Journal of Healthcare Informatics Research (2022) 6:253–294
https://doi.org/10.1007/s41666-022-00115-0

RESEARCH ARTICLE

The HoPE Model Architecture: a Novel Approach to Pregnancy Information Retrieval Based on Conversational Agents

João Luis Zeni Montenegro[1] · **Cristiano André da Costa**[1]

Received: 15 July 2021 / Revised: 26 January 2022 / Accepted: 16 February 2022 /
Published online: 6 April 2022

Abstract

Conversational agents are used to communicating with humans in a friendly manner. To achieve the highest level of performance, agents need to respond assertively and fastly. Transformer architectures are shown to produce excellent performances on recent tasks; however, for tasks involving conversational agents, they may have a lower speed performance. The main goal of this study is to evaluate and propose a HoPE (Healthcare Obstetric in PrEgnancy) model that is tailored to pregnancy data. We carried out a dataset extraction and construction process based on collections of health documents related to breastfeeding, childcare, pregnant care, nutrition, risks, vaccines, exams, and physical exercises. We evaluated two pre-trained models in the Portuguese language for the conversational agent architecture proposal and chose the one with the best performance to compose the HoPE architecture. The BERTimbau model, which has been trained on data augmentation strategies, proves to be able to retrieve information quickly and most accurately than others. For the fine-tuning process, we achieved a Spearman correlation of 95.55 on BERTimbau augmented with a few pairs (1.500 pairs). The HoPE model architecture achieved an F1-Score of 0.89, outperforming other combinations tested in this study. We will evaluate this approach for clinical studies in future studies.

Keywords Sentence-BERT · Data augmentation · Information retrieval · Conversational agents · Natural language processing · Public health informatics

✉ Cristiano André da Costa
cac@unisinos.br

João Luis Zeni Montenegro
joaomontenegro18@edu.unisinos.br

[1] SOFTWARELAB, Applied Computing Graduate Program, Universidade do Vale do Rio dos Sinos Tel, Av. Unisinos, 950, Sao Leopoldo, RS, Brazil

1 Introduction

Conversational agents (CA) are a type of software that is widely used for human-computer communication. Additionally, these systems make use of artificial intelligence (AI) software, which may involve a natural message exchange with the user. In general, this technology may be advantageous for marketing purposes, determining a geographic location, enhancing customer service, and automating certain operations [63].

Recent research indicates that the user's intention to use CA to access health information can be extremely beneficial and is strongly influenced by the ability to provide realistic advice. The CA must be familiar with the correct interpretation and understand how to communicate this information to the user clearly and assertively [48].The conversational agent can be used in pregnant education as a tool for identifying and mitigating preconception health risks, thereby assisting African-American women with less education in this area [44].

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 [120]. Agents could promote breastfeeding education through the use of counseling techniques, according to this study[120]. Counseling on social networking sites during the prenatal and postnatal period may result in improved 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 [14] [12].

Interactions with conversational agents frequently reveal useful information. Information can help humans make better decisions in a variety of situations. For example, accurate information during pregnancy can help in the prenatal and postpartum periods. According to the Ministry of Health, cesarean deliveries accounted for 42% of births in Brazil in 2018, and in the majority of cases, the decision is made due to a patient's lack of information. Numerous studies on patient education, in general, have been conducted over the years [12, 13, 44, 68, 72, 99, 120]. 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 [120]. Agents could promote breastfeeding education through the use of counseling techniques, according to this study[120]. Counseling on social networking sites during the prenatal and postnatal period may result in improved 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 [12, 14].

The article's main scientific contribution is to propose a model of conversation agent called HoPE (Healthy Obstetrician for Pregnancy). The HoPE model is a conversational agent hybrid architecture in health that focuses on information

delivery to your target audience: pregnant women. The proposal for information delivery is intended to promote pregnant women's health literacy on topics that we have determined to be priorities during the thousand days of pregnancy. We examined some of the studies involving architectures and models of conversational health agents that have been conducted in these areas over the years.

We intend to bring two significant contributions to health computing: a new approach of conversational agent model architecture using transformers and ontology structure to support pregnant doubts and a new corpus of natural language inference to the field of pregnancy health in Portuguese. The study evaluates the performance of the HoPE architecture through information retrieval assessment, and performance evaluation for pregnancy guidelines fine-tuning.

The article is organized into six sections: Section 2 presents the background of key concepts, Section 3 presents the related work on architectures for information retrieval, Section 4 shown all architecture components of HoPE framework while Section 5 shows details of the proposed architecture and the experiment methodology, Section 6 shows the results, Section 7 discussion, and Section 8 the main work considerations.

2 Background

In this section, we present the concepts associated with HoPE architecture, particularly theories that support the main contributions, such as the thousand days of pregnancy, transformer architectures, lexical retrieval, and ontology structure.

2.1 Thousand Days Period

Based on the concept of the Lancet series [73], "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 2 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 [102].

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) [64].

When confronted with these and other symptoms, women tend to seek information to alleviate them. During pregnancy, they frequently use the internet as a source of information (70–97%) [15]. According to one study of pregnant women, the web was frequently used for seeking pregnancy information, verifying information

received from health professionals, social networking, social support, and electronic commerce (e.g., e-commerce) [52].

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 [92]. Examining the impact of this habit on pregnant women's decision-making can help to enhance their decision-making process. To create meaningful online tools, healthcare clinicians and web developers must understand how and why pregnant women use the internet while making decisions [52].

The study of [23] carried out a survey with some of the main topics and ways that 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. According to the vast majority of women across all demographics, family members are a crucial source of information about healthy pregnancy and childbirth. Women with children in early childhood reported that they often follow their intuition for decision-making. All women use the internet as a source of information. It was discovered that Google and other search engines are frequently used. However, trusted websites, such as health guidelines-based apps, are quite popular among the groups surveyed [23].

2.2 Transformer Architecture

The use of word embedding systems has been used as a feature for machine learning systems, which enables new techniques to contextualize raw text data [62]. Recent dataset enhancements such as GLUE [107] and SQuAD [82] have driven the development of natural language understanding (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 [65].

When compared to embeddings learned from scratch, NLU's few-shot strategy to transfer pre-trained neural language representations improves downstream task scores [4]. More recent work, including but not limited to [1, 17], has added end-to-end fine-tuning of language models for downstream tasks, as well as extraction of contextual word representations, expanding on these ideas even further.

Due to this advanced engineering, and large compute availability, state-of-the-art NLU's transformer architecture has evolved from word embedding to transferring language models with billions of parameters achieving unprecedented results across natural language processing tasks [57].

The original transformer is a six-layered encoder-decoder model that generates a target sequence based on the encoder's output. The encoder and decoder, at a high level, have a self-attention layer and a feed-forward layer. By adding an attention layer between them, the decoder can map its relevant tokens to the encoder for translation purposes. Self-attention enables the look-up of remaining input words at

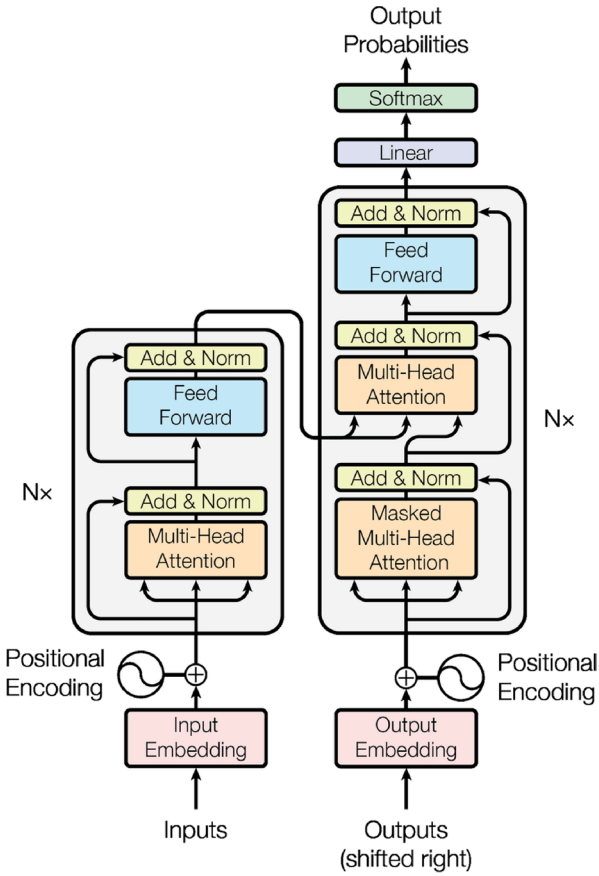


Fig. 1 State-of-art transformer architecture

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 of all words [96]. We present an illustration of this architecture in Fig. 1.

Transformers address a significant issue for previous architectures in terms of word embedding system evolution: the models were static, which meant that each word had a single vector regardless of context. This created dozens of new problems, not the least of which is that all possible meanings of a polysemic word will use the same image. Transformer architectures generated contextualized word representations and context-aware word vectors [27, 75, 85]. Many architectures such as GPT-3 [32], BERT [27], and ELMO [75] have already been used frequently in recent studies.

The BERT (Bidirectional Encoder Representations from Transformers) model architecture has been regarded as the cutting-edge in tasks such as text extraction, question answering, and entity recognition [54]. 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 a word integration for every

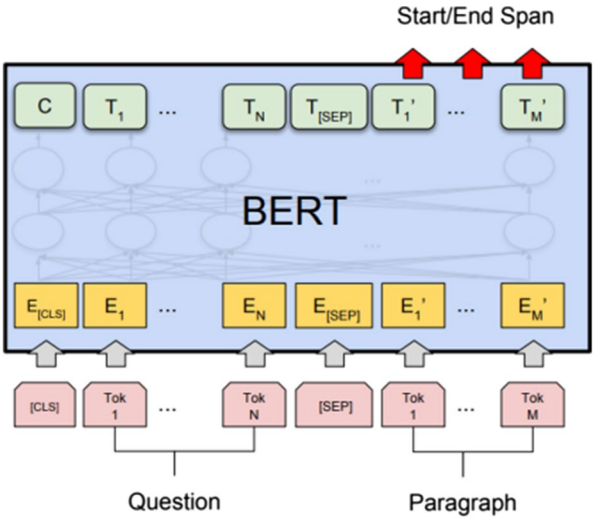


Fig. 2 Example of Bidirectional Encoder Representations from Transformers for question and answering

word within the context of the sentence. This means that the sentences "I came to the bench today" and "I'm sitting on the park bench" offer the term "bank" with distinct representations in each of them [27]. The BERT architecture is designed in such a way that during training and predicting embedding, both previous and next words are taken into account, and attention methods are employed to retain the most and least significant word information in the phrase [27].

The BERT architecture is shown in Fig. 2. Being a transformer encoder, it is a bidirectional model because of the complexity of the encoder self-attention within the transformer architecture. For a large number of NLP functions, BERT provides an advanced method for obtaining contextualized word embedding. BERT for NLP proposals outperformed previous state-of-the-art results in eleven different tasks, including a question-answer [27].

A pre-trained BERT model acts as a way of inserting words into a given sentence, taking into account their context: the last word in the secret state of the transformer's encoder [27, 105]. As has been evaluated in other articles [54, 110], the BERT model obtained good results when used for text mining in the medical literature.

