8.

MIKOLOV, Tomáš; YIH, Wen-tau; ZWEIG, Geoffrey. Linguistic regularities in continuous space word representations. In: **Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies**. [S.l.: s.n.], 2013. p. 746–751.

MILITELLO, Lisa et al. Delivering perinatal health information via a voice interactive app (smile): Mixed methods feasibility study. **JMIR formative research**, JMIR Publications Inc., Toronto, Canada, v. 5, n. 3, p. e18240, 2021.

MNASRI, Maali. Recent advances in conversational nlp: Towards the standardization of chatbot building. **arXiv preprint arXiv:1903.09025**, 2019.

MOKMIN, Nur Azlina Mohamed; IBRAHIM, Nurul Anwar. The evaluation of chatbot as a tool for health literacy education among undergraduate students. **Education and Information Technologies**, Springer, p. 1–17, 2021.

MONTENEGRO, João Luis Zeni et al. A proposal for postpartum support based on natural language generation model. In: IEEE. **2018 International Conference on Computational Science and Computational Intelligence (CSCI)**. [S.l.], 2018. p. 756–759.

MONTENEGRO, Joao Luis Zeni; COSTA, Cristiano André da; RIGHI, Rodrigo da Rosa. Survey of conversational agents in health. **Expert Systems with Applications**, Elsevier, v. 129, p. 56–67, 2019.

MORA, Sara et al. A nlp pipeline for the automatic extraction of microorganisms names from microbiological notes. In: **pHealth 2021**. [S.l.]: IOS Press, 2021. p. 153–158.

MOYER, Cheryl A et al. Pregnancy-related anxiety during covid-19: a nationwide survey of 2740 pregnant women. **Archives of Women's Mental Health**, Springer, v. 23, n. 6, p. 757–765, 2020.

MUGOYE, Kevin; OKOYO, Henry; MCOYOWO, Sylvester. Smart-bot technology: Conversational agents role in maternal healthcare support. In: IEEE. **2019 IST-Africa Week Conference (IST-Africa)**. [S.l.], 2019. p. 1–7.

MUKHERJEE, Subhabrata et al. Clues: Few-shot learning evaluation in natural language understanding. **arXiv preprint arXiv:2111.02570**, 2021.

MURAKAMI, Kentaro et al. Education, but not occupation or household income, is positively related to favorable dietary intake patterns in pregnant japanese women: the osaka maternal and child health study. **Nutrition Research**, Elsevier, v. 29, n. 3, p. 164–172, 2009.

NAWABI, Farah et al. Health literacy in pregnant women: A systematic review. **International journal of environmental research and public health**, Multidisciplinary Digital Publishing Institute, v. 18, n. 7, p. 3847, 2021.

NAZIR, Aisha et al. A novel approach for ontology-driven information retrieving chatbot for fashion brands. **Int. J. Adv. Comput. Sci. Appl. IJACSA**, v. 10, n. 9, 2019.

NGUYEN, Tri et al. Ms marco: A human generated machine reading comprehension dataset. In: **CoCo@ NIPS**. [S.l.: s.n.], 2016.

NI, Lin et al. Mandy: Towards a smart primary care chatbot application. In: SPRINGER. **International Symposium on Knowledge and Systems Sciences**. [S.l.], 2017. p. 38–52.

NIKFARJAM, Azadeh et al. Early detection of adverse drug reactions in social health networks: a natural language processing pipeline for signal detection. **JMIR public health and surveillance**, JMIR Publications Inc., Toronto, Canada, v. 5, n. 2, p. e11264, 2019.

NIKITINA, Svetlana; CALLAIOLI, Sara; BAEZ, Marcos. Smart conversational agents for reminiscence. **arXiv preprint arXiv:1804.06550**, ACM, v. 1, n. 4, p. 6, 2018.

NOURI, Sarah S; RUDD, Rima E. Health literacy in the “oral exchange”: An important element of patient–provider communication. **Patient education and counseling**, Elsevier, v. 98, n. 5, p. 565–571, 2015.

NOVIELLI, Nicole. Hmm modeling of user engagement in advice-giving dialogues. **Journal on Multimodal User Interfaces**, Springer, v. 3, n. 1-2, p. 131–140, 2010.

