Informatics in Medicine Unlocked 43 (2023) 101381

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Exploring named entity recognition and relation extraction for ontology and medical records integration

Diego Pinheiro da Silva [a,*], William da Rosa Fröhlich [b], Blanda Helena de Mello [a], Renata Vieira [c], Sandro José Rigo [a]

[a] Universidade do Vale do Rio dos Sinos, Av. Unisinos, 950, Cristo Rei, São Leopoldo, 93022-750, Brazil
[b] Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 90619-900, Brazil
[c] Universidade de Évora, Largo dos Colegiais 2, 7004-516, Évora, Portugal

ARTICLE INFO

Keywords:
Named entity recognition
Deep learning
Electronic health record
Relation extraction
Natural language processing
Ontology

ABSTRACT

The available natural language data in electronic health records is of noteworthy interest to health research and development. Nevertheless, their manual analysis is not feasible and poses a challenge to accessing valuable information in these records. This paper presents an approach to automatically extract information from these unstructured medical records using Domain Entity Recognition and Relation Extraction, structuring the results through a domain ontology. We developed our work in the oncology domain, an attention-demanding field. The main contribution of this work lies in integrating multiple resources in a complete methodology to accomplish this task. We developed a new entity and relation annotated dataset of medical evolutions in Brazilian Portuguese, containing 1622 documents, 146,769 entities, and 111,716 relations. We attained 78.24 % accuracy for entity and relation extraction in the exams domain. Healthcare specialists evaluated the approach regarding entity recognition and relation extraction positively and considered the methodology valuable to health professionals.

1. Introduction

Medical records typically describe health events by statements expressed in natural language. We observe these records transitioning to a digital format [1], therefore fostering the existence of large datasets of valuable health information. The data within Electronic Health Records (EHRs), such as clinical reports and medical observations, can be used for disease registries, epidemiological studies, drug safety surveillance, clinical trials, and health audits, among others [2,3]. Approaches in which medical information is manually extracted from patient records by clinical experts have an evident limitation in scalability and time, besides high costs [4,5,3]. In this context, EHRs' possibilities for data reuse have fostered a necessity to process free-text narratives automatically. This procedure enables new possibilities to support healthcare professionals [6,7].

Named Entity Recognition (NER) and Relation Extraction (RE) techniques have emerged as promising approaches for automatically extracting valuable information from unstructured health data records. NER involves identifying named entities from medical annotations, such as patient names, disease names, and medication names. As a complement, RE involves identifying and extracting relations connecting these entities, such as patient-diagnosis or medication-dosage relations. Both techniques have been increasingly employed in healthcare applications to enhance information retrieval, decision-making, and patient care [8,9,10,11,12].

Particular techniques can be applied to support NER and RE, such as BERT or BI-LSTM models. BERT models have shown promising results in Natural Language Processing (NLP) tasks focused on NER, particularly in specific domains, such as health [9,13,14,15]. The BiLSTM model can leverage forward and backward contextual information to capture complex patterns in RE [16,17]. Although the BERT and the BI-LSTM models are widely adopted, exploring these in Portuguese-speaking scenarios still requires more research and presents limitations regarding available datasets.

Although well-known methods for extracting information exist, there is a lack of methodologies dedicated to integrating different and complementary techniques. Hence, we developed a method for integrating natural language data in a domain ontology for oncology. The approach

* Corresponding author.
E-mail address: diego192@edu.unisinos.br (D.P. da Silva).

https://doi.org/10.1016/j.imu.2023.101381
Received 22 August 2023; Received in revised form 9 October 2023; Accepted 16 October 2023
Available online 21 October 2023

involves Deep Learning techniques, with Transformers models like BERT, and the Bi-LSTM model, for implementing tasks of named entity recognition and relation extraction. We integrated the extracted data into an ontology.

The methodology developed in this work contributed several aspects to the area. First, by highlighting the composition of unpublished datasets with entities and relations of medical evolutions, containing 1622 annotated documents, 146,769 entities, and 111,716 relations. Second, by creating an ontology structure in the oncology domain to structure related medical data. Third, by proposing an approach to integrating resources for NER and RE with the domain ontology. It is significant to highlight the training of real oncology data in Portuguese as a differential. The results pave the way for future applications that may benefit the medical field, providing a structured and reliable knowledge base to help health professionals and researchers.

The remainder of this text is organized as follows. Section 2 presents related works. Section 3 describes the research context. Section 4 describes the construction of datasets in Portuguese from real data. In section 5, we describe the experiments. Section 6 illustrates and discusses the results of the experiments. The conclusions and future work indications are presented in section 7.

2. Related works

We conducted a systematic literature review to promote a deeper understanding of related work scenarios. After filtering 961 publications, we selected and analyzed 23 studies on Entity Recognition, Relation Extraction, and medical information integration. We developed a protocol combining a protocol referring to the computing area [18] with a protocol directed to the health area [19]. This method increased the assertiveness of the results compared with those obtained from direct literature reviews [20].

The review identified the broad investigation of BERT-based models for NER and RE in the health domain due to their effectiveness [9,14]. One key advantage of BERT models is the ability to fine-tune specific datasets, which has proven beneficial in scenarios with limited annotated data [21,22]. Transformer-based approaches can also be practical in biomedical domain data mining [23]. Although BERT serves as a versatile model for various NLP tasks, specific models such as BioBERT [21], presents excelent results for applications in the biomedical domain. Therefore emphasizing the relevance and impact of transformer models in NER and RE research [24]. The model conception can benefit from a broader context, as presented in MedBERT [25], a contextualized embedding model trained on the MedNLI medical natural language inference dataset. MedBERT improves NER and RE tasks by capturing domain-specific context, showing significant gains in accuracy. Furthermore, domain-specific contextualized embeddings are tailored to the nuances of medical language, making them valuable resources for information extraction in EHRs [26].

Although the study showed relevant works regarding using BERT, we identified some research gaps. The first is the Portuguese language support, which currently has few available resources. The second refers to the types of data used, since most works utilizing data outside the health area or not originated from real contexts. The third issue deals with integrating data based on formal resources, such as ontologies. This integration can enhance information extraction by providing a more structured and semantically meaningful representation of entities and their relation.

3. Research context

In this section, we describe the main elements of the proposed methodology for integrating natural language data in a domain ontology for oncology. To carry out this study, we partnered with a software company with an oncology EHR system called the Gemed Oncology System. The company works with clinics and hospitals, providing solutions for the electronic recording of medical, nursing, and pharmaceutical care data. Therefore, in this research, we developed a case study in the oncology domain fostered by real data and systems from this partner company.

Fig. 1 presents an overview of our research context, considering medical clinics or hospitals and characterizing the needs of health professionals. The adopted methodology considers some key elements, which are the availability of a medical record generated by a healthcare professional in an EHR system, the processing of these medical records in natural language for automatically named entities and relation extraction, and finally, the insertion of the extracted data in a domain ontology. The main objective of this approach is to support health professionals in better use of their medical notes.

The first step [1] in our methodology considers the work of health specialists registering patients' evolutions and the results of these actions. The relevant information for the patient's treatment is registered in the EHR GEMED system in the form of clinical notes composed of free text and structured data. The system's medical notes contain clinical events registered by healthcare professionals. The healthcare professional uses a customizable system form to record each clinical event information. The records are stored in a database, as the next step [2]. All patient health information is stored, such as medical notes, electronic prescriptions, and patient consultation history.

The next step of methodology [3] involves automatic tasks for data preparation for automatic named entity recognition and relation extraction. One of the tasks is data anonymization, which is necessary to remove the association between the identifying dataset and the data subject. Another task is the general pre-processing. This research applies an anonymization process to avoid the disclosure and unique identification of patients and health professionals. Once finalized, the data is ready for annotation and dataset generation. In the data annotation process [4], this research had the support of Interprocess GEMED and Universidade do Vale do Rio dos Sinos health experts [5]. The annotated data was extracted from the GEMED database to annotate the clinical data trials according to previously defined labels [6]. The annotated dataset was called Dataset Gemed Onco (DGO).

Next, in the third step, automatic data extraction is performed [7]. The architecture of our approach consists of integrating two models: BERT and Bi-LSTM. We trained BERT to perform the recognition of entities [8] and extraction of features [9] that will represent the interest elements of the text. On the other hand, Bi-LSTM is responsible for classifying the relation between the recognized entities.