Studies such as [112] used the BERT network to evaluate different methods for a Q&A system trained on Chinese medical data. SCI-BERT [10], which leveraged unsupervised pre-training on a large multi-domain corpus of scientific publications, was introduced in and BioBERT, which was pre-trained on biomedical domain corpora (e.g., PubMed abstracts and PubMed Central full-text articles), was proposed in [54].

The study by [9] seeks to use the BERTimbau model trained in Brazilian Portuguese to solve industry problems in a chatbot architecture. NLP models trained in Brazilian Portuguese are infrequent, and therefore the BERTimbau model [97] plays a vital role in embedding models. This model was also used in an NLU architecture for conversational agents that support the population against COVID-19 [45]. The use of derivation of the original BERT as siamese structures was used for

information retrieval task in Brazilian Portuguese language in architecture for conversational agents [74]. Although some models trained in European Portuguese can be used, the nuances between languages necessitate the development of new studies and clinical datasets in Brazilian Portuguese [70].

While BERT has achieved new state-of-the-art outputs for downstream natural language processing tasks, the models' 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 result in a slow process sometimes [83, 86]. In a conversation with humans, the time response of conversational agents is critical [6]. Next section, we present Sentence-BERT (SBERT) [83], a derivation of BERT embedding that overcomes this type of difficulty.

2.3 Sentence-BERT

Regarding the different methods of the sentence embedding, the siamese neural network's architecture presents itself as a valid alternative to derive embedding from semantically significant sentences [83]. These structures applied to pre-trained BERT models have often been associated with semantic research tasks, the similarity of sentences, and information retrieval [37, 83, 98].

In [108], siamese-recurrent architecture and siamese-convolutional architecture are used during the preliminary investigation to discover sentence similarities in the Chinese language. This architecture outperforms recurrent's architectures in terms of accuracy.

The efficacy of these strategies with BERT was also demonstrated in the [115] 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 generate 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 [83]. Information retrieval works on the following principle: if you feed it 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 embedding runs into trouble when dealing with a few number of large documents. For this reason, a fine-tuning process is important to enrich this model for more accuracy.

The standard fine-tuning process uses a bi-encoder network (SBERT) on the labeled target dataset [83]. It works passing sentence pairs (A / B) in a neural network where each sentence (A/B) yields the embedding u and v as shown in Fig. 3. The similarity of these embeddings is calculated using cosine similarity and compared to the gold similarity score. This allows the network to be fine-tuned and recognize sentence similarity. The fine-tuning of data is limited to the upper layers of the pre-trained model to perform "characteristic extraction," which allows the model to use the representations of each model [83].

Using a data augmentation strategy, we can train SBERT on datasets comprised of a few pairs (1k–3k). SBERT augments annotated or unannotated datasets and significantly improves results in models fine-tuned with a few data points.

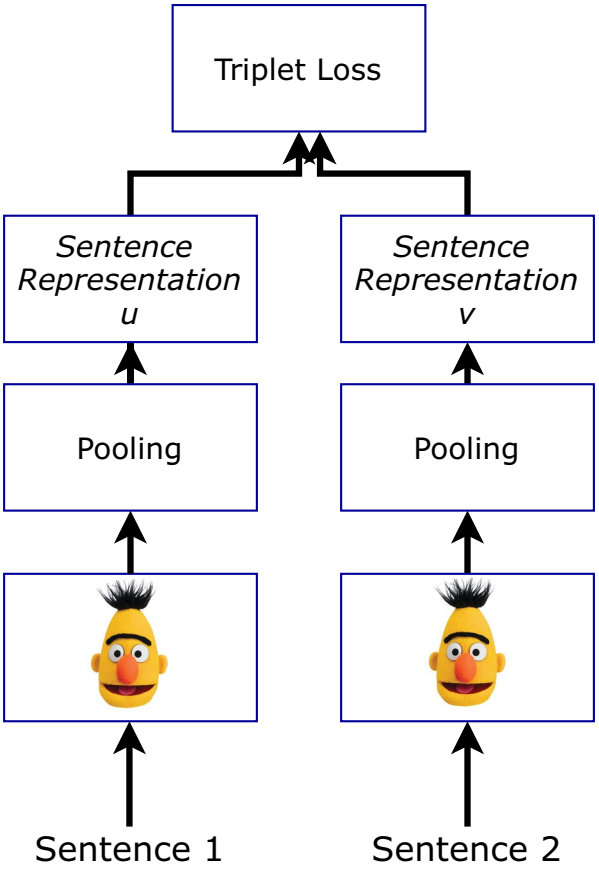


Fig. 3 SBERT's bi-encoder structure allows you to fine-tune data to pre-trained models

Fig. 4 Sentence-BERT strategy for data augmentation

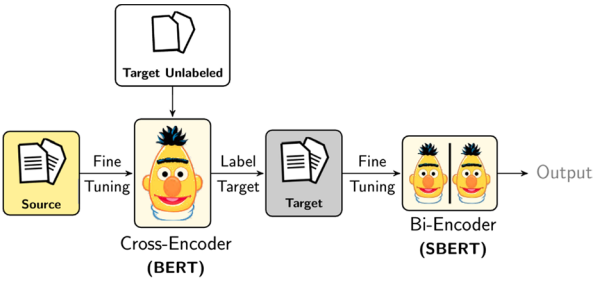


Figure 4 illustrates the data augmentation training process. A cross-encoder network is used in this strategy to label pairs of unlabeled sentences using a traditional BERT model. The cross-encoder can be trained on datasets from the sentence text similarity benchmark and then applied to smaller domain data [101].

2.4 BM25 Okapi

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 [46], newspaper [53], medical summaries [35], and other things. Documents are often used to refer to certain items. When referring to a broader range of retrieval activities, such as document or text retrieval, the term "information retrieval" may be used (IR) [40].

One of the state-of-art IR models is BM25. The BM25 model has different variations [58], such we have the BM25 Okapi. The Okapi BM25 is still a popular benchmark for similar jobs. The Okapi BM25 provides a TF-IDF benchmark that we can employ. 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 [89]. The Okapi BM25 formula is a baseline method purposed for Terrier [111].

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

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

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 that are usually chosen in the absence of advanced optimization. *(Q_i)* is the number of documents that it contains.

2.5 Ontological Knowledge Base

The development and design of ontologies is a complex process that requires knowledge management and subject matter experts. Because of the development of the semantic web, this type of representation has grown in popularity in recent years. Ontology systems rely on knowledge gained through the use of a formal language, which can be represented in a variety of ways. These systems are intended to generate new knowledge from previously collected data [59, 77]. In [69], it 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 through defining individual instances of these classes by filling in specific properties value information and by limiting more properties [69].

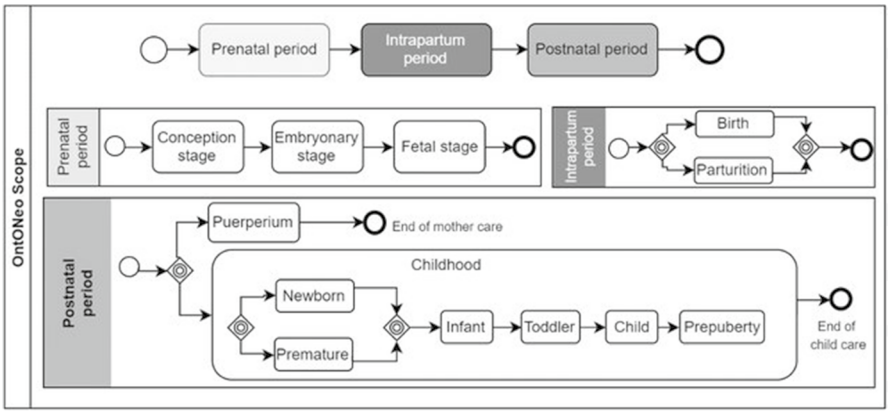


Fig. 5 Pregnancy and child development stages represented

Ontologies can also be combined into a single structure by sharing 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 another ontology's from various subjects [69].

Data is stored in two forms to model knowledge: A more complex structure is *OWL* (Ontology Web Language) and is a Web Language representation, which maps all things that the agent can infer around a domain, and a simpler structure is *RDF* (Resource Description Framework). Graphs from the RDF that specify facts and relationships in a straightforward manner [69]. *SPARQL* is a query language that is used to query ontology. It is a query language for knowledge extraction that connects the RDF structure of an ontology to the SQL language of a typical database [8, 81].

Three types of *RDF* data exist: IRIs, blank nodes, and literals [22] . 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 [80]. As a result, collections of triples are frequently referred to as *RDF* graphs.

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 in nature, as it contains information from the pregnant woman's medical record [28] (Fig. 5).

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 was decided throughout its development to reuse existing ontologies

Table 1 OntONeo ontology metrics

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

from the OBO Foundry to improve interoperability with existing biological ontology and to leverage previously established ones. OntONeo is being built incrementally and iteratively. The ontology is constructed incrementally over time, with each iteration's scope predetermined. Each new version of the ontology introduces new entities and relationship [31]. Table 1 summarizes the current ontology metrics.

According to the authors of ontology [31], OntONeo, the results may not be applicable in other contexts because they were developed using examples from specific EHRs. The content of the ontology, on the other hand, is focused on the representation of general entities, implying that it can be used in a variety of contexts within a specific domain.

3 Related Work

This section will discuss the most recent and significant works on the subject of our study. We intend to approach techniques and architectures about the central concepts that sustain the HoPE concept.

Sentence-BERT has been used to research information retrieval systems and conversational agents. In [30], the CO-Search, a semantic search engine designed to manage complex inquiries into COVID-19 literature used siamese BERT-based and TF-IDF as encoders for paragraphs embeddings to perform the task.

In [114], the proposal for an end-to-end model for a question and answer system integrating SQUAD to Anserini toolkit was made. 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. Using the BERT with data augmentation technique [115], the stage-wise method is applied to fine-tune BERT on a multitude of datasets, beginning with data that are "furthest" from the test data and ending with the "closest." The results presented offer performances superior to datasets in English and Chinese question-answering (QA).

New data augmentation strategies were demonstrated to dynamically annotate paragraphs as positive or negative instances to accompany training data, which were then combined to fine-tune BERT. This research provides evidence that two English and two Chinese QA datasets can do well together [113]. The BERT model is also used in conjunction with conversational agents to evaluate document prediction tasks, which involve a new set of public data [34]. The proposed results aim to strengthen the use of deep learning techniques for information retrieval tasks.

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 [91], 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 useful feedback documents and attempts to understand the relationship between query and documents in terms of document usefulness for query extension.

In [3], a hybrid approach to query expansion that combines statistical and semantic approaches is presented. 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 lengthy natural language questions, demonstrating BERT's ability to extract valuable information from complex inquiries. In [71], 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 [103], 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.

We intend to contribute scientifically in the field of health computing by proposing an architecture model based on a conversational agent to support pregnancy doubts and general questions and to offer insights during this period. The studies of [60, 63], and [94] had already warned of the need for further studies exploring the combination of different techniques for conversational agent architectures. Using a search string **("conversational agents" OR "chatbots" OR "virtual agents") AND ("hybrid") AND ("architecture" OR "model") AND ("pregnant" OR "pregnancy")** related to our aim in the period between 2017 and 2021 on Google Scholar platform, we found no studies that resemble our architectural model proposal. In the next section, we will discuss our model, and how it aims to help pregnant women gain reliable insights and recommendations in a variety of contexts.