NOY, Natalya F; MCGUINNESS, Deborah L et al. **Ontology development 101: A guide to creating your first ontology**. [S.l.]: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA, 2001.

OCHS, Magalie et al. An architecture of virtual patient simulation platform to train doctor to break bad news. In: **Conference on Computer Animation and Social Agents (CASA)**. [S.l.: s.n.], 2017.

PADAKI, Ramith; DAI, Zhuyun; CALLAN, Jamie. Rethinking query expansion for bert reranking. In: SPRINGER. **European Conference on Information Retrieval**. [S.l.], 2020. p. 297–304.

PALANICA, Adam et al. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 21, n. 4, p. e12887, 2019.

PATEL, Vikram et al. The lancet's series on global mental health: 1 year on. **The Lancet**, Elsevier, v. 372, n. 9646, p. 1354–1357, 2008.

PETERS, Matthew E et al. Deep contextualized word representations. **arXiv preprint arXiv:1802.05365**, 2018.

PETTICREW, Mark; ROBERTS, Helen. **Systematic reviews in the social sciences: A practical guide**. [S.l.]: John Wiley & Sons, 2008.

PICALAUSA, François; VANSUMMEREN, Stijn. What are real sparql queries like? In: **Proceedings of the International Workshop on Semantic Web Information Management**. [S.l.: s.n.], 2011. p. 1–6.

PODGORNY, Igor; KHABURZANIYA, Yason; GEISLER, Jeff. Conversational agents and community question answering. In: **CHI 2019 Workshops, Glasgow, United Kingdom**. [S.l.: s.n.], 2019.

POPOVA, Svetlana et al. Alcohol industry–funded websites contribute to ambiguity regarding the harmful effects of alcohol consumption during pregnancy: A commentary on lim et al.(2019). **Journal of studies on alcohol and drugs**, Rutgers University, v. 80, n. 5, p. 534–536, 2019.

QIAN, Yuhua; LIANG, Jiye; DANG, Chuangyin. Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. **International Journal of Approximate Reasoning**, Elsevier, v. 50, n. 1, p. 174–188, 2009.

QUAMAR, Abdul et al. Conversational bi: an ontology-driven conversation system for business intelligence applications. **Proceedings of the VLDB Endowment**, VLDB Endowment, v. 13, n. 12, p. 3369–3381, 2020.

RAGAB, Mahmoud et al. Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques. **Wireless Communications and Mobile Computing**, Hindawi, v. 2021, 2021.

RAHMANI, Narges et al. Comparison of health literacy between the pregnant women referring to health care centers and those referring to private offices. **Journal of Health Literacy**, Mashhad University of Medical Sciences. Iranian Association of Health . . . , v. 4, n. 2, p. 35–43, 2019.

RAJI, PS; SURENDRAN, Subu. Rdf approach on social network analysis. In: IEEE. **2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS)**. [S.l.], 2016. p. 1–4.

RAJOSOA, Mickael et al. Hybrid question answering system based on natural language processing and sparql query. 2019.

RAJPURKAR, Pranav et al. **SQuAD: 100,000+ Questions for Machine Comprehension of Text**. 2016.

RAMESH, Kiran et al. A survey of design techniques for conversational agents. In: SPRINGER. **International conference on information, communication and computing technology**. [S.l.], 2017. p. 336–350.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

REN, Jin; WANG, Hengsheng; LIU, Tong. Information retrieval based on knowledge-enhanced word embedding through dialog: A case study. **International Journal of Computational Intelligence Systems**, Atlantis Press, v. 13, n. 1, p. 275–290, 2020.

RENTEA, Victor et al. Prevention assistant–risk evaluation based on sparse data. In: IEEE. **Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on**. [S.l.], 2012. p. 158–165.

RING, Lazlo; BICKMORE, Timothy; PEDRELLI, Paola. An affectively aware virtual therapist for depression counseling. In: **ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health**. [S.l.: s.n.], 2016.

____. Real-time tailoring of depression counseling by conversational agent. **Iproceedings**, JMIR Publications Inc., Toronto, Canada, v. 2, n. 1, p. e27, 2016.

RIZZO, Albert; KENNY, Patrick; PARSONS, Thomas D. Intelligent virtual patients for training clinical skills. **JVRB-Journal of Virtual Reality and Broadcasting**, v. 8, n. 3, p. 9, 2011.