Once the information extraction step is completed, we propose the data representation step by structuring the extracted information in a domain ontology [10]. We opted for constructing and using an oncology domain ontology to represent the entities, concepts, and specific relations identified in the information extraction. With the data structured, the EHR system consults the ontology [11] and uses its data for distinct tasks, mainly data query, but also as support for diagnosis or event prediction [12].

We defined an assessment approach to evaluate the model in two contexts: in the health area and computational aspects. Regarding the health area, specialists in the oncology domain validated the model, and its results in applied case studies. Experiments were carried out with data collection in questionnaires and follow-ups with professionals to validate their perceptions regarding using the model. Concerning the computational context, we used standardized metrics to evaluate the model responses, using Accuracy, Recall, F1-Score, Macro AVG, and Weighted AVG.

4. Developed datasets

We produced two datasets. The first one describes relations in exams. Interprocess professionals and students in the last semester of medicine at the Universidade do Vale do Rio dos Sinos annotated this dataset. The second one aims at the domain of Diagnostic Features of breast cancer,

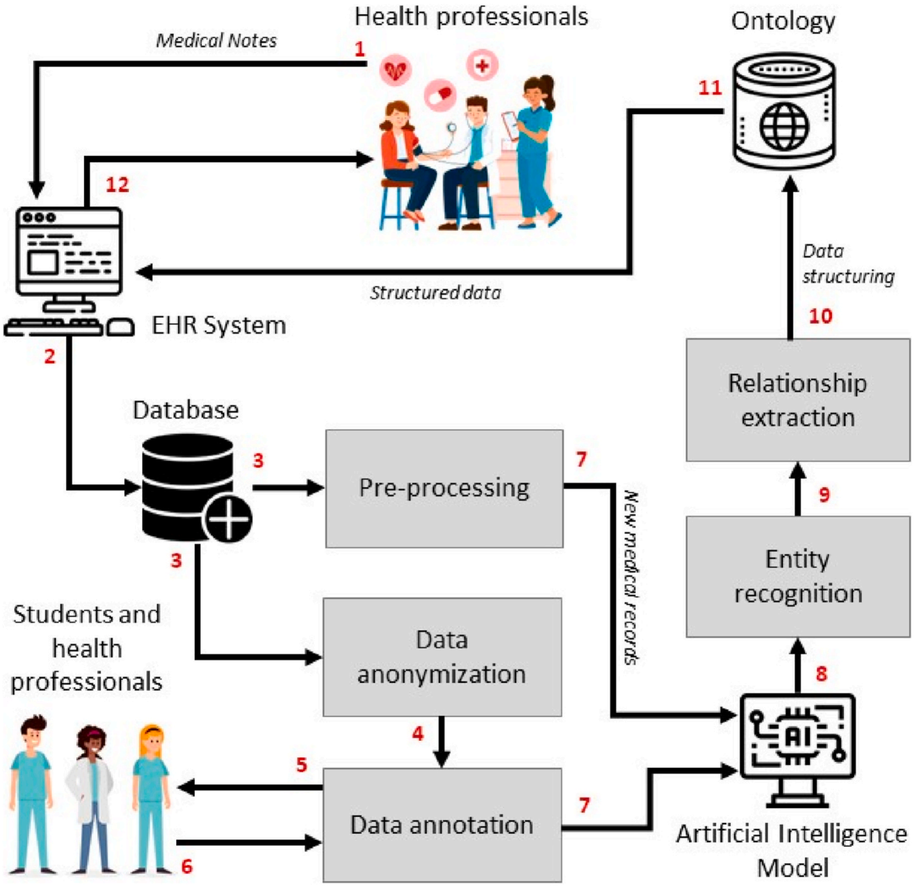


Fig. 1. Overview of the research context.

and health specialists annotated it. The research ethics committee evaluation of the methodology was not required in this case because, in the overall process, the original used data is anonymized, and there is no data association with the patients, leading to no individual exposure.

For constructing the first dataset, the Gemed Onco - Exams, here mentioned as DGO-E, we studied 1022 documents on the evolution of an oncology clinic. These documents contain real data prescribed by healthcare professionals. Before using the documents, they underwent an anonymization process in which we removed any names or identifications of health professionals or patients. Meetings were held with oncology specialists to define the focus of the construction of the data set. We decided to carry out an approach with exam data, as it is possible to identify these data in the text, which are essential for a company prototype.

The following entities of interest were defined as the focus of data annotation for the DGO-E dataset:

- *Exame* (Exam): Indicates the name of the exam. Example: PSA;
- *Resultado* (Result): Specifies the result values of an exam. Examples: 193mil (193 thousand), 1547;
- *Data* (Date): Informs the date of an exam. Examples: 07/18/2022 (2022/18/07), 07/2022 (2022/07);
- *Membro* (Member): Annotates a member of the human body, seeking to identify the member on which the exam was performed. Examples: Crânio (Skull), lombar (lumbar);
- *Tempo* (Time): Indicates the time of an action, trying to identify how long ago an exam was performed. Examples: 1 mês atrás (1 month ago), no último ano (last year).

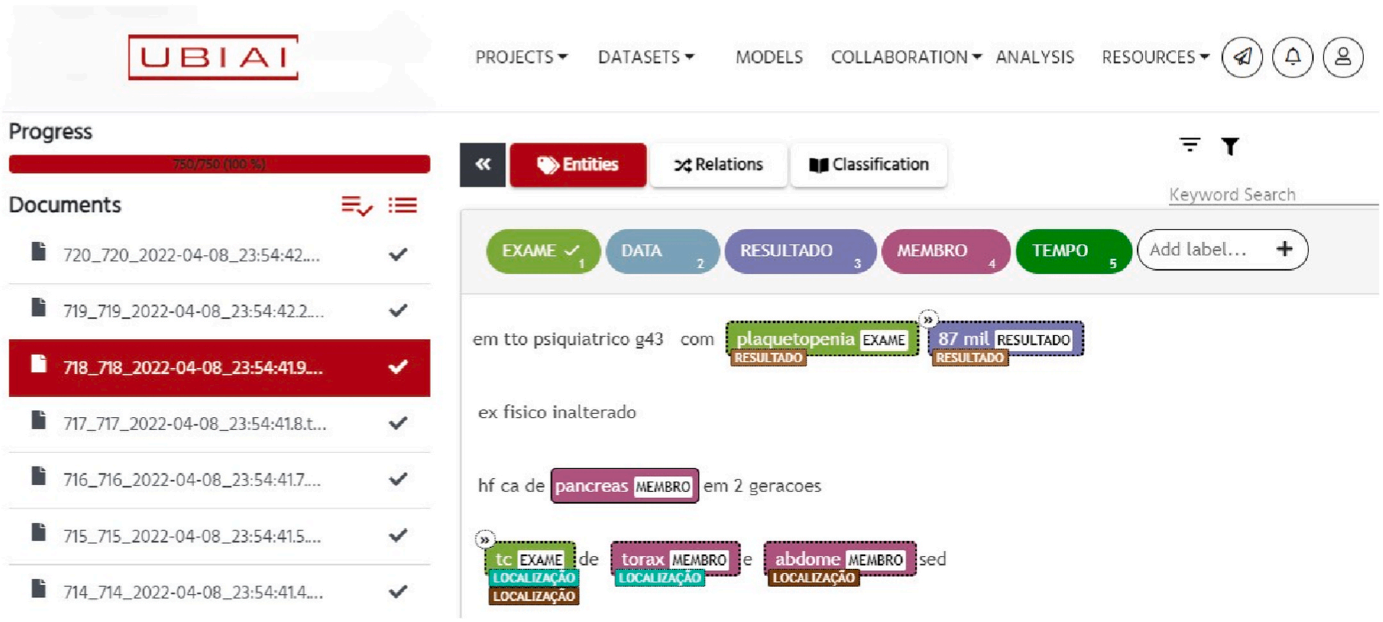
The following relations of interest for the annotation were defined:

- *Resultado* (Result): Relates the result value of an exam to the kind of exam;
- *Quando* (When): Relates the date and time of an exam to the exam;
- *Localização* (Location): Lists a member to the exam.