4 HoPE Model Architecture

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

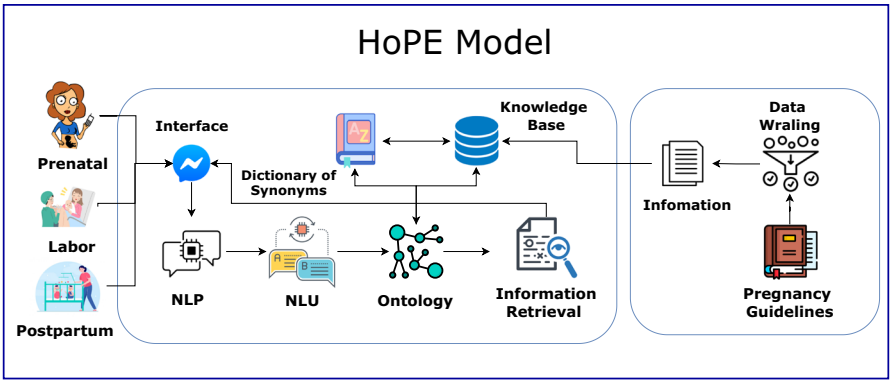


Fig. 6 Support for pregnancy doubts using the HoPE model

during several critical periods. The goal of this proposal is to combine semantic strategies to retrieve the most assertive information possible as response to questions from pregnant women.

The framework shown in Fig. 6 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 in *JSON* format, which can be parsed to extract the required information.

The dialog structure proposed by HoPE uses a composite of pre-defined rules, NLP machine learning engines, and ontology-oriented dialogs. Rule-based strategy is responsible for basic input or output in our conversation agent. They are rigid structures aimed at providing pragmatism in the conversation: greetings, goodbyes, initial explanations, agent feedback, and 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 [61]. NLP machine learning engines are usually associated with pre-defined 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 [16]. 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 in recognizing 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.

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

The agent uses the intent configuration to maintain conversational coherence by establishing contextual inputs and outputs [16].

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, difficulty with unforeseen contexts, and stagnation in the process of detecting user intent [29, 76].

As a contribution, the HoPE model seeks to add components that will 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 used for pre-processing 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 dialogues, and we do so by utilizing ontology. Its application to management tasks has been thoroughly investigated in [20, 78, 100]. The OntONeo ontology is composed of entities that correspond to domains in a pregnant woman's electronic health record. This structure incorporates pre-defined relationships defined by specialists and was also used to structure and store content from pregnancy guidelines in our model.

This process was also aided, using a terminology dictionary based on the content of these guidelines. This artifact aids in the ontology queries, acting as a refinement for stored contents search. The information retrieval module is the component in charge of responding to user queries. This module includes a neural network model that has been pre-trained on large datasets and has a high capacity for 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 with data from 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 pregnancy time was the focus of these materials. For our purposes, we determined that scientific articles and case studies were inadequate. In the next sections, we present the modules referring to the structure of the conversational agent in production. We divide it into three main sections: intent recognition, dialog management, and information retrieval.

4.1 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

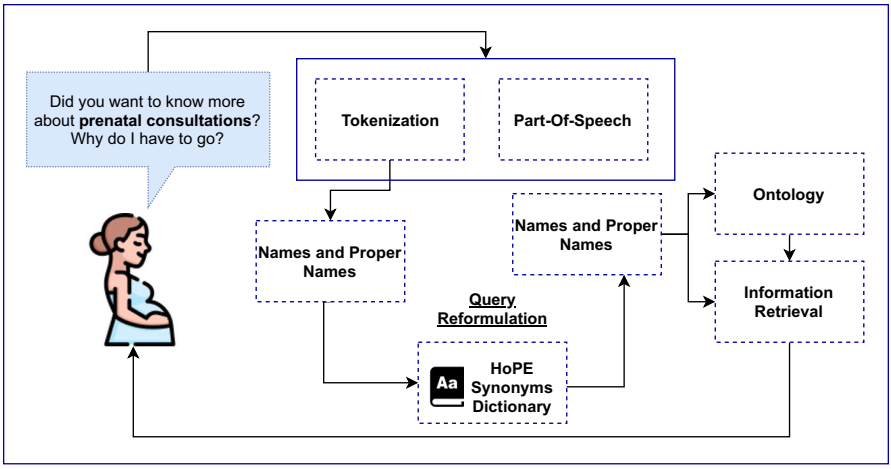


Fig. 7 Entity extraction and treatment process by the dialog manager in HoPE architecture

the incorrect answer) and recall (the agent does not recognize the intention with satisfactory reliability).

The HoPE model proposes strategies for supplementing the conversational agent with a new NLU, disambiguation, and classification process to increase the accuracy and recall coefficients. An overview of the HoPE dialog management module for the conversational agent architecture is provided in the next section.

4.2 Dialog Manager

The dialog management module aims to orchestrate actions after the intention recognition phase (Fig. 7). We obtained the intent and/or recognized entities in the previous module. It is possible, however, that the intention was misunderstood or that an entity was missing from it. Additional natural language processing techniques are required to process the text and use it for full sentence comprehension. Tokenization refers to the breakdown of sentences into tokens, whereas Part-of-Speech refers to the distinction between proper and common names. We also performed a steaming operation and normalized all entities.

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, they are compared in the synonym field to their synonyms. If we come across it, we will replace it with a term from our ontology. Otherwise, the conversation is routed directly to the information retrieval module.

The entity "physical exercise" for example exists in our corpus because it is a common term in health guidelines. We added synonyms such as gym, cross-fit, and yoga to this term. To ensure that the entity recognized in the previous module is familiar to the conversational agent, it must first be checked in 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, we use the entities as input for ontology research. The strategy is to retrieve responses present in individuals of the ontology that contain 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. In case we do not find the correct entity, the sentence will go to the information retrieval module directly. A summary of the process is presented in Fig. 7.

4.3 Information Retrieval

In this section, we show the recovery process for a user query. We incorporate the retrieved sentences from ontology in a list and index them in the pre-trained Sentence-Bert model.

Initially, the input sentence is consolidated with the possible answers from the ontologies in a bi-encoder model. The model's output should bring a retrieval of 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. The sentence with the highest score is chosen as a response to user input.

This module can also be activated without using the ontology output, when the entry sentence is "out-of-scope." If no entity or name is found, we try to respond to the user using the Sentence-BERT model with pre-computed response embedding in clusters. In this case, we preload the representations of the paragraphs in indices and cluster them using the approximate nearest neighbor (ANN) search.

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 [87]. 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 ontology management. 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.

Following that, the HoPE information retrieval module will present three options: respond with a relevant and correct score (greater than 65%), respond with a relevant and incorrect score, or respond with 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, as well as through the use of static data such as weighted averages and standard deviation.

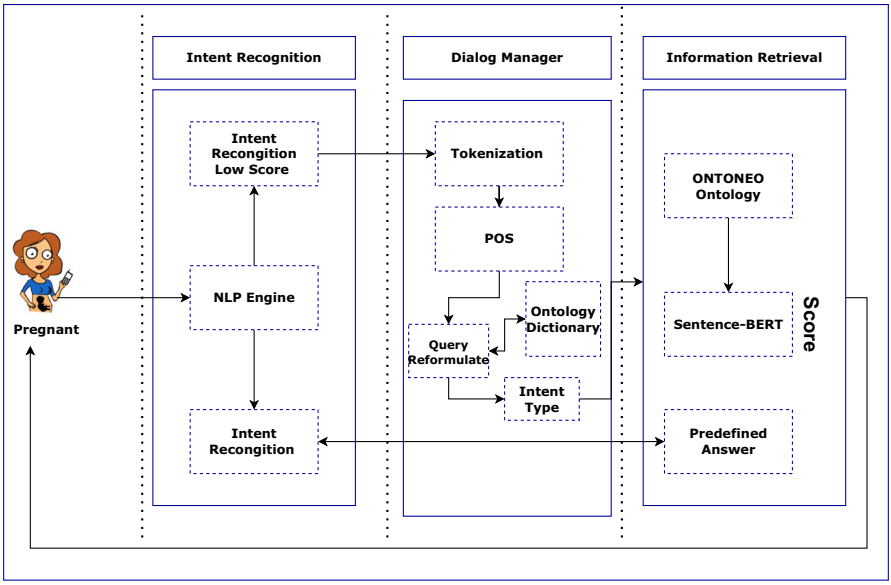


Fig. 8 Conversational agent architecture in production environments

5 Materials and Methods

This section details the materials and methods used in our study. The HoPE architecture offers different contributions in its structure. The present work aims at the application of two evaluations: the first one is based on Sentence-BERT evaluation, seeking to identify which model had the best performance to integrate in HoPE. The second evaluation aimed to perform inference tests, using the structure with the proposed dialog manager. Details about the corpus construction, ontology, and dictionary structure, as well as parameters for training the models and evaluation metrics, are provided below (Fig. 8).

5.1 Corpus Construction

The first step in the text processing process is scanning digital documents. The AWS TextExtract[2] service is used to extract text from these documents and convert it to the *.txt* file format, which allows us to manipulate the document. Likewise, editable pdf files have been converted to *.txt* format. The texts were then tokenized, with any unnecessary images or tables removed. This is accomplished by combining Pypdf2[3]), functions with the tokenization and structuring capabilities of the NLTK[4] and Pandas[5] packages.

[2] https://aws.amazon.com/pt/textract/
[3] https://pythonhosted.org/PyPDF2/
[4] https://www.nltk.org/
[5] https://pandas.pydata.org/

Table 2 The distribution of percentages by corpus topic

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%

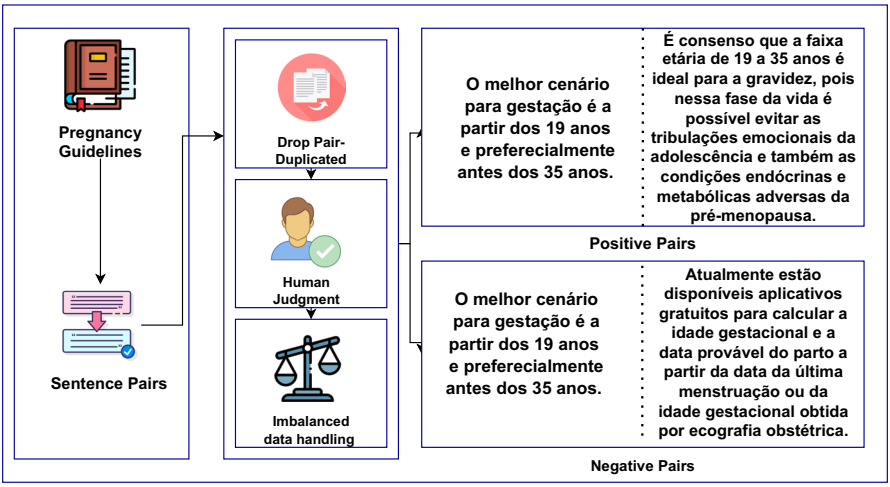


Fig. 9 An example of positive and negative pairs is present in our base. Positive pairs (pregnancy is best starting at the 1 year and ending before the age of 35 years. The best scenario for pregnancy is from the age of 19 and preferably before the age of 35). Negative pairs (pregnancy is best starting 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)

Following that, these documents were reviewed manually to ensure no line break adjustments or extraction failures occurred. Some sentences carry the context of the preceding sentence, resulting in tokenized paragraphs. We chose human review over other computational approaches for text adjustment because it is more assertive. The procedure resulted in the creation of a corpus containing 7.077 documents. This content was used for two purposes: building a knowledge base for conversational agents and fine-tuning Sentence-BERT model training. Table 2 presents the distribution with the main topics present in the corpus.