RODRIGUES, Rui; COUTO, Paula; RODRIGUES, Irene. Ipr: The semantic textual similarity and recognizing textual entailment systems. In: **ASSIN@ STIL**. [S.l.: s.n.], 2019. p. 39–48.

RODRIGUES, Ruan Chaves et al. Portuguese language models and word embeddings: Evaluating on semantic similarity tasks. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2020. p. 239–248.

ROGERS, Anna; KOVALEVA, Olga; RUMSHISKY, Anna. A primer in bertology: What we know about how bert works. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 8, p. 842–866, 2021.

ROQUE, Geicianfran da Silva Lima et al. Content validation and usability of a chatbot of guidelines for wound dressing. **International Journal of Medical Informatics**, Elsevier, v. 151, p. 104473, 2021.

RUBINSTEIN, Aviad. Hardness of approximate nearest neighbor search. In: **Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing**. [S.l.: s.n.], 2018. p. 1260–1268.

RYCHALSKA, Barbara; GLABSKA, Helena; WROBLEWSKA, Anna. Multi-intent hierarchical natural language understanding for chatbots. In: IEEE. **2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.], 2018. p. 256–259.

SAFAIE, Shayna Zade et al. The relationship between maternal health literacy and pregnancy outcome in postnatal wards. **J Biochem Tech**, 2019.

SAFDER, Iqra; HASSAN, Saeed-Ul. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. **Scientometrics**, Springer, v. 119, n. 1, p. 257–277, 2019.

SAMIMI, Parnia; RAVANA, Sri Devi. Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: a review. **The Scientific World Journal**, Hindawi, v. 2014, 2014.

SANKHAVARA, Jainisha. Feature weighting in finding feedback documents for query expansion in biomedical document retrieval. **SN Computer Science**, Springer, v. 1, n. 2, p. 1–7, 2020.

SARICA, Serhad; LUO, Jianxi. Stopwords in technical language processing. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 8, p. e0254937, 2021.

SARWAR, Shihab et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. **NPJ digital medicine**, Nature Publishing Group, v. 2, n. 1, p. 1–7, 2019.

SAYAKHOT, Padaphet; CAROLAN-OLAH, Mary. Internet use by pregnant women seeking pregnancy-related information: a systematic review. **BMC pregnancy and childbirth**, BioMed Central, v. 16, n. 1, p. 1–10, 2016.

SEBASTIAN, Joel; RICHARDS, Deborah. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. **Computers in Human Behavior**, Elsevier, v. 73, p. 479–488, 2017.

SENESE, Matteo Antonio et al. Mtsi-bert: A session-aware knowledge-based conversational agent. In: **Proceedings of The 12th Language Resources and Evaluation Conference**. [S.l.: s.n.], 2020. p. 717–725.

SHAH, Chirag; CROFT, W Bruce. Evaluating high accuracy retrieval techniques. In: **Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.: s.n.], 2004. p. 2–9.

SHAKED, Nava A. Avatars and virtual agents–relationship interfaces for the elderly. **Healthcare technology letters**, IET, v. 4, n. 3, p. 83, 2017.

SHAWAR, B Abu; ATWELL, Eric. Arabic question-answering via instance based learning from an faq corpus. In: **Proceedings of the CL 2009 International Conference on Corpus Linguistics. UCREL**. [S.l.: s.n.], 2009. v. 386.

SHEVAT, Amir. **Designing bots: Creating conversational experiences**. [S.l.]: " O'Reilly Media, Inc.", 2017.

SINGH, Iknoor; SCARTON, Carolina; BONTCHEVA, Kalina. Multistage bicross encoder: Team gate entry for mlia multilingual semantic search task 2. **arXiv preprint arXiv:2101.03013**, 2021.

SINGH, Sushant; MAHMOOD, Ausif. The nlp cookbook: Modern recipes for transformer based deep learning architectures. **IEEE Access**, IEEE, v. 9, p. 68675–68702, 2021.

SONG, Xinmeng; XIONG, Ting. A survey of published literature on conversational artificial intelligence. In: IEEE. **2021 7th International Conference on Information Management (ICIM)**. [S.l.], 2021. p. 113–117.

SOPHIA, J Jinu et al. A survey on chatbot implementation in health care using nltk. **Int. J. Comput. Sci. Mob. Comput**, v. 9, 2020.