After defining the necessary parameters, the annotation construction process for the dataset began. For this, we selected three scholarship students who are at the end of the medical course at Unisinos and have knowledge in the area, in addition to a nurse and a biomedical doctor, both collaborators of Empresa Interprocess. We used UBIAI Tools[1] software to take notes. This tool enables collaboration in taking notes and has an intuitive interface. With UBIAI Tools, it is possible to perform manual annotations of entities, relations, and classifications, with features such as Entity Dictionary, Rule-Based Matching, and relation Dictionary for automatic annotations. For this study, we have chosen manual annotations in order to achieve a high level of quality in the data set. Fig. 2 shows an image with an annotation example, in which the software used can also be seen.

For the construction of the second dataset, named Dataset Gemed Onco - Diagnostic Features (DGO-DF), we established a partnership with professionals specialized in annotation and health who supported the accomplishment of the labeling tasks. After the initial analysis, we decided to direct efforts toward creating a dataset focused on the diagnosis of Breast Cancer, aiming to provide valuable information in this crucial area of health. We covered entities that in this dataset include:

[1] https://ubiai.tools.



720_720_2022-04-08_23:54:42....
719_719_2022-04-08_23:54:42.2...
718_718_2022-04-08_23:54:41.9....
717_717_2022-04-08_23:54:41.8.t...
716_716_2022-04-08_23:54:41.7....
715_715_2022-04-08_23:54:41.5....
714_714_2022-04-08_23:54:41.4....

em tto psiquiatrico g43 com plaquetopenia EXAME RESULTADO 87 mil RESULTADO RESULTADO

ex fisico inalterado

hf ca de pancreas MEMBRO em 2 geracoes

tc EXAME LOCALIZAÇÃO LOCALIZAÇÃO de torax MEMBRO LOCALIZAÇÃO e abdome MEMBRO LOCALIZAÇÃO sed

Fig. 2. UBIAI software with an annotated document.

- *Característica do diagnóstico* (Diagnosis Features): Indicates attributes and relevant information about a specific case of cancer or oncological condition.
  Example: *Ca de cólon Ell* ***alto risco*** (Ell colon cancer **high risk**).
- *Órgão* (Organ): Note the organ affected by the cancer.
  Example: *Paciente notou nódulo em* ***mama*** *Esquerda* (Patient noticed lump in **breast** Left).
- *Local do Órgão* (Organ Location): Expresses the precise location of the affected organ.
  Example: *Identificou um nódulo em mama* ***Direita*** (Identified a breast lump **Right**).
- *Estadiamento* (Staging): Indicates the staging classification of the cancer.
  Example: *Ca de ovário* ***EIV*** *resistente a platina* (Platinum resistant ovarian cancer **EIV**).
- *Tempo* (Time): Informs the time elapsed since the detection or onset of symptoms.
  Example: *Paciente notou* ***há 6 meses*** *nódulo na mama E com crescimento moderado* (Patient noted **6 months ago** moderately growing breast E lump).

Due to the health professionals' specialty, only the entity annotations were developed in this dataset. For this, 600 documents of oncological evolutions were exported from an oncological clinic, already filtered with a diagnosis of Breast Cancer. All documents were anonymized to guarantee patient privacy and compliance with ethical guidelines.

We used the software MAE Annotation Tool[2] to perform the annotations. Multi-document Annotation Environment (MAE) is an annotation tool offering comprehensive features for marking text entities. MAE has advanced features, such as working with multiple documents simultaneously, making it easy to process large data sets.

The manually annotated Portuguese datasets allow the training of deep learning models specific to the Portuguese language, which is fundamental for NLP tasks. This manual annotation, carried out by specialists, guarantees the accuracy and quality of the labeled information, which considerably improves the performance and effectiveness of the trained models, increasing confidence in the results and allowing for faster advances in health.

To ensure the quality of the dataset, the set of 1022 annotated documents of the DGO-E underwent a review and validation process carried out by specialists in the field of oncology. During the review and validation process, experts agreed in consensus about the consistency and accuracy of annotations by verifying that entities and relations have been correctly identified and tagged in documents. This step is critical to ensure the quality and reliability of the annotated data set, providing valuable information for further analysis. The graph in Fig. 3 visually represents these results. The color bars represent the number of annotated entities for each category. The categories *Resultado* (result) and *Exame* (Exams) are the two numerous ones. Categories *Membro* (member), *Data* (date), and *Tempo* (time) are the less numerous ones, with a very small number of annotations for this last category. This distribution is coherent with the most frequent clinical notes observed in the use case material.

The dataset comprises many annotated entities emphasizing the categories *Exame* and *Resultado*, which have a significant amount of annotations (60,820 and 47720, respectively). These numbers indicate that the dataset covers various related information in the EHRs. In

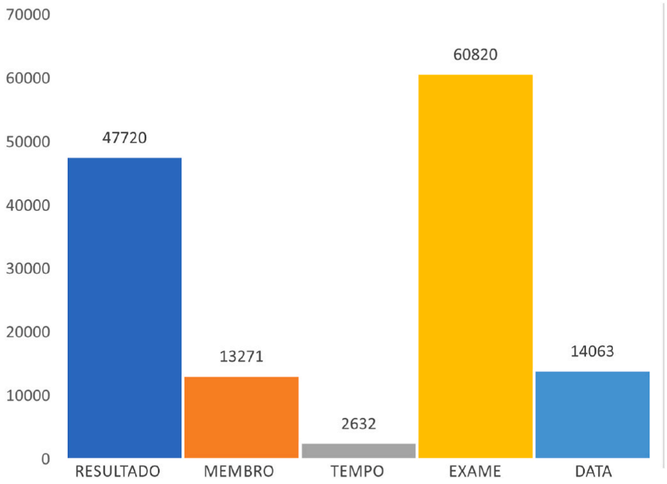


Fig. 3. DGO-E examination entities comparison.

[2] http://keighrim.github.io/mae-annotation/.

addition, *Membro* and *Data* entities had fewer annotations (13,271 and 14,063). These entities are relevant to the medical context and can provide important information about affected body parts and dates related to examinations. The small number of *Tempo* annotations (2,632) poses a challenge to the model generation, nevertheless, it represents the data distribution.

In the context of Relation Extraction (RE) oncological evolutions, the objective was to identify the relevant relations between the annotated entities, such as, for example, the relation between an *Exame* and its *Resultado*, or the relation between a *Data* and a specific *Exame*. The graph in Fig. 4 compares the relations.

As for the relations annotated in the dataset, we observed that the relation *Resultado* has a significant amount of annotations (87,386), while the relations *Quando* and *Localização* have smaller amounts (10,302 and 9,080, respectively). This information was essential to assist in interpreting and analyzing test results.

Data labeling for NER uses the IOB format widely. It provides a consistent and standardized structure for marking the boundaries of entities in a text. This format allows for accurate text analysis and prevents overlapping or missing words relating to an entity, ensuring that all relevant parts are correctly identified.

As for the DGO-DF, upon completing the annotation process, the 600 annotated documents underwent review and validation by oncology experts. This step is essential to ensure the quality and consistency of the notes, ensuring that the relevant information has been correctly identified and labeled in the documents. Finally, we exported the dataset in IOB format. A comparative graph of the entities is illustrated in Fig. 5.

This dataset has a considerable amount of annotated entities, emphasizing the *Estadiamento* entity with 1148 occurrences. This entity is significant in diagnosing breast cancer, as it provides valuable information about the stage of the disease and its progression. In addition, other relevant entities were noted, such as *Caracteristica do diagnóstico* with 1096 occurrences, and *Órgão* with 855 occurrences. These entities provide information about the specific characteristics of the disease and the affected organ.

On the other hand, the *Local do órgão* entity had fewer occurrences, with 342 records. This amount may suggest that specific information about the location of the organ affected by breast cancer may be less emphasized in annotated medical texts. However, it is necessary to consider that the importance of this entity may vary depending on the specific medical context.

The *Tempo* entity registered 116 occurrences, providing information about the time related to breast cancer diagnosis. Although it is a smaller amount compared to other entities, time is a relevant factor in the analysis of breast cancer. It can provide vital information for the treatment and prognosis of patients.