The data must be organized in a specific format for the evaluation phases. Figure 9 illustrates a comprehensive view of this process. We begin by organizing the domain's data in the following manner: *(Q1, Q2, Score)*. The process of training 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 not provided in the labeled form, but rather in pairs. We used unsupervised Sentence-BERT models to speed up, producing positive sentence pairs through semantic similarity. The objective was to find similar or almost similar pairs, which we empirically understood would be the minority. We used 12 guidelines in our study, and we expect that there would be similar issues between the documents.

We employ a semi-supervised annotation process based on [101] 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 a few more preprocessing steps such as removal of duplicate pairs, human judgment, and class balancing. The removal of duplicate pairs was the removal of identical sentences *(A, B)* that resulted from the pair generation procedure. Six hundred sixty-four pairs remained. We reviewed the pairs with a human judgment step to declare whether they were similar or not. We identified about 325 paragraphs with similar content written in different ways. The authors and two medical researchers re-written 275 phrases to increase the positive examples. We focused on providing sentences with the same meaning but rewritten and of smaller size to adjust the model to deal with size asymmetries between short and long sentences. Previous studies [2, 51, 119] pointed out that for better evaluation results, balancing examples was an important feature. After this stage of data organization, we had 600 positive and 900 negative random pairs. This dataset is sufficient for fine-tuning the Sentence-BERT neural network [2, 83].

5.2 Pregnancy Dictionary of Entities

We created a query reformulation strategy as well as an entity dictionary. We extracted personal and proper names from each paragraph corpus. These terms were used to search the WordNetPT ontology[6] for synonyms for each entity. The authors added the synonyms that were not found in ontology (e.g., specific medical terms). Synonyms are normalized and stemmed before being added to the dictionary as values alongside the corresponding term. The strategy, in this case, is to map whether the input entities exist in our dictionary.

5.3 Ontology Procedures

To help understand the user query and be more assertive in retrieving responses, we use ontology as a structure for the chatbot. The OntONeo ontology was used to play

[6] https://github.com/recognai/spacy-wordnet

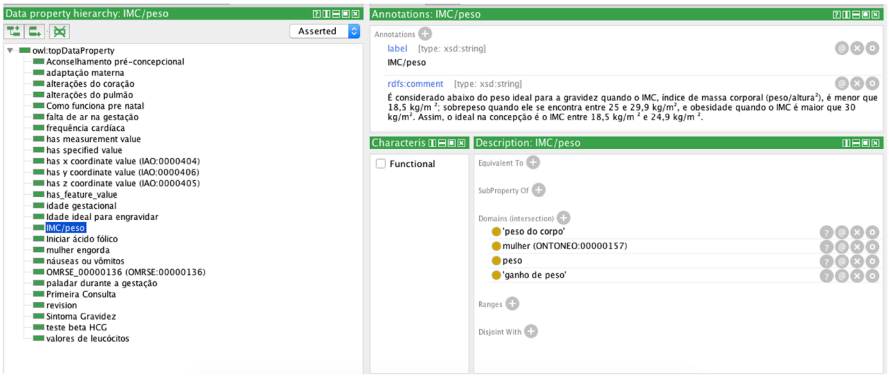


Fig. 10 Relationship of data properties and classes

this role in our work. This ontology is based on electronic medical records of pregnant women, and it contains concepts and relationships that are similar to the content of the knowledge base used in our study. The offline process involving the Data Property population of ontology with our data. We store each sentence in a Data Property in the ontology. Each sentence is stored as a text *rdf: comment*, as well as a label to identify the intent of that paragraph. An example: *rdf:comment* Avoid eating fried foods and bacon every day rdf:label (Can I eat bacon during pregnancy?). Thus, for each semantic search based on entities extracted from the sentence, the query will return a set of documents related to those terms. The information retrieval module will perform the task of re-ranking and understanding which answer is correct and if there is a correct answer (Fig. 10).

5.4 Fine-Tune Procedures

In the HoPE architecture proposed in our 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 as well as the data augmentation strategy proposed [83] for our experiments. The data augmentation strategy is the most appropriate for the corpus type we are using here. We chose this approach due to the lack of labeled data.

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 three phases of data augmentation. The first phase of training can begin with the design of the cross-encoder. For this purpose, the approach suggests selecting a benchmark dataset within the chosen task. This dataset is also referred to as a "gold dataset."

The ASSIN2 benchmark was chosen because it is already used for sentence text similarity (STS) tasks and be trained in general vocabulary examples. The goal here is to train a cross-encoder model using knowledge transfer. The ASSIN2 dataset is

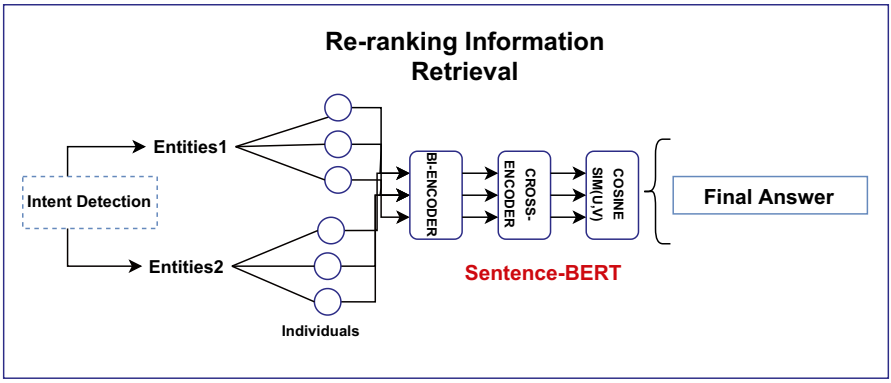


Fig. 11 Information re-ranking system using a bi-cross-encoder structure

widely used in Brazil to evaluate supervised STS systems. The training and validation datasets contained 6.500 and 500 pairs of annotated Brazilian Portuguese sentences, respectively, for inference and semantic similarity. The semantic similarity values ranged from 1 to 5, with the entailment and no inference classes. The test dataset contains approximately 3.000 pairs of sentences that contain the same linguistic phenomena and annotations as the training dataset. All data was gathered manually.

The second step is to label our dataset, which is also referred to as the "silver dataset." As suggested by the strategy, we must normalize the data scores (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. This step is responsible for conducting semantic research, which entails comprehending research content via lexical correspondence, context, and synonyms. As a result, we reranked models and optimized retrieval according to our answers. All processes are depicted in Fig. 11.

For all training performed, we used 256 max length for the tokenization layer and a MEAN-pooling strategy. We run 1000 evaluations steps and use the ADAM optimizer for all models. These parameters are frequently used in studies that perform training in Sentece-BERT networks [83].

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) [39]. The hyper-parameters that were customized for each specific model are listed in Table 3.

Batch sizes were 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 was adjusted based on the model's performance. We iterated over this hyper-parameter until we found the optimal coefficient without

Table 3 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

Table 4 Pre-trained BERT models used in fine-tuning cross-encoder

N	Model	Citation	Corpus
1	BERTimbau[a]	[106]	brWaC
2	Paraphrase-Multilingual-MiniLM[b]	[83]	Microsoft-Multilingual
3	BERT-Multilingual[c]	[27]	Wikipedia

[a]https://huggingface.co/neuralmind/bert-base-portuguese-cased

[b]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[c]https://huggingface.co/bert-base-multilingual-cased

over-fitting. The learning rate for the para was reduced because it performed better at this rate.

5.5 First Evaluation: Pre-trained Portuguese Sentence-BERT Models for Retrieval Pregnancy Information

We conducted two assessments: the first evaluated the quality of embedding generated, and the second assessed the models' performance on our validation and test data using similarity metrics. We chose three major models to participate in this evaluation to facilitate comparison. Table 4 summarizes them.

The BERTimbau model is pre-trained on data from the brWaC corpus. This corpus is a multidomain dataset, composed of 2.7 billion tokens annotated with tagging and parsing information. The number of web pages contributing to this set is 120,000 [106]. Another model used is the Sentence-BERT paraphrase Multilingual. It is pre-trained on a machine reading comprehension (MS-Marco) corpus proposed in [67], which offers a large dataset extracted from real web documents using the most advanced version of the Bing search engine. About 100,000 queries are present in this dataset, which is widely 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 [27]. 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 models which were translated.

For this experiment, we split the data into sets of training, validation, and test. From 1.500 paragraphs/sentences in our corpus. We allocated 80% to training, 10% to validation, and 10% to 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 [55, 83].

Spearman's correlation coefficient, a rank-based alternative to Pearson's correlation coefficient that works with non-normally distributed and non-linear variables, is a rank-based alternative to Pearson's correlation coefficient. Its application is not limited to continuous data analysis; it can also be applied to ordinal attribute analyses.[24].

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

The Pearson correlation Pc 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's 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 [11].

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

Some distances are used in the bi-encoder evaluation. We have collected a few of them here. The Euclidean distance is a distance metric between two points or vectors in a two- or multidimensional (Euclidean) space that is based on Pythagoras' theorem. Squaring the sum of the squared pair-wise distances in each dimension yields the distance [56].

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

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 [56].

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

The Manhattan distance is used to determine the distance between two real-valued vectors or points. It is calculated by adding the absolute differences in their

Table 5 Three examples of ground truth pairs for the retrieval information evaluation

N	Questions	Correct 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.

Cartesian coordinates. The Manhattan distance, defined for a plane containing a data point p1 with coordinates (x1 y1) and its nearest neighbor p2 with coordinates (x1 y1), can also be used (x2, y2). A comparable connection can be defined in higher-dimensional space [36].

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

The last metric used for evaluation in this experiment was the dot product, which is equal to the sum of the product of the horizontal components and the product of the vertical components [50].

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

5.6 Second Evaluation: HoPE Architecture Evaluation

The evaluation of the HoPE architecture is covered in this section. To assess the effectiveness of the HoPE model, we conducted two experiments: test collection and inference speed test. The test collection experiment tested the models' accuracy by utilizing sixty question/answer pairs as ground truth pairs. As previously stated in [90], this experiment is composed of 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 1.500 paragraphs/sentences 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 trials. We attempted to combine questions that contained short and long sentences, sentences with minor orthographic

Confusion Matrix	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%***	

Fig. 12 Confusion matrix for evaluating a system for information retrieval

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) were annotated. Three of these pairs are illustrated in Table 5.

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 [104]. We calculate the time from 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 are used: pre-computed embedding and online embedding. The embedding representation of 247 paragraphs/sentences was indexed offline, serialized, and grouped by the ANN strategy. Online embeddings refer to the stage of ontology retrieval paragraphs/sentences on the fly in the system. The Google Collaborative[7] environment, which provides open-source GPUs and CPUs, is used for the speed test experiments.