SOUZA, João Vitor Andrioli de et al. Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2020. p. 357–367.

TAHAMI, Amir Vakili; GHAJAR, Kamyar; SHAKERY, Azadeh. Distilling knowledge for fast retrieval-based chat-bots. In: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2020. p. 2081–2084.

TANAKA, Hiroki et al. Detecting dementia through interactive computer avatars. **IEEE journal of translational engineering in health and medicine**, IEEE, v. 5, p. 1–11, 2017.

____. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. **PloS one**, Public Library of Science, v. 12, n. 8, p. 15, 2017.

TANANA, Michael J et al. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 21, n. 7, p. e12529, 2019.

TEIXEIRA, Milene Santos; MARAN, Vinícius; DRAGONI, Mauro. The interplay of a conversational ontology and ai planning for health dialogue management. In: **Proceedings of the 36th Annual ACM Symposium on Applied Computing**. [S.l.: s.n.], 2021. p. 611–619.

THAKUR, Nandan et al. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. **arXiv preprint arXiv:2010.08240**, 2020.

TIELMAN, Myrthe L et al. How should a virtual agent present psychoeducation? influence of verbal and textual presentation on adherence. **Technology and Health Care**, IOS Press, n. Preprint, p. 1–16, 2017.

TOKUNAGA, Seiki et al. Deploying service integration agent for personalized smart elderly care. In: IEEE. **Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on**. [S.l.], 2016. p. 1–6.

TRAYLOR, Claire S et al. Effects of psychological stress on adverse pregnancy outcomes and non-pharmacologic approaches for reduction: an expert review. **American Journal of Obstetrics & Gynecology MFM**, Elsevier, p. 100229, 2020.

TRIVEDI, Sapna et al. Evaluation of a concept mapping task using named entity recognition and normalization in unstructured clinical text. **Journal of Healthcare Informatics Research**, Springer, v. 4, n. 4, p. 395–410, 2020.

TURUNEN, Markku et al. Multimodal and mobile conversational health and fitness companions. **Computer Speech & Language**, Elsevier, v. 25, n. 2, p. 192–209, 2011.

UGURLU, MUHITTIN; ORHAN, HIKMET. Knowledge, attitude and practices of dentists about oral health care during pregnancy: A cross-sectional study from turkey. **Journal of Clinical & Diagnostic Research**, v. 13, n. 4, 2019.

VAIRA, Lucia et al. Mamabot: a system based on ml and nlp for supporting women and families during pregnancy. In: **Proceedings of the 22nd International Database Engineering & Applications Symposium**. [S.l.: s.n.], 2018. p. 273–277.

VASWANI, Ashish et al. Attention is all you need. **arXiv preprint arXiv:1706.03762**, 2017.

VENKATESH, Viswanath; THONG, James YL; XU, Xin. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. **MIS quarterly**, JSTOR, p. 157–178, 2012.

WANG, Alex et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: **International Conference on Learning Representations**. [s.n.], 2019. Disponível em: <https://openreview.net/forum?id=rJ4km2R5t7>.

WANG, Catharine et al. Acceptability and feasibility of a virtual counselor (vicky) to collect family health histories. **Genetics in Medicine**, Nature Publishing Group, v. 17, n. 10, p. 822, 2015.

WANG, Kun et al. Medical question retrieval based on siamese neural network and transfer learning method. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. [S.l.], 2019. p. 49–64.

WANG, Ruyi et al. Supervised machine learning chatbots for perinatal mental healthcare. In: IEEE. **2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)**. [S.l.], 2020. p. 378–383.

WANG, Wenhui et al. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. **Advances in Neural Information Processing Systems**, v. 33, p. 5776–5788, 2020.

WANG, Yuxia et al. Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. In: **Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing**. [S.l.: s.n.], 2020. p. 105–111.

WANG, Yanshan et al. A comparison of word embeddings for the biomedical natural language processing. **Journal of biomedical informatics**, Elsevier, v. 87, p. 12–20, 2018.

____. An ensemble model of clinical information extraction and information retrieval for clinical decision support. In: **TREC**. [S.l.: s.n.], 2016.