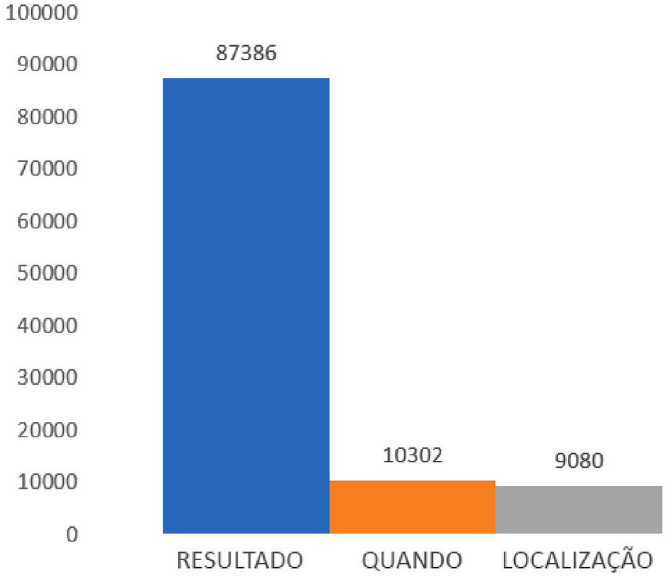


Fig. 4. Comparison of DGO-E exam lists.

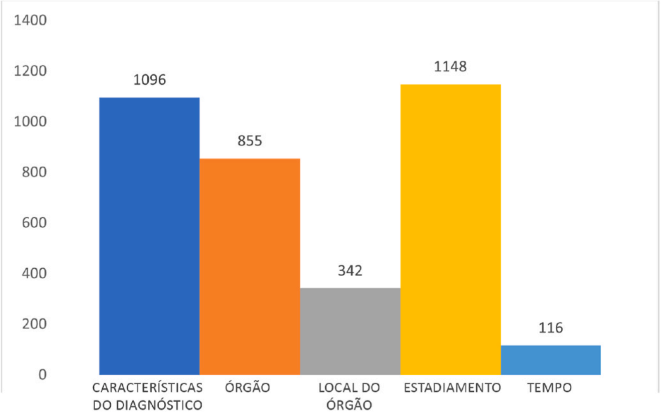


Fig. 5. Comparison of DGO-DF entities.

5. Experiments

This section describes the experiments and their contributions in two main aspects. The first one is regarding the NER and RE models developed, and the second is considering the construction and the use of the ontology for information integration.

5.1. Named entity recognition

In the first experiment, we trained the BERT model to perform the NER and extract features representing the text. Our NER architecture was developed using the BertForTokenClassification model,[3] proposed in Ref. [27]. This model has extensive library support, including a Python tokenizer, besides being compatible with Jax, PyTorch, and TensorFlow.

We experimented with different models in the experiment. The proposal was to use the entity recognition experiment's best result to perform the relations classification with BiLSTM later. The BERT models used for testing this experiment are: BERT-base-multilingual-cased [28], BioBERT-PT [28,29] and BERTimbau [30,31].

Building the model for NER was initially performed using the Spacy[4] library, implemented in Python. Spacy is an NLP library that offers advanced features for parsing and extracting information from text. We used this library to prepare and process data during training. Two distinct pieces of training were carried out, each with 50 epochs, allowing the model to learn to identify the entities present in the data gradually. The model choice was the BERT model, an architecture based on Transformers. We conducted the training in a supervised learning environment, where the training data have texts labeled with their respective classes.

In the first training, we used 716 annotated samples, corresponding to 70 % of the dataset available for NER. We read the data from JSON files that contained the texts and their respective classes and were organized in Pandas data tables (a two-dimensional data structure with tabularly aligned data in rows and columns) to facilitate the processing and manipulation.

Before training, we performed a data cleaning and treatment step to remove possible outliers or corrupted data. However, we did not find any problem with the data used. In addition, tokenization was applied to

[3] www.huggingface.co/transformers.

[4] https://spacy.io.

the texts, dividing them into smaller units, such as words or subwords. Then, the texts were pre-processed using BERT's tokenizer, converting them into sequences of tokens and adding unique tokens to mark the beginning and end of the sequences. The strings of tokens have been adjusted to have the same specified maximum length. The model was loaded from a pre-trained model using the Transformers library.

Due to the BERT input size limitation, the algorithm was modified to go through the text in blocks of 128 tokens, considering the classification of each token at the moment when it was closest to the center. In short, the maximum sequence size was set to 512 with a step of 128. On a token 400, for example, in the first iteration, it was at a distance of 144 tokens from the center of the read (which ranges from 0 to 511). The second iteration contained 16 tokens from the center (the text advanced 128 tokens). In the third iteration, the distance was 112 tokens; in the fourth iteration, it was 240 tokens; and in the fifth iteration, the token 400 was no longer part of the text read by BERT. Therefore, for token 400, the output given in the second run was considered.

The training was performed in 50 epochs, with a batch size of 16. The optimization was done using the AdamW algorithm [32] with parameters of $\beta 1 = 0.9$ and $\beta 2 = 0.999$. A weight decay of 0.1 was applied, and the initial learning rate was set to 1e-5, with a linear decay and a warmup period in the first 100 steps.

During training, we applied the fine-tuning process. We trained the model with the dataset DGO-E and its examples labeled for the specific task. In this case, the final BERT classification layer was replaced by a new layer suitable for the task, such as an entity classification layer. Fine-tuning involves training the model with the annotated data, using optimization algorithms to adjust the model weights and minimize the loss function. Randomness was applied to the data, avoiding possible overfitting of the model. This approach helps ensure the model can generalize to new data and become independent of training examples. The entities were processed together, allowing the model to learn to identify named entities. Throughout the training process, several libraries were used, such as re, random, numpy, pandas, matplotlib, sklearn, TensorFlow, and Keras, which provided additional functionality and facilitated model development and training.

After training, we evaluated the model using the validation data described in section 6.1. The trained model can be used to perform inferences on new texts. Test texts were processed, and we obtained the predictions from the model. In addition to predictions, embeddings are extracted from texts using BERT's token [CLS] representation, capturing contextual information from the text.

To evaluate the model's performance, we conducted additional training using the DGO-DF dataset without any modifications. Given that the DGO-DF contains only annotated entities, the evaluation was focused exclusively on the NER architecture.

5.2. Relation extraction

For the Relations extraction experiment, we applied an approach with BiLSTM to identify pairs of entities to classify the existing relations between them. Classification is performed based on the types of entities in the text and the entities selected for the relation. BiLSTM runs several times separately for each pair, trio, or group of entities. We used unique tokens to indicate the start and end of entities, which the BERT model recognizes.

We implemented this approach with the PyTorch library from the API Hugging Face, adding a token-level classifier on top of BERT models. Data from the DGO-E, mainly related to the previous experiment, was used. Initially, we prepared the data for training. We transformed the data into an appropriate format, such as a Datatable. It was essential to generate a data structure to correctly represent the relations between the entities present in the text. Then, we coded the words of the text into numerical representations. We performed this coding using the Word embeddings Word2Vec and GloVe techniques. Numerical representations capture semantic information about words, allowing the model to learn relations between words.

We built the BILSTM architecture with layers of LSTM cells. The bidirectional structure implies a layer processing the sequence in the forward direction and another in the reverse direction. With this, the model captured contextual information concerning each word in the sequence, improving the overall understanding of the context. During training, a loss function was defined to measure the discrepancy between the relations extracted by BILSTM and the real relations present in the data.

We applied a random initialization to the BILSTM weights, with values close to zero, as suggested by Xavier initialization [33]. This approach helps to prevent gradient saturation or burst problems during training. BILSTM weights are adjusted using the backpropagation [34] algorithm, which propagates the loss function gradients through the network. We conducted the training in several epochs, completing iterations through the training data. For each epoch, the network weights are updated based on the calculated gradients, allowing the network to learn the relations between entities more accurately.

To guide the process of updating the network weights, the Adam optimizer [32] was used. Adam has hyperparameters such as learning rate, learning rate decay, momentum ($\beta 1$), and momentum decay ($\beta 2$). We updated the learning rate to control how quickly weights are gained during training. Learning decay allows for reducing the learning rate as training progresses gradually. Momentum controls the contribution of the first-order moment in updating the weights, while the decay of the moment allows gradually reducing this contribution. Therefore, to optimize BiLSTM, the following parameters were adopted: $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\epsilon = 1\text{e-}7$. The initial learning rate was set to 1e-3, with a decay of 1e-5. We performed experiments to determine the values of S = 128 and P = 64, which showed superior performance for the model. The BiLSTM base architecture was tested with these parameters while the other hyperparameters were adjusted. BiLSTM training was carried out for 75 epochs.

For BILSTM regularization, the L1 and L2 regularization techniques were applied, which add terms to the loss function to penalize large weights, together with the Dropout technique, which randomly turns off a set of units during training, reducing the coadaptation between them. We used these regularization techniques to avoid overfitting and improve the model's generalizability. During training, the data were divided into batches (or mini-batches) aiming at computational efficiency. This division allows performing calculations in parallel and optimizing available resources.