The metrics used to evaluate the previously mentioned experiments are also reported here. The confusion matrix [26] is a machine learning construct that stores information about a classification system. A confusion matrix is bi-dimensional, with one dimension representing the object's actual class and another representing the class predicted by the classifier. The confusion matrix that was used in this investigation is depicted below in Fig. 12:

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 by the human, but relevant by the system; (3) fallback, classified as relevant by the human but irrelevant to the system; and (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 [79]. The F1-Measure performance metric is the most frequently used for text classification. Precision and recall are defined as

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

the harmonic mean of their precision and recall. It is known to be more informative and valuable than classification due to the widespread phenomenon of class imbalance in text classification [37]. In the context of information retrieval, the F1-Measure is a special case of measure with an equal weighting of recall and precision. It has a maximum value of 1 and a minimum value of 0.

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

For this experiment, we apply a stratified k-fold cross-validation [33] with k five (the class distribution remains identical for each fold) in our corpus to allow generalizing our results. To assess inference speed, we use two distinct measures: speed for queries requiring online vs offline encoding, and inference speed to get the answer to the user.

6 Results

This section contains the findings of our evaluations. In all of the experiments, we used data from pregnancy-specific health guidelines as a knowledge base. We began the evaluations 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, and finally we evaluated the information retrieval capacity using the HoPE model.

6.1 Corpus Evaluation

We found several common topics in the dataset that are relevant to the context of the thousand days. In this way, we can see that the content extracted from the various materials is consistent with what we intend for this work. We chose a smaller sample size for the experimental steps because it contained more critical information about the post and prenatal periods. We selected this sample based on an assessment of two assistant physicians. Two hundred forty-seven sentences were separated and then classified by topic. Figure 13 depicts the distribution of this sample.

Ten topics were listed for the sample used as a knowledge base in the experiments. The most common topics were Gestational Symptoms, Gestational Information, and Nutrition. Medication and physical activity had the lowest representation in our sample.

Gestational Symptoms address prenatal symptoms, whereas Gestational Information topics cover general pregnancy information such as frequently asked questions and curiosities. Nutritional information was related to pregnant women's eating habits and feeding recommendations for newborns.

Pregnancy alteration addresses aspects that change during pregnancy in the pregnant woman's body. Illness is a topic that covers information about the most

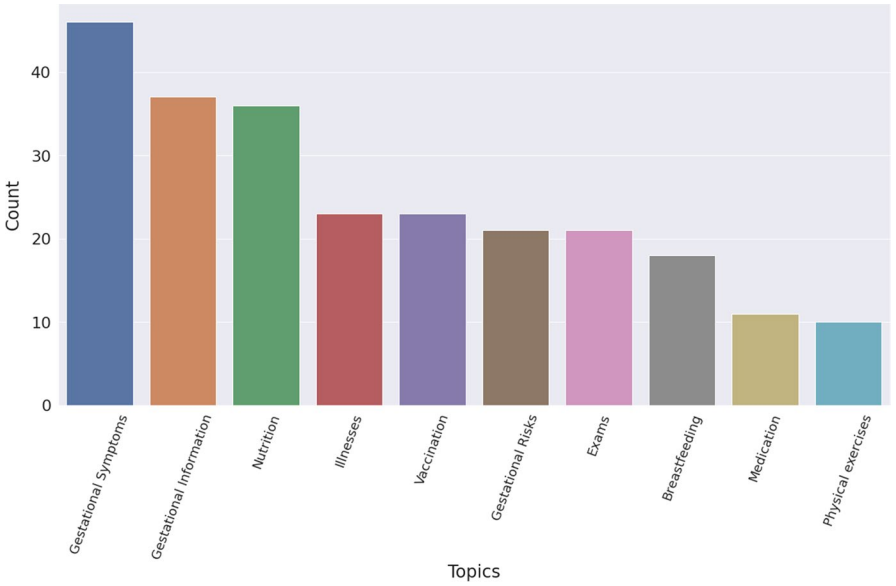


Fig. 13 Subjects that appear in our documents more frequently

common diseases that occur during pregnancy, whereas vaccination covers immunization information for pregnant women and newborns. The topic Gestational Risks refers to information about preventive behaviors during pregnancy and exams contents assist with 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 guidelines for specific exercises.

We also looked at the frequency of each topic to see which terms were used the most frequently across all of them. Childbirth plans, gestational factors, age doubts, and menstruation were the most frequently mentioned gestational symptoms, according to our corpus's frequency analysis.

The majority of gestational information focuses on consultations, gestational age, and pregnancy phases. Nutritional concerns about foods and beverages for children and pregnant women were frequently expressed. The topic about illnesses brought up more frequent topics 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.

Table 6 Fine-tune cross-encoder models vs literature models

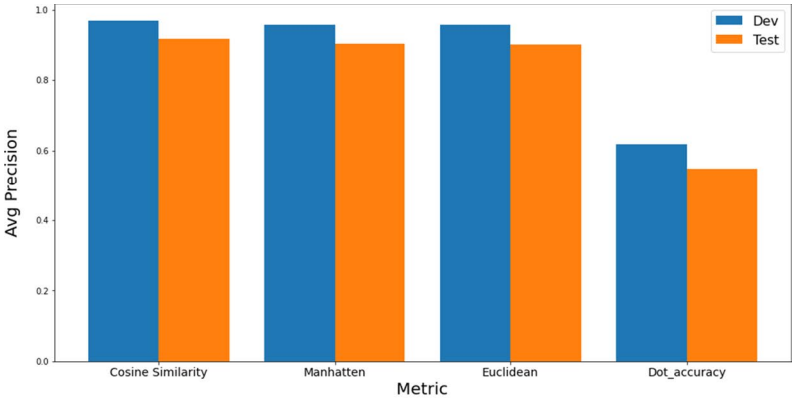
Models	Spearman
Sentence-BERTimbau Augmented	**90.55**
Sentence-BERT Multilingual Augmented	**90.33**
Sentence-BERT Multilingual	**89.21**
Sentence-BERTimbau	**83.97**
Not trained on STS	
Avg. GloVe embeddings	58.02
InferSent	46.35
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	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

6.2 Evaluation of Pre-trained in Brazilian Portuguese SBERT Model's Applied for Retrieval Pregnancy Information

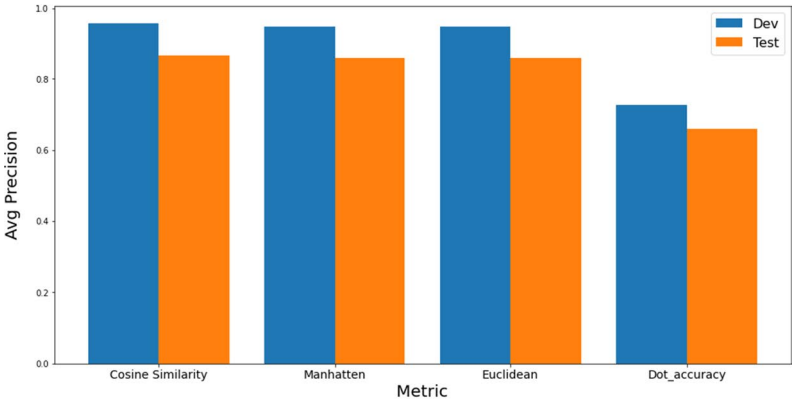
This section presents the findings of an information retrieval models evaluation that have been fine-tuned for use in health guidelines. We assess the embedding generation and validation processes. In the first assessment, we measured the models embedding the performance for fine-tuning cross-encoder. We measured the Spearman rank correlation between the cosine similarity of the sentence embedding and the gold labels. Spearman's rank correlation has been used to measure semantic textual similarity in other studies [19, 55, 116]. We show the performance in Table 6, 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 have fed these results to the table to provide an overview of the performance of cross-encoder networks for this type of task.

Models trained using the data augmentation strategy have a slightly higher coefficient than models trained using the traditional bi-encoder architecture. The Sentence-BERTimbau Augmented model, with a Spearman coefficient (SC) of 90.55, performed best in assessment when compared to the BERT Multilingual Augmented model, which had a coefficient of 90.33. Models without data augmentation, such as the Sentence-BERT Multilingual fine-tuned in-domain, had a coefficient of 89.21, while the BERT-Multilingual reached 83.97. For each epoch, we also calculated the Pearson coefficient.

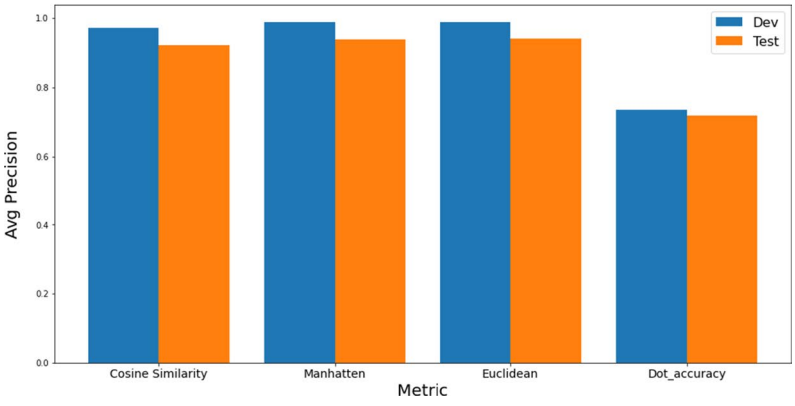
The second experiment aimed to evaluate the models by putting them through their paces on our data domain. We used all previous models as well as a new one



(a)



(b)



(c)

Fig. 14 Average precision metric for bi-encoder evaluation. **a** Sentence-BERTimbau Augmented. **b** Sentence-BERT Multilingual Augmented. **c** Sentence-BERT Paraphrase-Multilingual-MiniLM

for this test: Paraphrase-Multilingual-MiniLM. We chose this model because it performed great in another semantic search experiment. 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. Each of these metrics used average precision as the final coefficient.

We labeled our data using a cross-encoder BERTimbau Augmented as the model base for this evaluation. We chose this model due to it performing the best in previous experiments for Spearman and Pearson coefficients. We present the results in Fig. 14.

The combination of cross-encoder Sentence-BERTimbau Augmented with Sentence-BERT Paraphrase-Multilingual-MiniLM model performed best in the experiment. The joint use of the two networks, bringing in their background different vocabularies and training strategies, provided an average precision of 0.96. This combination was chosen as the best for incorporation into the HoPE architecture, which was evaluated in the next section.

6.3 HoPE Architecture Evaluation

In this section, we evaluate 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. 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, 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 [41]. In Fig. 15, 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 dataset. The HoPE model presented the best results from the F1-Measure analysis (0.896). With slightly worse performance, the Sentence-BERTimbau-Paraphrase-Multilingual reaches an F1-Score of (0.816). The BM25 lexical model obtained a coefficient (0.728) while the paraphrase model reaches (0.728). The results can be seen in Table 7.

The last experiment evaluates 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. For a fair comparison, we used the same set of phrases for each experiment, resetting the kernel every time after inference a model to finish.