WARGNIER, Pierre et al. Field evaluation with cognitively-impaired older adults of attention management in the embodied conversational agent louise. In: IEEE. **Serious Games and Applications for Health (SeGAH), 2016 IEEE International Conference on**. [S.l.], 2016. p. 1–8.

____. Towards attention monitoring of older adults with cognitive impairment during interaction with an embodied conversational agent. In: IEEE. **Virtual and Augmented Assistive Technology (VAAT), 2015 3rd IEEE VR International Workshop on**. [S.l.], 2015. p. 23–28.

WELLS, Kristen J et al. Acceptability of an embodied conversational agent-based computer application for hispanic women. **Hispanic health care international: the official journal of the National Association of Hispanic Nurses**, NIH Public Access, v. 13, n. 4, p. 179, 2015.

WINTER, Joost FC de; DODOU, Dimitra. Five-point likert items: t test versus mann-whitney-wilcoxon (addendum added october 2012). **Practical Assessment, Research, and Evaluation**, v. 15, n. 1, p. 11, 2010.

XIE, Yuqing et al. Distant supervision for multi-stage fine-tuning in retrieval-based question answering. In: **Proceedings of The Web Conference 2020**. [S.l.: s.n.], 2020. p. 2934–2940.

YAGHOUBZADEH, Ramin et al. Virtual agents as daily assistants for elderly or cognitively impaired people. In: SPRINGER. **International Workshop on Intelligent Virtual Agents**. [S.l.], 2013. p. 79–91.

YAN, Jun; LIU, Ning; CHEN, Zheng. **Learning user intent from rule-based training data**. [S.l.]: Google Patents, 2011. US Patent App. 12/783,457.

YANG, Wei et al. End-to-end open-domain question answering with bertserini. **arXiv preprint arXiv:1902.01718**, 2019.

____. Data augmentation for bert fine-tuning in open-domain question answering. **arXiv preprint arXiv:1904.06652**, 2019.

YASAVUR, Ugan; LISETTI, Christine; RISHE, Naphtali. Intelligent virtual agents and spoken dialog systems come together to deliver brief health interventions. **Journal on Multimodal User Interfaces, in press**, v. 1, n. 1, p. 19, 2014.

YIN, Xiaoya et al. Improving sentence representations via component focusing. **Applied Sciences**, Multidisciplinary Digital Publishing Institute, v. 10, n. 3, p. 958, 2020.

YOO, SoYeop; JEONG, OkRan. An intelligent chatbot utilizing bert model and knowledge graph. **Journal of Society for e-Business Studies**, v. 24, n. 3, 2020.

YOON, Sangwoong et al. Image-to-image retrieval by learning similarity between scene graphs. **arXiv preprint arXiv:2012.14700**, 2020.

ZAIB, Munazza; SHENG, Quan Z; ZHANG, Wei Emma. A short survey of pre-trained language models for conversational ai-a new age in nlp. In: **Proceedings of the Australasian Computer Science Week Multiconference**. [S.l.: s.n.], 2020. p. 1–4.

ZAVERI, Amrapali et al. Quality assessment for linked data: A survey. **Semantic Web**, IOS Press, v. 7, n. 1, p. 63–93, 2016.

ZHANG, Junlei et al. S-simcse: Sampled sub-networks for contrastive learning of sentence embedding. **arXiv preprint arXiv:2111.11750**, 2021.

ZHANG, Xinzhi et al. A survey on modularization of chatbot conversational systems. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. [S.l.], 2020. p. 175–189.

ZHANG, Yuhao et al. Biomedical and clinical english model packages for the stanza python nlp library. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 28, n. 9, p. 1892–1899, 2021.

ZHANG, Zhe et al. Maintaining continuity in longitudinal, multi-method health interventions using virtual agents: the case of breastfeeding promotion. In: SPRINGER. **International Conference on Intelligent Virtual Agents**. [S.l.], 2014. p. 504–513.

ZHANG, Zhe; BICKMORE, Timothy W; PAASCHE-ORLOW, Michael K. Perceived organizational affiliation and its effects on patient trust: Role modeling with embodied conversational agents. **Patient education and counseling**, Elsevier, v. 100, n. 9, p. 1730–1737, 2017.

ZIA, Haris Bin; RAZA, Agha Ali; ATHAR, Awais. Urdu word segmentation using conditional random fields (crfs). **arXiv preprint arXiv:1806.05432**, 2018.