5.3. Information integration with ontology

We based the ontology construction on the methodology described in Ref. [35]. Initially, 3309 documents containing medical consultations and health professionals' notes were studied to identify the relevant domain terms. These documents totaled 212,829 words on 28,178 lines. From the study of these contents, we identified sets of information that can be obtained automatically by extracting information. This supplementary information is often needed by healthcare professionals in their daily work. Examples of these situations are observed in sentences describing the perception of symptoms or medication use.

A procedure involving contact with specialist professionals in health and technology was carried out, obtaining a set of meaningful phrases and concepts that served as a starting point for this work. We conducted interviews to identify a critical set of relevant issues and provide the basis for modeling the domain knowledge. We performed manual textual analysis to relate this information with the main concepts and relations defined in the knowledge base. This process was adopted to ensure that the knowledge base was accurate and reliable, serving as a starting point for future initiatives to extract textual information automatically.

The ontology was developed using the Protégé[5] software and consists of 181 classes, 14 data, and 12 object properties. It can describe classes such as diagnosis, course, test, drug, patient, procedure, protocol, symptom, and organ relevant to the oncology domain. In addition, an ontology defines relations between different entities, such as the relation between patients and symptoms, patients and drugs, drugs and measurement units, exams or procedures, locations in the body, patients and diagnoses, patients and exams. Data properties are used to represent specific data characteristics, such as exam identifiers, exam characteristics, treatment cycles, event dates, cancer staging, exam types, symptom intensity, body locations, patient identifiers, periods, medication dosage, and test results.

The ontology provides an organized structure to represent and relate data relevant to oncology, allowing a better understanding and analysis of clinical data. To exemplify, in Fig. 6, the central concepts of *Edema* are highlighted with their main relations, with examples of some instances of the ontology created. Further on, Fig. 9 in section 6.2 illustrates an example of an individual in the ontology.

The ontology's formally structured data facilitates the integration of unstructured clinical data, helping health professionals in the interpretation, and analysis of cancer patient's context. With well-defined relations, the ontology can serve as a basis for developing medical decision support systems, contributing to improvements in the treatment and monitoring of cancer patients.

Integration with the ontology was performed using libraries developed in this work. These tools assist in registering and manipulating data in the ontology, allowing the extracted entities and relations to be properly registered. Automatic insertion tests were performed.

The model uses the experiment with the DGO-E to extract the entities and relations of the exams, data that, after the extraction process, are available for registration in the ontology. By default, an ontology instance is created with the evolution information, such as the patient's GEMED ID (not extracted in this experiment due to anonymization), the evolution date, and the exams extracted through the "Present" relation. The extracted exams and their relations are placed in a new instance. By default, instance names are comprised of the exam or evolution name, the date, time, and patient ID in GEMED (ignored for now due to anonymization). It was decided to test the integration with the DGO-DF in future work since it does not have annotated relations.

The automatic ontology integration with the model was developed in Python 3 using the Owlready2[6] library. Owlready2 is a library that enables ontology-oriented programming in Python, providing transparent access to OWL ontologies. It includes an optimized quad store (a type of database optimized for storing and querying data in RDF format (Resource Description tablework), suitable for storing and querying large amounts of data) based on the SQLite3 (an open-source embedded relational database library), which offers good performance and efficient memory consumption, allowing you to handle large ontologies. In addition, Owlready2 has integration with UMLS (Unified Medical Language System) and medical terminologies through the submodule PyMedTermino2.

The algorithm is also integrated with the HermiT [36] reasoner, a logical reasoner designed to process OWL ontologies and make inferences about their concepts and relations. It supports automated reasoning, applying description logic to represent knowledge and perform inferences in an ontology. It follows the OWL ontology description language specifications and is capable of performing tasks such as instance classification, property inference, inconsistency detection, and equivalence inference.

Thus, integrating the model with the ontology enables ontological knowledge to enrich the models' analyses and inferences. In addition, using the structured ontology and the automated reasoning capabilities offered by HermiT helps the semantic interpretation of the knowledge represented in the ontologies. It facilitates the manipulation of the data generated by the model in a coherent and structured way.

After obtaining the model outputs, the OWL ontology is loaded into the algorithm. Next, the outputs are mapped to the concepts and properties defined in the ontology. This involves creating instances and establishing relations between them. Through Owlready2, it is possible to access and consult the ontology's concepts, properties, and relations, allowing the model data to be inserted in the structured format defined by the ontology.

6. Results and discussion

Regarding evaluating the results, two main fronts were considered: computing and health. The NER and RE experiments, the construction and integration with the ontology were evaluated in computational terms. At the same time, usability analysis and comparison of results were carried out in collaboration with health professionals.

6.1. Evaluation of NER and RE experiments

In this section, we delve into the evaluation of the performance of NER and RE models in the context of medical text analysis. Specifically, we focus on the results obtained from two datasets: DGO-E and DGO-DF. The main metrics considered for assessment include Accuracy, Recall, and F1 Score, along with Support, Macro Mean, and Weighted Mean.

To evaluate the DGO-E NER results, 305 annotated evolutions (30 % of the dataset) were used to test the model. The results of Accuracy, Recall, F1 Score, Support, Accuracy, Macro Mean, and Weighted Mean are illustrated in Table 1. In this table, we can observe that the entities *Exame*, *Órgão*, *Resultado*, *Data* and *Tempo* were recognized with accuracies ranging between 70.01 % and 78.92 %. The recall of these entities ranges from 70.38 % to 77.27 %, while the F1-score ranges from 70.59 % to 78.77 %. Support, which represents the number of instances of each entity in the dataset, ranges from 3311 to 14,648. The overall performance of the model, measured by accuracy, is 78.24 %. These values indicate that the model performed as expected in detecting and classifying these entities.

During training, an evaluation of the BILSTM was performed on a validation set, which consists of data not used during training (30 % of DGO-E). The BiLSTM output is compared with the actual labels of the relations present in the training examples. To measure the discrepancy between the model predictions and the true labels, the cross-entropy loss function was used. The results obtained are presented in Table 2.

In Table 2, the results of the recognition of relations in the DGO-E dataset are presented. The *Resultado*, *Quando* and *Localização* relations reached accuracies from 73.45 % to 76.75 %, recalls from 73.55 % to 75.87 % and F1-scores from 74.12 % to 75.28 %. Ratio support ranges from 2100 to 20,983. The overall accuracy of the model is 76.17 %. These results indicate that the model performed accordingly to the literature in detecting and classifying relations in this dataset.

To evaluate the result, the model was trained without any modification with the DGO-DF. Table 3 presents the NER results of the DGO-DF.

In Table 3, the results of NERs in the DGO-DF dataset are presented. The entities *Característica do Diagnóstico*, *Orgão*, *Local do órgão*, *Estadiamento* and *Tempo* achieved accuracies from 70.12 % to 73.71 %, recalls from 70.17 % to 72.74 % and F1-scores from 70.99 % to 72.98 %. Entity support ranges from 24 to 230. The overall accuracy of the model is 72.87 %. These results indicate that the model performed according to the obtained results in the literature in detecting and classifying these entities in the DGO-DF dataset.

Overall, the results obtained are satisfactory. When analyzing the metrics, it can be observed that the categories present similar performance, with values of Accuracy, recall and F1-score close. Most

[5] https://protege.stanford.edu/.

[6] https://owlready2.readthedocs.io/en/latest.