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 paragraph embeddings of ANN are used. To understand the time difference between the two methods, we compare them. The results are shown

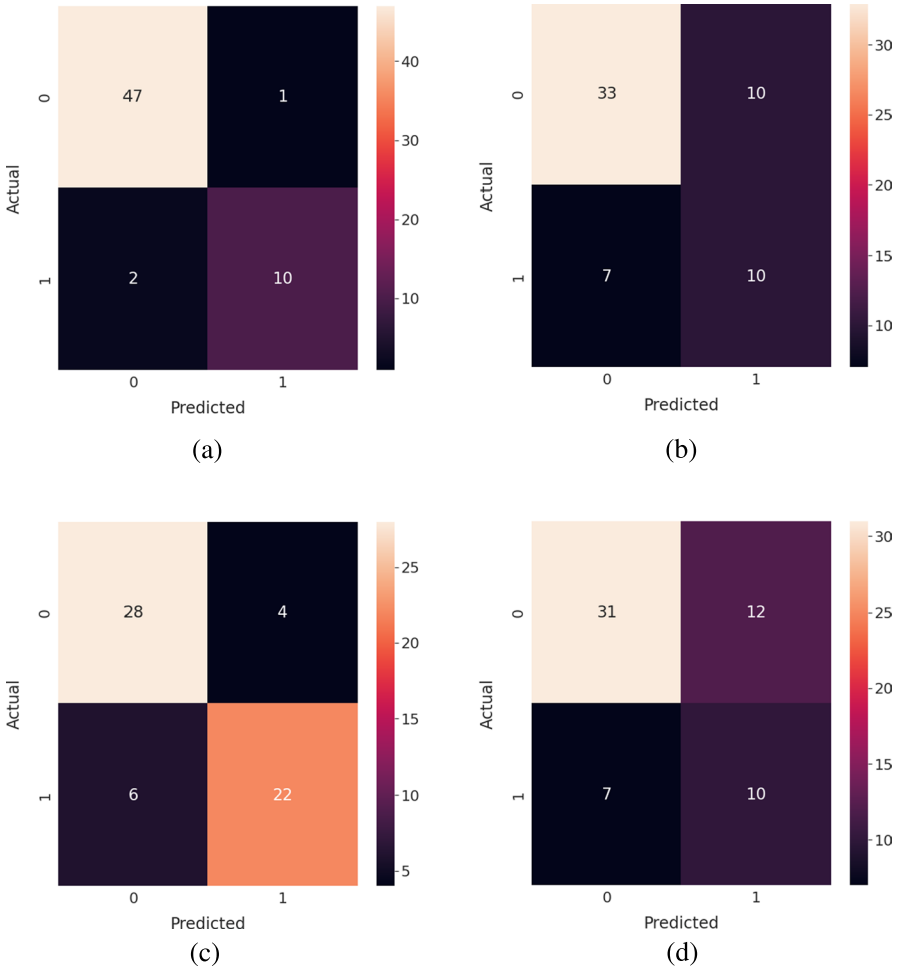


Fig. 15 Confusion matrix for test collection experiment. **a** HoPE model. **b** Sentence-BERTimbau-Paraphrase-Multilingual. **c** BM25 Okapi. **d** Paraphrase-Multilingual-MiniLM

Table 7 Evaluation of model's accuracy in performing information retrieval tasks using five *K*-folds

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

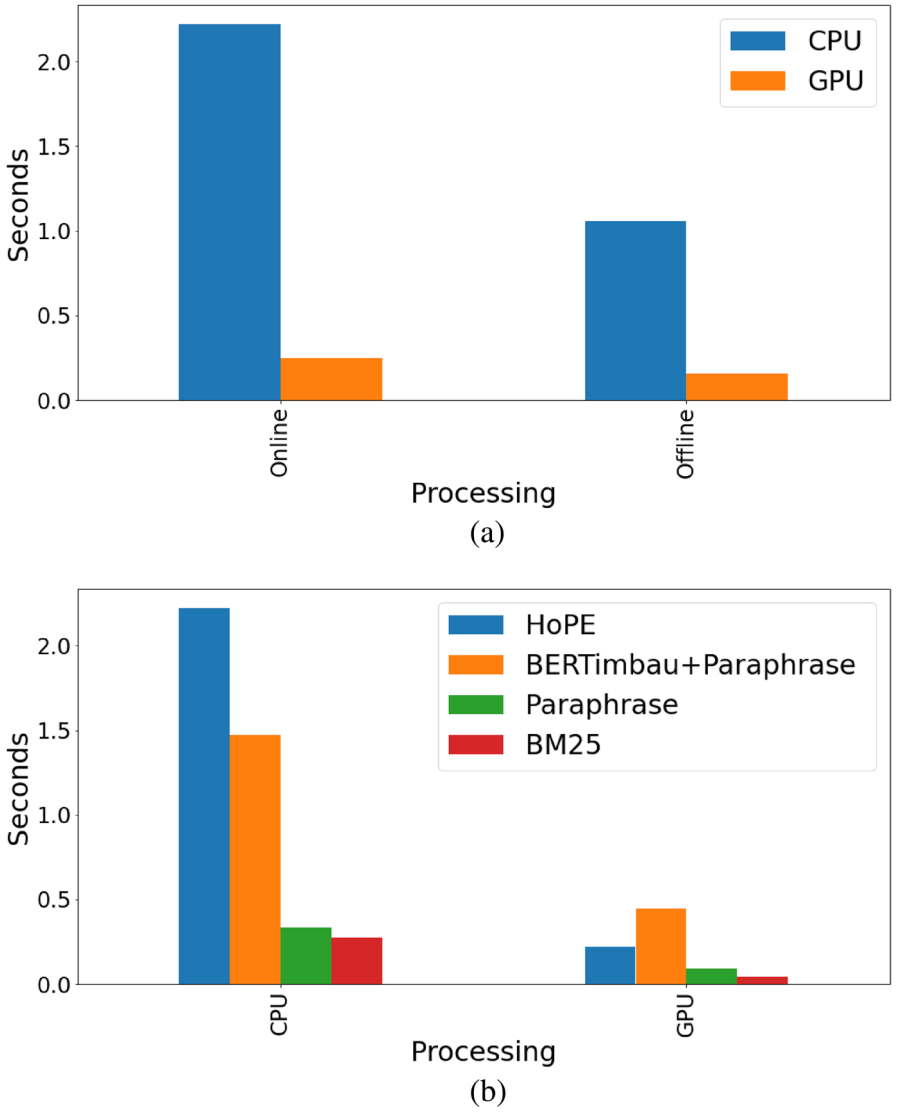


Fig. 16 HoPE performance for inference and encoding time. **a** Encoding performance. **b** Inference performance

in Fig. 16. Once compared to a CPU environment, the pre-computed embedding method outperformed the online method by an average of 2 s. When using a composite GPU environment, the difference shrinks.

In addition, we compared the inference speed of the models tested in previous experiments to the HoPE architecture. The experiment ran the inference exercise in the models for five sentences ranging in length from 0 to 100 characters and then calculated the average time to recover the answer. The results show that the HoPE

architecture takes longer to process than other models, with an average difference of 1 s for information retrieval on CPU-powered systems. The system improves significantly with GPU and has a time that is almost identical to the BERTimbau-Augmented model when used alone.

7 Discussion

This chapter will discuss the results of the experiments conducted. In contrast to the previous results section, we decide to address all of the results in the same section. We will review the findings in the order in which they were presented in the preceding chapter.

The first study evaluates the fine-tuned models on pregnancy guidelines. We assess the quality of incorporation in the initial experiment by fitting the models using sentence pairs through the siamese structure. The strategy of using pre-trained data has already been used in several studies with BERT networks in the health area [25, 109].

We validated our model's data augmentation strategy using a Brazilian Portuguese language benchmark and general knowledge transfer questions. This experiment began with fine-tuning in a cross-encoder network based on previously trained BERT models. The hyper-parameters of each model were individually adjusted for each model. At this point, we confirmed that the BERTImbau model performed best after being trained in eight epochs, whereas the BERT-Multilingual model over-fitted 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 important factors in the selection of this dataset. Multilingual models in the traditional BERT framework were fitted to data from ASSIN2 and someone else obtained great results [84]. Pre-trained BERT models had already demonstrated good performance in the ASSIN2 [18] data classification task, surpassing the state-of-art.

Spearman's and Pearson's correlation coefficient for augmented models was superior to non-augmented models. The results showed a significant difference between the Pearson coefficient for augmented and non-augmented models. Non-augmented data did not receive a fine-tuning of the benchmark data, which reinforces our belief that using a previous set can generate good results, improving the quality of embeddings. This behavior has also been corroborated by [21]. 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 embeddings. In this sense, we verified the performance gain with the transferring knowledge from the STS benchmark.

The next evaluation sought to fine-tune a bi-encoder over a dataset labeled from cross-encoders. The labeling method used was binary scores, with 0 for distant sentence pairs and 1 for close sentence pairs. 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-MiniLM model. This model is native in

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, and it performed well in this experiment, with the best overall average among the similarity metrics for both the test and validation sets. Similar results related to lower learning rate were observed in [49].

For cosine similarity, one of the main metrics for this assessment, the best average precision was for the Sentence-BERTimbau-Paraphrase-Multilingual model. In comparison to the obtained results, the Augmented BERTimbau model proved to be more consistent for this metric with a bi-encoder classification with cosine similarity reaching 90.55. 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.

When these two models are combined, they produce favorable results for response and re-rank recovery systems. The [101] study combines a cross-encoder with full attention over the input pair with bi-encoders that map each input independently to a dense vector space. The cross-encoder strategy makes no assumptions about the similarity scoring function that exists between the input and the candidate label. Instead, the concatenation of input and a candidate serves as a new input to a non-linear function that scores its match based on the desired dependencies [42]. In [42], a poly-encoder caches candidates for a given label, thereby determining a shorter inference time, while the cross-encoder extracts more information.

The HoPE model was evaluated for the inference task in our study using a 60-question dataset. The HoPE architecture is composed of two layers: a dialog management and comprehension layer comprised of the OntONeo ontology populated with preaching content extracted from guidelines, and a recently trained bi-cross model on Sentence-BERTimbau-Paraphrase-Multilingual.

To begin, the evaluation sought to illustrate the confusion matrix generated for the set of user query tests. The confusion matrix indicated that the HoPE model provided exceptional precision for user queries. The model performed admirably when it came to classifying hits and fallback, with a low percentage of false positives and negatives. All of the models evaluated made the most common classification errors with sentences: misspellings and extremely brief sentences.

The ambiguity or lack of context associated with short words may be a cause of poor prediction, as BERT models rely on attention mechanisms to identify relationships between the words contained in the sentence [105]. Sentences with rare terms in the corpus also had lower scores. The vocabulary with greater sentence variability for fine-tuned and pre-processing strategies should improve the coefficient predictions. Other assumptions raised in other discussions [95, 104] evaluate the combination of pre-trained bi-encoder models with a cross-encoder layer for this specific task of retrieval information.

All Sentence-BERT models had a similar performance for false negatives. The BM25 model however predicted few model hits and an expressive number of fallback. 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. Because the models were trained on a smaller set of in-domain data, we considered the results were reasonable. As with [38], future studies may benefit from a variety of examples to achieve a more accurate classification.

The Sentence-BERTimbau-Paraphrase-Multilingual 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 ontology's in chatbots already has known positive effects for natural language generation and information retrieval tasks [7, 66].

However, the use of this model combined with convolution networks for conversational agents is still little explored [93, 117]. 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.

The final experiment proposed in our article was the speed of inference. We propose a comparison of information retrieval encoding and inference methods. Loading offline pre-computed examples appeared to be faster than the online encoding of small datasets. Within its operation, the HoPE architecture employs both forms. It was discovered that, even with a large number of indexed paragraphs, the pre-computed approach via some indexing system such as ANN can provide twice the speed in CPUs and GPUs. This behavior has been observed in other systems where embedding pre-computed is up to ten times faster than online computing [118].