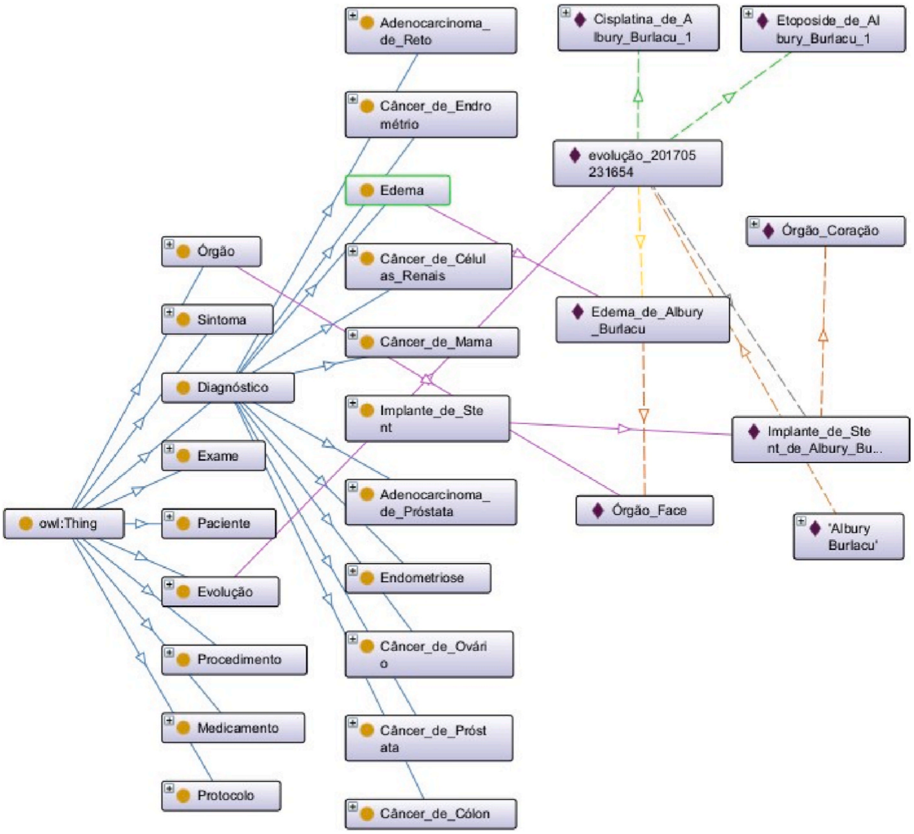


Fig. 6. Core Concepts of *Edema*. The concept *Edema* is related as a kind of *Diagnóstico*, linked to the organ *Órgão Face*. The evolution details regarding this kind of concept are expressed as ontology instances.

Table 1
NER performance of the DGO-E.

	Accuracy	Recall	F1-score	Support
Exame	78,92 %	77,27 %	78,77 %	14648
Membro	78,26 %	77,74 %	78,25 %	3311
Resultado	70,01 %	70,38 %	70,59 %	11504
Data	77,76 %	75,14 %	76,33 %	3501
Tempo	73,33 %	72,17 %	73,56 %	592
Accuracy			78,24 %	35994
Macro avg	72,78 %	76,01 %	75,10 %	35994
Weighted avg	78,32 %	78,76 %	78,66 %	35994

Table 3
DGO-DF NER performance.

	Accuracy	Recall	F1-score	Support
Característica do Diagnóstico	70,22 %	70,37 %	71,24 %	230
Órgão	73,71 %	72,74 %	72,98 %	189
Local do Órgão	70,12 %	70,17 %	70,99 %	71
Estadiamento	71,73 %	70,13 %	71,00 %	246
Tempo	71,70 %	71,19 %	71,46 %	24
Accuracy			72,87 %	760
Macro avg	72,78 %	72,00 %	72,10 %	760
Weighted avg	72,33 %	72,17 %	72,33 %	760

Table 2
DGO-E relation recognition performance.

	Accuracy	Recall	F1-score	Support
Resultado	76,61 %	75,87 %	75,24 %	20983
Quando	73,45 %	73,55 %	74,12 %	2100
Localização	76,75 %	75,01 %	75,28 %	2501
Accuracy			76,17 %	25584
Macro avg	76,78 %	76,00 %	75,10 %	25584
Weighted avg	76,33 %	78,17 %	78,22 %	25584

categories reached values above 70 % in these metrics. This indicates that the model is able to correctly identify entities and relations consistently.

In general terms, the BERT model demonstrated a good performance in the NER and RE in the two evaluated domains. It is important to emphasize that these results must be interpreted considering the specific characteristics of the data sets used. In the context of the medical domain, the RNE and the RE can be challenging due to the complexity and diversity of medical texts. Therefore, the results obtained can be considered satisfactory, taking into account the difficulty of the task.

It is important to consider that these results are specific to the datasets and tasks in question. They are not directly comparable with other experiments in the field due to the particular characteristics of the DGO-E and DGO-DF datasets, which are specific to the healthcare domain and contain unique company and patient data. However, even though a direct comparison is not feasible, the obtained results are consistent with the state of the art, as demonstrated by Ref. [37], who conducted NER on psychiatry records in Portuguese and achieved an accuracy of 64 %. Furthermore [38], conducted research on RE in clinical progress notes with pharmacogenomic data in Portuguese, obtaining an accuracy of 70 %.

Unlike the models in the literature, which are usually pre-trained on large general datasets, such as Wikipedia or fictitious data, the use of proprietary datasets, annotated by health experts, allows for the provision of domain-specific context. This represents a significant contribution to research. Although the results are not directly comparable to previous studies due to the reasons mentioned, they contribute to the advancement of the area and can serve as a reference in future comparative studies. Furthermore, it is important to consider that the results can be influenced by the amount of data available for training. If there is more data annotated, it is possible that the performance of the model will improve, as it will have more information to learn and generalize.

6.2. *Ontology integration*

In this section, we present the evaluation steps performed for the DGO-E experiment with the ontology. The study was conducted in a scenario study with healthcare professionals, including biomedical practitioners, specialist healthcare nurses, and medical students. After adequate preparation, two separate assessments were performed. The first evaluation aimed to calculate the level of accuracy of the model in relation to the professionals' judgment. The second evaluation aimed to understand the acceptance of specialists in relation to the integration of unstructured data in an ontology, and the details of this evaluation are presented in section 6.2.1.

For the integration evaluation, 22 evolutions were randomly exported from the SGO. A healthcare professional was in charge of evaluating the texts and manually marking the entities present in each of the documents. Later, these same evolutions were inserted into the model, allowing the comparison of the level of success of the model in relation to the expert.

Fig. 7 illustrates a comparative example of evolution 12, where the results of the marking carried out by the health professional and the entities recognized by the model are presented. On the left side of the figure, there are the markings made by the health professional, while on the right side, the entities identified by the model are indicated. The entities are identified by means of colored stripes: red represents the *exames*, yellow the *resultados*, blue the *membros*, pink the *datas* and green the *tempos*.

As shown in Fig. 7, the model presented the performance in identifying the entities marked by the health professional, with notable successes in the categories of *Membro*, *Tempo* and *Data*. However, the model failed to correctly identify the *US* (Ultrassom - Ultrasound) exams (*exame*), as well as the result (*resultado*) "nodule in the left breast measuring 6.3 cm". Table 4 presents the total comparative result between the markings made by the health professional and those made by the model.

The model showed a general average of 73.52 % correct answers, indicating a reasonable performance in relation to the markings made by the health professional. Specific evolutions were also examined, revealing cases in which the model obtained discrepant results in comparison with the markings made by health professionals. When analyzing the specific categories, it is observed that the model had the lowest accuracy in marking *Resultado* (33.33 %), indicating difficulties in recognizing and extracting detailed information from exams, which are generally complex, expressed in specific technical terms and abbreviations. For example, in evolution 9, the model was correct only 8 of the 28 *Resultado* markings made by the health professional, indicating a particular difficulty in this context. The results identified by the clinician often stem from a long sentence of text without certain patterns. This is a point that can be further explored in future work.

On the other hand, it obtained the highest accuracy in marking *Tempo* (100 %), but only 1 sample of this category was marked. The *Membro* and *Data* tags showed an accuracy closer to the general average (84.28 % and 94.33 %, respectively), indicating greater consistency of the model in these categories. In the *Exame*, many acronyms were used, but even so, the model had an average performance, presenting 55.68 % of correct answers. An evolution in which the model achieved 100 % accuracy in all categories is evolution 22. However, this result can be attributed to the fact that this evolution has only 2 marks in each category, making it a simpler case for the coping model. Finally, Fig. 8 illustrates a comparison by category.

Another important aspect raised in the analysis is the relation between the complexity of the information present in the evolutions and the performance of the model. Evolutions written in plain text format showed better results, while those extracted from tables or with data without continuity had lower performance presents an example of this problem), which indicates possible challenges for the model in question. relation to the structure and continuity of the data.

After the model's accuracy analysis, the evolutions were integrated into the ontology. This process is exemplified below with a specific example.

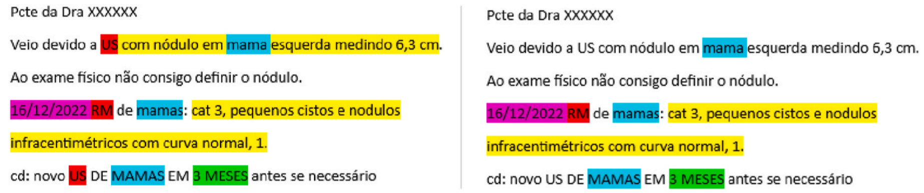
After processing these data by the model, the entities are extracted. As an example, the following category entities and values are recognized and extracted by the model: *data(26/12/2018)*, *data(01/07/2019*), *exame(cea)*, *resultado(252)*, *exame(vid d)*, *resultado(249)*. Then, the relations are identified and inserted to the ontology.

The experiment creates an evolution individual in the ontology, classified as *Evolução*, which includes the *CEA* (identified as *Antígeno Carcinoembrionário* - Carcinoembryonic Antigen') and *VIT D* (*Vitamina D* - Vitamin D) exams, along with the evolution date. In addition, two separate individuals were created for each exam extracted from the text, as well as for the respective results. Fig. 9 presents the evolution relations extracted in graph format.

The ontology allows a more precise and semantic structuring of the extracted data, in addition to enabling the creation of relations between entities, enriching the analysis, and allowing advanced queries. With the integration of entities and relations from the medical texts in the ontology, it was possible to obtain a more complete and structured view of the unstructured data, facilitating the retrieval and understanding of this information.

A preliminary proposal for implementing a prototype, based on alignment meetings with Interprocess, consists of automating the process of structuring the information extracted from the exams written by the physicians during the patient's evolution to fill in the SGO automatically. Currently, this process is performed manually by health professionals. This module is part of the patient's EHR, containing main information, evolutions, prescriptions, schedules, referrals, documents, activities, and personal data. In the exam module, the SGO presents a list of the exams performed by the patient, their results, the laboratory, and the exam date.

The proposal uses an information extraction model to analyze the registered medical exams and structure the relevant information based on these data. In this way, the exams were presented in an organized list. This will eliminate the need for manual completion and provide a more transparent and accessible view for healthcare professionals. This approach aims to optimize the workflow, save time for healthcare



Pcte da Dra XXXXXX

Veio devido a US com nódulo em mama esquerda medindo 6,3 cm.

Ao exame físico não consigo definir o nódulo.

16/12/2022 RM de mamas: cat 3, pequenos cistos e nodulos

infracentimétricos com curva normal, 1.

cd: novo US DE MAMAS EM 3 MESES antes se necessário

Pcte da Dra XXXXXX

Veio devido a US com nódulo em mama esquerda medindo 6,3 cm.

Ao exame físico não consigo definir o nódulo.

16/12/2022 RM de mamas: cat 3, pequenos cistos e nodulos

infracentimétricos com curva normal, 1.

cd: novo US DE MAMAS EM 3 MESES antes se necessário

Fig. 7. Comparison of evolution 12.

Table 4
Model accuracy comparison. The first column indicates the number of each document. In the columns *Exame*, *Resultado*, *Tempo*, *Membro* and *Data*, the values on the left column represent the number of appointments made by the health professional (HP), while the values on the right column indicate the number of markings made by the model (Mod). Additionally, an average accuracy of the markings performed on the document indicated in each line is presented.

	Exame		Resultado		Tempo		Membro		Data		Accuracy
	HP	Mod	HP	Mod	HP	Mod	HP	Mod	HP	Mod	
1	5	3	15	4	0	0	0	0	18	18	65,79 %
2	30	21	29	9	0	0	12	12	14	14	65,88 %
3	15	9	13	4	0	0	5	3	8	8	58,54 %
4	2	2	2	0	0	0	1	1	1	1	66,67 %
5	6	3	6	4	0	0	2	2	6	5	70,00 %
6	5	2	5	2	0	0	0	0	8	7	61,11 %
7	2	0	2	1	0	0	3	1	2	2	44,44 %
8	2	0	6	3	0	0	2	2	14	14	79,17 %
9	28	17	28	8	0	0	12	11	16	16	61,90 %
10	2	1	2	0	0	0	1	0	7	6	58,33 %
11	4	2	3	1	0	0	1	1	5	2	46,15 %
12	3	1	2	1	1	1	3	3	1	1	70,00 %
13	2	1	6	3	0	0	2	1	3	2	53,85 %
14	3	0	3	1	0	0	1	1	10	9	64,71 %
15	21	9	20	6	0	0	7	7	15	15	58,73 %
16	2	2	3	1	0	0	1	0	8	8	78,57 %
17	5	2	5	3	0	0	3	3	3	3	68,75 %
18	5	3	5	1	0	0	5	4	4	4	63,16 %
19	2	1	2	0	0	0	1	1	1	1	50,00 %
20	5	2	5	1	0	0	0	0	7	6	52,94 %
21	16	10	16	5	0	0	7	5	6	6	57,78 %
22	2	2	2	2	0	0	1	1	2	2	100,00 %
Average		55,68 %		33,33 %		100 %		84,28 %		94,33 %	**73,52 %**

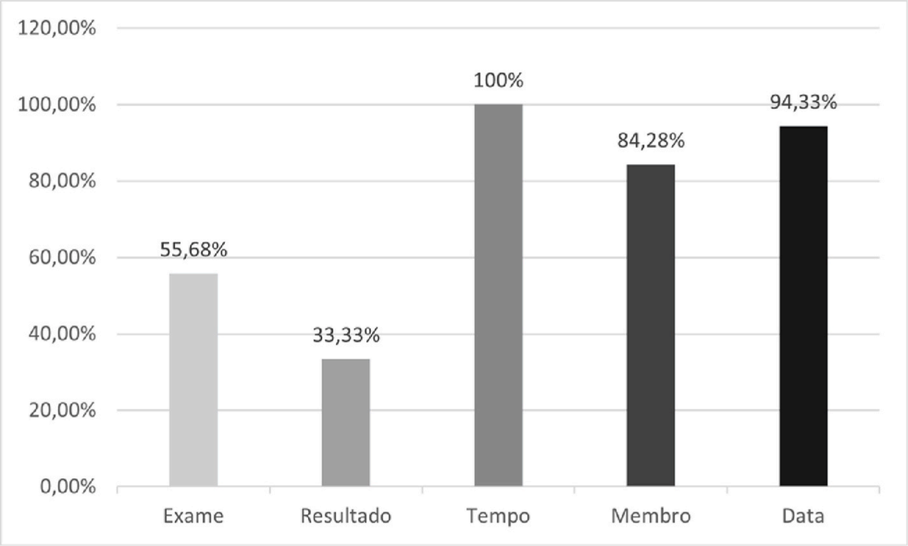


Fig. 8. Comparison of accuracy by category.

professionals, and reduce possible filling errors.

It is important to note that the prototype integrated with the SGO was a first version, subject to adjustments and refinements based on feedback and needs identified during its use. The partnership with Interprocess will allow the implementation to be carried out collaboratively. With the adoption of this prototype, we expect that the SGO will become even more complete and adequate, helping to improve patient care and the management of clinical information in a safe and integrated way.

6.2.1. Assessment of acceptance by professionals

The ontology was evaluated based on a scenario model. When describing the scenario model, an evaluator systematically describes the user's interaction scenarios with a system and evaluates the actions required to complete each scenario with a cognitive model [39]. Due to the objectives described in terms of scenarios and taking into account the answers obtained, the evaluator can effectively identify critical issues that may harm the achievement of objectives and identify aspects of acceptance of the use of resources in the system.

The five specialists in the health area mentioned above answered the questionnaire. Before the evaluations, these professionals received a detailed presentation about the developed system, an explanation of ontologies, and examples of model execution and results. The acceptance assessment sought to evaluate the experts' acceptance of integrating unstructured data into an ontology. For this, they answered a questionnaire designed to gather the necessary evidence. The questions were the following: I know of some software that performs this type of task; The use of the model can facilitate the formalization of knowledge in the health area; The use of the model can enable the development of new applications in the health area; The proposed model is helpful for

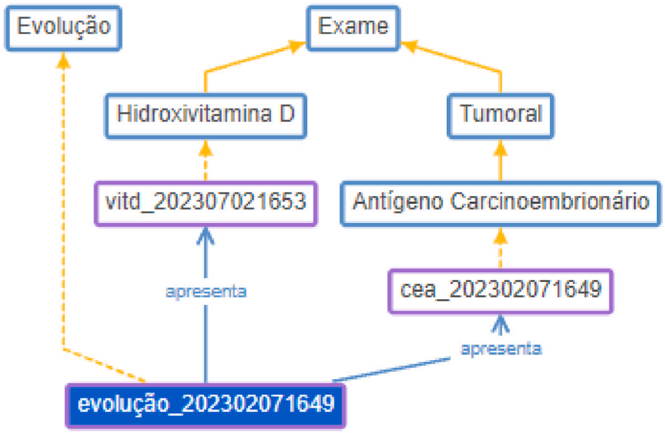


Fig. 9. Flow of extracted evolution relations.

the health area; The use of ontology as a basis for structuring data can simplify the understanding and visualization of health information; If you need structured information, you will use the template; It is easy to understand the structuring of the data with the model.