The use of online encoding, on the other hand, increases assertiveness due to the reduction of content indexed to the prediction model. Because it encodes more likely paragraphs for the input query, it is a viable option for conversational agents. Furthermore, the use of a GPU system increases the system's speed by three times. In addition, we compare the model inference times. This type of analysis is essential for conversational agents because response time is an important factor in end-user engagement [47].

The in-domain model's without augmentation on benchmark outperformed all others with query response times of less than 1 s. This experiment made use of the online encoding capability of the HoPE model. When a GPU was used, performance was increased by up to two times slower compared to other models.

Despite the inferior performance for timing performance, the HoPE architecture was within a 5-s threshold for inference, considered a benchmark for conversational agents. Despite the poor performance in terms of time performance, the HoPE architecture was within a 5-s threshold for a reply, considered a benchmark for conversation agents [43].

8 Conclusion

The main objective of this study was the proposition of the HoPE model architecture for the task of retrieving information for the period of pregnancy. HoPE is an architecture for conversational agents that uses in its structure ontology-based modeling and Sentence-BERT networks adjusted on pregnancy guidelines data to support pregnant women in obtaining more reliable information during the baby's thousand days period.

We seek to elucidate the functioning of the architecture and its components, followed by the evaluation methods. We propose two evaluations in our study: a comparative study between Sentence-BERT networks adjusted to pregnancy guidelines data and a study to evaluate the ability of the HoPE architecture to predict responses assertive compared to other information retrieval models.

In general terms, the evaluation of Sentence-BERT models adjusted to the corpus data presented a satisfactory result, mainly the models that used data augmentation strategies. These models have performed better in all of the proposed evaluations.

The HoPE architecture performed best in terms of assertiveness in our ablative study with other models. The use of ontology acted as a knowledge clustering mechanism for the corpus content, thus, helping to direct the dialogues to the domains involved in the user's question. The use of semantic models trained on large datasets supported this system to be more assertive from user inputs with no recognized entity. This study also presented a corpus in an unstructured data extraction pipeline and a method of structuring this set for training processes in Sentence-BERT networks.

Despite its advantages, the HoPE architecture was slow when used in an environment without GPUs, and it had a lower limited capacity for sentences with infrequent terms or spelling. The HoPE model demonstrated positive characteristics such as good predictive ability at a reasonable speed and being able to deal with a variety of input sentence structures. It did, however, have limitations when it came to dealing with infrequent words in the dataset and lexicons. We believe that a fine-tuning process with higher term variability and a supervised content grouping phase can significantly improve our architecture.

We have some ideas for future studies: firstly, we want to look at strategies for dealing with imperative and entity-less questions. For example, the phrase "can you tell me more about this?" lacks a defined subject entity and relies on previous contexts to be resolved, still, a strategy for detecting multiple intentions, which we see as a challenge that has received little attention in the context of conversational agents [5, 88]. Another objective is to carry out modifications in the cross-encoder training phase, using a strategy with weighted scores instead of binary scores. Finally, through the development of a conversational agent tool, we validated our architecture in a clinical trial with pregnant women and health professionals.

Acknowledgements The authors would like to thank the Brazilian National Council for Scientific and Technological Development - CNPq (Grant Numbers 303640/2017-0 and 405354/2016-9) for supporting this work.

References

1. Alambo A, Padhee S, Banerjee T, Thirunarayan K (2021) COVID-19 and mental health/substance use disorders on reddit: a longitudinal study. In: International conference on pattern recognition, pp 20–27. Springer
2. Alfeo AL, Cimino MG, Vaglini G (2021) Technological troubleshooting based on sentence embedding with deep transformers. J Intell Manuf 32:1–12
3. ALMarwi H, Ghurab M, Al-Baltah I (2020) A hybrid semantic query expansion approach for Arabic information retrieval. J Big Data 7(1):1–19
4. Alomari A, Idris N, Sabri AQM, Alsmadi I (2021) Deep reinforcement and transfer learning for abstractive text summarization: a review. Comput Speech Lang 71:101276
5. Altinok D (2018) An ontology-based dialogue management system for banking and finance dialogue systems. arXiv:1804.04838
6. Amith M, Anna Z, Cunningham R, Rebecca L, Savas L, Laura S, Yong C, Yang G, Julie B, Roberts K et al (2019) Early usability assessment of a conversational agent for HPV vaccination. Stud Health Technol Inform 257:17
7. Avila CVS, Calixto AB, Rolim TV, Franco W, Venceslau AD, Vidal VM, Pequeno VM, De Moura FF (2019) MediBot: an ontology based chatbot for Portuguese speakers drug's users
8. Bakouan M, Kone T, Kamagate BH, Oumtanaga S, Babri M (2018) A chatbot for automatic processing of learner concerns in an online learning platform. Int J Adv Comput Sci Appl 9(5):168–176
9. Barbosa A, Godoy A (2021) Augmenting customer support with an NLP-based receptionist. arXiv:2112.01959
10. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 3615–3620. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1371. https://aclanthology.org/D19-1371
11. Benesty J, Chen J, Huang Y, Cohen I (2009) Pearson correlation coefficient. In: Noise reduction in speech processing, pp 1–4. Springer
12. Bickmore TW, Pfeifer LM, Byron D, Forsythe S, Henault LE, Jack BW, Silliman R, Paasche-Orlow MK (2010) Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 15(S2):197–210
13. Bickmore TW, Pfeifer LM, Jack BW (2009) Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1265–1274. ACM
14. Bickmore TW, Utami D, Matsuyama R, Paasche-Orlow MK (2016) Improving access to online health information with conversational agents: a randomized controlled experiment. J Medical Internet Res 18(1):e1
15. Bjelke M, Martinsson AK, Lendahls L, Oscarsson M (2016) Using the Internet as a source of information during pregnancy-a descriptive cross-sectional study in Sweden. Midwifery 40:187–191
16. Boonstra L (2021) Getting started with dialogflow essentials. In: The definitive guide to conversational AI with dialogflow and google cloud, pp 29–57. Springer
17. Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, Dai L (2021) ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. Complexity
18. Cabezudo MAS, Inácio M, Rodrigues AC, Casanova E, de Sousa RF (2019) NILC at ASSIN 2: exploring multilingual approaches. In: ASSIN@ STIL, pp 49–58
19. Carlsson F, Gogoulou E, Ylipää E, Gyllensten AC, Sahlgren M (2021) Semantic re-tuning with contrastive tension. In: International conference on learning representations
20. Chang WC, Yu HF, Zhong K, Yang Y, Dhillon I (2019) Taming pretrained transformers for extreme multi-label text classification. arXiv:1905.02331
21. Choi H, Kim J, Joe S, Gwon Y (2021) Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. In: 2020 25th International conference on pattern recognition (ICPR), pp 5482–5487. IEEE
22. Consortium WWW, et al (2014) Rdf 1.1 concepts and abstract syntax

23. Criss S, Baidal JAW, Goldman RE, Perkins M, Cunningham C, Taveras EM (2015) The role of health information sources in decision-making among Hispanic mothers during their children's first 1000 days of life. Matern Child Health J 19(11):2536–2543
24. Croux C, Dehon C (2010) Influence functions of the Spearman and Kendall correlation measures. Statistical Methods & Applications 19(4):497–515
25. Dai Z, Wang X, Ni P, Li Y, Li G, Bai X (2019) Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. In: 2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei), pp 1–5. IEEE
26. Deng X, Liu Q, Deng Y, Mahadevan S (2016) An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences 340:250–261
27. Devlin J, Chang MW, Lee K, Toutanova K (2019) 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
28. Emygdio JL, Almeida MB (2019) Representações formais do conhecimento aplicadas à interoperabilidade semântica de terminologias clínicas. Múltiplos Olhares em Ciência da Informação **9**(2)
29. Engelmann D, Damasio J, Krausburg T, Borges O, Colissi M, Panisson AR, Bordini RH (2021) Dial4jaca–a communication interface between multi-agent systems and chatbots. In: International conference on practical applications of agents and multi-agent systems, pp 77–88. Springer
30. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, Socher R (2021) COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. NPJ Digital Medicine 4(1):1–9
31. Farinelli: ONTONEO (2018). http://bioportal.bioontology.org/ontologies/ONTONEO
32. Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Minds and Machines 30(4):681–694
33. Fushiki T (2011) Estimation of prediction error by using k-fold cross-validation. Statistics and Computing 21(2):137–146
34. Ganhotra J, Roitman H, Cohen D, Mills N, Gunasekara C, Mass Y, Joshi S, Lastras L, Konopnicki D (2020) Conversational document prediction to assist customer care agents. arXiv:2010.02305
35. Goeuriot L, Jones GJ, Kelly L, Müller H, Zobel J (2016) Medical information retrieval: introduction to the special issue. Information Retrieval Journal 19(1–2):1–5
36. Greche L, Jazouli M, Es-Sbai N, Majda A, Zarghili A (2017) Comparison between Euclidean and Manhattan distance measure for facial expressions classification. In: 2017 International conference on wireless technologies, embedded and intelligent systems (WITS), pp 1–4. IEEE
37. Guo J, Fan Y, Pang L, Yang L, Ai Q, Zamani H, Wu C, Croft WB, Cheng X (2020) A deep look into neural ranking models for information retrieval. Information Processing & Management 57(6):102067
38. Han X, Eisenstein J (2019) Unsupervised domain adaptation of contextualized embeddings for sequence labeling. arXiv:1904.02817
39. Henderson M, Al-Rfou R, Strope B, Sung YH, Lukács L, Guo R, Kumar S, Miklos B, Kurzweil R (2017) Efficient natural language response suggestion for smart reply. arXiv:1705.00652
40. Hersh W, Hersh W (2020) Information retrieval: a biomedical and health perspective. Springer
41. Huertas-García Á, Huertas-Tato J, Martín A, Camacho D (2021) Countering misinformation through semantic-aware multilingual models. In: International conference on intelligent data engineering and automated learning, pp 312–323. Springer
42. Humeau S, Shuster K, Lachaux MA, Weston J (2019) Poly-encoders: transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv:1905.01969
43. Inamdar VA, Shivanand R (2019) Development of college enquiry chatbot using snatchbot. DEVELOPMENT 6(07):1615–1618
44. Jack B, Bickmore T, Hempstead M, Yinusa-Nyahkoon L, Sadikova E, Mitchell S, Gardiner P, Adigun F, Penti B, Schulman D et al (2015) Reducing preconception risks among African American women with conversational agent technology. The Journal of the American Board of Family Medicine 28(4):441–451
45. Júnior VODS, Branco JAC, De Oliveira MA, Da Silva TLC, Cruz LA, Magalhaes RP (2021) A natural language understanding model COVID-19 based for chatbots. In: 2021 IEEE 21st International conference on bioinformatics and bioengineering (BIBE), pp 1–7. IEEE

46. Kadri Y, Nie JY (2006) Effective stemming for Arabic information retrieval. In: Proceedings of the challenge of arabic for NLP/MT conference, Londres, Royaume-Uni, pp 68–74
47. Kankaria RV, Agrawal A, Barot H, Godbole A (2021) RAAH.ai: an interactive chatbot for stress relief using deep learning and natural language processing. In: 2021 12th International conference on computing communication and networking technologies (ICCCNT), pp 1–7. IEEE
48. Kasilingam DL (2020) Understanding the attitude and intention to use smartphone chatbots for shopping. Technology in Society 62:101280
49. Khattab O, Zaharia M (2020) 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, pp 39–48
50. Kilmer S, Marshall C, Senger S (2020) Dot product chains. arXiv:2006.11467
51. Kim T, Yoo KM, Lee SG (2021) Self-guided contrastive learning for BERT sentence representations. arXiv:2106.07345
52. Lagan BM, Sinclair M, George Kernohan W (2010) Internet use in pregnancy informs women's decision making: a web-based survey. Birth 37(2):106–115
53. Larkey LS, Ballesteros L, Connell ME (2002) 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, pp 275–282
54. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
55. Li B, Zhou H, He J, Wang M, Yang Y, Li L (2020) On the sentence embeddings from pre-trained language models. arXiv:2011.05864
56. Li D, Zeng W (2018) Distance measure of Pythagorean fuzzy sets. Int J Intell Syst 33(2):348–361
57. Liu X, Wang Y, Ji J, Cheng H, Zhu X, Awa E, He P, Chen W, Poon H, Cao G, et al (2020) The microsoft toolkit of multi-task deep neural networks for natural language understanding. arXiv:2002.07972
58. Lv Y, Zhai C (2011) When documents are very long, BM25 fails! In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, pp 1103–1104
59. Malizia A, Onorati T, Diaz P, Aedo I, Astorga-Paliza F (2010) SEMA4A: an ontology for emergency notification systems accessibility. Expert Systems with Applications 37(4):3380–3391
60. Maroengsit W, Piyakulpinyo T, Phonyiam K, Pongnumkul S, Chaovalit P, Theeramunkong T (2019) A survey on evaluation methods for chatbots. In: Proceedings of the 2019 7th International conference on information and education technology, pp 111–119
61. Mellado-Silva R, Faúndez-Ugalde A, Lobos MB (2020) Learning tax regulations through rules-based chatbots using decision trees: a case study at the time of COVID-19. In: 2020 39th International conference of the chilean computer science society (SCCC), pp 1–8. IEEE
62. Mikolov T, Yih WT, Zweig G (2013) 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, pp 746–751
63. Montenegro JLZ, da Costa CA, da Rosa Righi R (2019) Survey of conversational agents in health. Expert Syst Appl 129:56–67
64. Moyer CA, Compton SD, Kaselitz E, Muzik M (2020) Pregnancy-related anxiety during COVID-19: a nationwide survey of 2740 pregnant women. Arch Women's Mental Health 23(6):757–765
65. Mukherjee S, Liu X, Zheng G, Hosseini S, Cheng H, Yang G, Meek C, Awadallah AH, Gao J (2021) CLUES: few-shot learning evaluation in natural language understanding. arXiv:2111.02570
66. Nazir A, Khan MY, Ahmed T, Jami SI, Wasi S (2019) A novel approach for ontology-driven information retrieving chatbot for fashion brands. Int J Adv Comput Sci Appl IJACSA 10(9):546–552
67. Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: a human generated machine reading comprehension dataset. In: CoCo@ NIPS
68. Nouri SS, Rudd RE (2015) Health literacy in the oral exchange: an important element of patient-provider communication. Patient Education and Counseling 98(5):565–571
69. Noy NF, McGuinness DL, et al (2001) Ontology development 101: a guide to creating your first ontology
70. Oliveira LE, Gebeluca CP, Silva AM, Moro CM, Hasan SA, Farri O (2017) A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents. In: 2017 IEEE International conference on bioinformatics and biomedicine (BIBM), pp 1072–1078. IEEE

71. Padaki R, Dai Z, Callan J (2020) Rethinking query expansion for BERT reranking. In: European conference on information retrieval, pp 297–304. Springer
72. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y (2019) Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. Journal of medical Internet research 21(4):e12887
73. Patel V, Garrison P, de Jesus Mari J, Minas H, Prince M, Saxena S (2008) The Lancet's series on global mental health: 1 year on. The Lancet 372(9646):1354–1357
74. Pereira I, Sousa J, Costa PB, Barbosa SD, Colcher S (2020) SucupiraBot: an interactive question-answering system for the Sucupira platform. In: Proceedings of the brazilian symposium on multimedia and the web, pp 277–280
75. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv:1802.05365
76. Podgorny I, Khaburzaniya Y, Geisler J (2019) Conversational agents and community question answering. In: CHI 2019 Workshops, Glasgow, United Kingdom
77. Qian Y, Liang J, Dang C (2009) Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. Int J Approximate Reason 50(1):174–188
78. Quamar A, Özcan F, Miller D, Moore RJ, Niehus R, Kreulen J (2020) Conversational BI: an ontology-driven conversation system for business intelligence applications. Proceedings of the VLDB Endowment 13(12):3369–3381
79. Ragab M, Abdel Aal AM, Jifri AO, Omran NF (2021) Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques. Wireless Communications and Mobile Computing 2021
80. Raji P, Surendran S (2016) RDF approach on social network analysis. In: 2016 International conference on research advances in integrated navigation systems (RAINS), pp 1–4. IEEE
81. Rajosoa M, Hantach R, Abbes SB, Calvez P (2019) Hybrid question answering system based on natural language processing and SPARQL query
82. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text
83. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using siamese BERT-networks. arXiv:1908.10084
84. Rodrigues R, Couto P, Rodrigues I (2019) IPR: the semantic textual similarity and recognizing textual entailment systems. In: ASSIN@ STIL, pp 39–48
85. Rodrigues RC, Rodrigues J, de Castro, PVQ, da Silva NFF, Soares A (2020) Portuguese language models and word embeddings: evaluating on semantic similarity tasks. In: International conference on computational processing of the portuguese language, pp 239–248. Springer
86. Rogers A, Kovaleva O, Rumshisky A (2021) A primer in BERTology: what we know about how BERT works. Transactions of the Association for Computational Linguistics 8:842–866
87. Rubinstein A (2018) Hardness of approximate nearest neighbor search. In: Proceedings of the 50th annual ACM SIGACT symposium on theory of computing, pp 1260–1268
88. Rychalska B, Glabska H, Wroblewska A (2018) Multi-intent hierarchical natural language understanding for chatbots. In: 2018 Fifth international conference on social networks analysis, management and security (SNAMS), pp 256–259. IEEE
89. Safder I, Hassan SU (2019) Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. Scientometrics 119(1):257–277
90. Samimi P, Ravana SD (2014) Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: a review. Sci World J 2014
91. Sankhavara J (2020) Feature weighting in finding feedback documents for query expansion in biomedical document retrieval. SN Computer Science 1(2):1–7
92. Sayakhot P, Carolan-Olah M (2016) Internet use by pregnant women seeking pregnancy-related information: a systematic review. BMC Pregnancy and Childbirth 16(1):1–10
93. Senese MA, Rizzo G, Dragoni M, Morisio M (2020) MTSI-BERT: a session-aware knowledge-based conversational agent. In: Proceedings of The 12th language resources and evaluation conference, pp 717–725
94. Sheth A, Yip HY, Iyengar A, Tepper P (2019) Cognitive services and intelligent chatbots: current perspectives and special issue introduction. IEEE Internet Computing 23(2):6–12

95. Singh I, Scarton C, Bontcheva K (2021) Multistage BiCross encoder: team GATE entry for MLIA multilingual semantic search task 2. arXiv:2101.03013
96. Singh S, Mahmood A (2021) The NLP cookbook: modern recipes for transformer based deep learning architectures. IEEE Access 9:68675–68702
97. Souza F, Nogueira R, Lotufo R (2020) BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Brazilian conference on intelligent systems, pp 403–417. Springer
98. de Souza JVA, Oliveira LESE, Gumiel YB, Carvalho DR, Moro CMC (2020) Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages. In: International conference on computational processing of the portuguese language, pp 357–367. Springer
99. Tanana MJ, Soma CS, Srikumar V, Atkins DC, Imel ZE (2019) Development and evaluation of ClientBot: patient-like conversational agent to train basic counseling skills. J Med Int Res 21(7):e12529
100. Teixeira, M.S., Maran, V., Dragoni, M (2021) The interplay of a conversational ontology and AI planning for health dialogue management. In: Proceedings of the 36th annual ACM symposium on applied computing, pp 611–619
101. Thakur N, Reimers N, Daxenberger J, Gurevych I (2020) Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. arXiv:2010.08240
102. Traylor CS, Johnson J, Kimmel MC, Manuck TA (2020) Effects of psychological stress on adverse pregnancy outcomes and non-pharmacologic approaches for reduction: an expert review. Am J Obstet Gynecol MFM, 100229
103. Trivedi S, Gildersleeve R, Franco S, Kanter AS, Chaudhry A (2020) Evaluation of a concept mapping task using named entity recognition and normalization in unstructured clinical text. J Healthc Inform Res 4(4):395–410
104. Vakili Tahami A, Ghajar K, Shakery A (2020) Distilling knowledge for fast retrieval-based chatbots. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in information retrieval, pp 2081–2084
105. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv:1706.03762
106. Wagner Filho JA, Wilkens R, Idiart M, Villavicencio A (2018) The brWaC corpus: a new open resource for Brazilian Portuguese. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)
107. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: International conference on learning representations. https://openreview.net/forum?id=rJ4km2R5t7
108. Wang K, Yang B, Xu G, He X (2019) Medical question retrieval based on siamese neural network and transfer learning method. In: International conference on database systems for advanced applications, pp 49–64. Springer
109. Wang Y, Liu F, Verspoor K, Baldwin T (2020) 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, pp 105–111
110. Wang Y, Rastegar-Mojarad M, Elayavilli RK, Liu S, Liu H (2016) An ensemble model of clinical information extraction and information retrieval for clinical decision support. In: TREC
111. Whissell JS, Clarke CL (2011) Improving document clustering using Okapi BM25 feature weighting. Information Retrieval 14(5):466–487
112. Wu Z, Liang J, Zhang Z, Lei J (2021) Exploration of text matching methods in Chinese disease q&a systems: a method using ensemble based on BERT and boosted tree models. J Biomed Inform 115:103683
113. Xie Y, Yang W, Tan L, Xiong K, Yuan NJ, Huai B, Li M, Lin J (2020) Distant supervision for multi-stage fine-tuning in retrieval-based question answering. Proceed Web Conference 2020:2934–2940
114. Yang W, Xie Y, Lin A, Li X, Tan L, Xiong K, Li M, Lin J (2019) End-to-end open-domain question answering with BERTserini. arXiv:1902.01718
115. Yang W, Xie Y, Tan L, Xiong K, Li M, Lin J (2019) Data augmentation for BERT fine-tuning in open-domain question answering. arXiv:1904.06652
116. Yin X, Zhang W, Zhu W, Liu S, Yao T (2020) Improving sentence representations via component focusing. Applied Sciences 10(3):958

117. Yoo S, Jeong O (2020) An intelligent chatbot utilizing BERT model and knowledge graph. J Soc e-Business Stud 24(3)
118. Yoon S, Kang WY, Jeon S, Lee S, Han C, Park J, Kim ES (2020) Image-to-image retrieval by learning similarity between scene graphs. arXiv:2012.14700
119. Zhang J, et al (2021) S-SimCSE: sampled sub-networks for contrastive learning of sentence embedding. arXiv:2111.11750
120. Zhang Z, Bickmore TW, Paasche-Orlow MK (2017) Perceived organizational affiliation and its effects on patient trust: role modeling with embodied conversational agents. Patient Educ Couns 100(9):1730–1737