It is possible to state that 60 % of the participants disagree entirely when questioned about their knowledge of any software capable of performing the proposed task, which indicates that most do not know the software that performs the task. This indicates a need for more knowledge and adoption in this area, which may represent a research opportunity. The remaining 40 % showed agreement, suggesting a small group of participants with some prior knowledge about this type of technology.

As for the possibility of the model facilitating the formalization of knowledge in the health area, 100 % of the participants agree entirely. This indicates that there is significant recognition among participants that using the model can contribute to the formalization of knowledge in this specific health field. This positive perception suggests a potential interest and acceptance of the model. Regarding the ability of the model to enable the development of new applications in the health area, 100 % of the participants agree entirely. This high agreement reflects a strong consensus that the proposed model can be a viable tool for the development of new applications in the health area.

In summary, the results show a positive recognition of the potential of the proposed model, mainly concerning ease of understanding, usefulness in the health area, feasibility for the development of applications, and simplification of data structuring. These insights can assist in deploying the model in the SGO, highlighting relevant discussion points.

7. Conclusion

This article presented a new method for integrating natural language data in a domain ontology. The study case involves Deep Learning techniques, with Transformers models like BERT, and the Bi-LSTM model, for implementing tasks of named entity recognition and relation extraction.

The methodology was evaluated by specialists in the health field and at the computational level. Experiments were carried out with NER and RE techniques in real medical texts in Portuguese. Our work provided a new annotated dataset in Brazilian Portuguese in the medical domain, which was used in the experiments. The developed model achieved results of 78.24 % accuracy in the exam domain and 72.87 % in the diagnosis domain in the NER and RE. In addition, an oncology-focused ontology was built and integrated into the model, encompassing approximately 181 classes, 14 data properties, 12 object properties, and more than 200 individuals. The evaluation with specialists in the health area obtained a success rate of 73.52 % concerning their analysis, and the usability research showed excellent acceptance.

The methodology developed in this work provides an efficient approach for unstructured NER and RE in the health area. The promising results pave the way for future applications that may benefit the medical field, providing a structured and reliable knowledge base to help health professionals and researchers. Through the results, one can observe the potential and importance of this work for the health area, as well as identify a range of possibilities to be worked on for an effective extraction of relations and aspects to be evaluated on how to structure these data received in natural language.

This study serves as a foundation for the development of applications that support various analysis and operational tasks by using natural language processing to extract diverse information from unstructured text and transform it into structured data with relational attributes. While the focus is on Portuguese text, the study provides insights that could be valuable for processing medical text in languages other than English.

7.1. Future work

In future work, we intend to expand the DGO-E with new annotations of entities and relations for the field of oncology. With the increase in available data, improving the model's performance would be possible. Another possibility is to augment the DGO-DF with the annotation of entity relations. After expanding the datasets, it would be possible to perform new training on the BiLSTM architecture using the annotated data. This approach could improve information extraction and understanding of the relations of EHR evolutions.

Exploring the ontology developed in this study more comprehensively is considered necessary. Rather than just focusing on specific entities and relations, information could be extracted from the ontology to generate oncology-related questions. It would create a system capable of interacting with users, providing answers and relevant information based on ontology concepts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel - CAPES (Finance Code 001) and the National Council for Scientific and Technological Development – CNPq (Grant Numbers 159593/2019-0, 309537/2020-7 and 404572/2021-9) for supporting this work.

References

[1] Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20(5):e262–73.
[2] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inf Assoc 2016;23(5):1007–15.
[3] Solares JRA, Raimondi FED, Zhu Y, Rahimian F, Canoy D, Tran J, Gomes ACP, Payberah AH, Zottoli M, Nazarzadeh M, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J Biomed Inf 2020;101:103337.
[4] Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inf Assoc 2019;26(4):364–79.
[5] Dhole G, Uke N. Nlp based retrieval of medical information for diagnosis of human diseases. Int J Renew Energy Technol 2014;3(10):243e8.
[6] Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inf 2017;73:14–29.
[7] Liu S, Wang X, Hou Y, Li G, Wang H, Xu H, Xiang Y, Tang B. Multimodal data matters: language model pre-training over structured and unstructured electronic health records. IEEE J Biomed Health Inform 2022;27(1):504–14.

[8] Bitterman DS, Miller TA, Mak RH, Savova GK. Clinical natural language processing for radiation oncology: a review and practical primer. Int J Radiat Oncol Biol Phys 2021;110(3):641–55.

[9] Zhou S, Wang N, Wang L, Liu H, Zhang R. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inf Assoc 2022;29(7):1208–16.

[10] Qiao B, Zou Z, Huang Y, Fang K, Zhu X, Chen Y. A joint model for entity and relation extraction based on bert. Neural Comput Appl 2022:1–11.

[11] Kim Y-M, Lee T-H, Na S-O. Constructing novel datasets for intent detection and ner in a Korean healthcare advice system: guidelines and empirical results. Appl Intell 2023;53(1):941–61.

[12] Leng J, Wang D, Ma X, Yu P, Wei L, Chen W. Bi-level artificial intelligence model for risk classification of acute respiratory diseases based on Chinese clinical data. Appl Intell 2022;52(11):13114–31.

[13] Li Z, Chen H, Qi R, Lin H, Chen H. Docr-bert: document-level r-bert for chemical-induced disease relation extraction via Gaussian probability distribution. IEEE J Biomed Health Inform 2021;26(3):1341–52.

[14] Qasim R, Bangyal WH, Alqarni MA, Ali Almazroi A, et al. A fine-tuned bert-based transfer learning approach for text classification. J Healthc Eng 2022.

[15] Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, Wu X-C, Durbin EB, Doherty J, Stroup A, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021;25(9):3596–607.

[16] Sun T, Wang D. Research on relation extraction method based on multi-channel convolution and bilstm model. In: 2020 IEEE intl conf on parallel & distributed processing with applications, big data & cloud computing, sustainable computing & communications, social computing & networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE; 2020. p. 1082–90.

[17] Chen L, Zhang W, Ye H. Accurate workload prediction for edge data centers: Savitzky-golay filter, cnn and bilstm with attention mechanism. Appl Intell 2022; 52(11):13027–42.

[18] B. A. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Tech. Rep. EBSE 2007-001, Keele University and Durham University Joint Report (07 2007). URL https://www.elsevier.com/_data/promis_misc/525444systematicreviewsguide.pdf.

[19] Prisma. Preferred reporting items for systematic reviews and meta-analyses. disponível em: http://prisma-statement.org/PRISMAStatement/Checklist.aspx. [Accessed 8 April 2021].

[20] Cooper ID. What is a “mapping study?”. J Med Libr Assoc: JMLA 2016;104(1):76.

[21] Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. Behrt: transformer for electronic health records. Sci Rep 2020; 10(1):1–12.

[22] S. Morgan, T. Ranasinghe, M. Zampieri, Wlv-rit at germeval 2021: multitask learning with transformers to detect toxic, engaging, and fact-claiming comments, arXiv preprint arXiv:2108.00057..

[23] Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P. Fine-tuning bert for joint entity and relation extraction in Chinese medical text. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2019. p. 892–7.

[24] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.

[25] C. Herlihy, R. Rudinger, Mednli is not immune: natural language inference artifacts in the clinical domain, arXiv preprint arXiv:2106.01491..

[26] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342..

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805..

[28] J. V. A. de Souza, E. T. R. Schneider, J. O. Cezar, L. E. Silva, Y. B. Gumiel, E. C. Paraiso, D. Teodoro, C. M. C. M. Barra, et al., A multilabel approach to Portuguese clinical named entity recognition, J Health Inform 12..

[29] Ji Z, Wei Q, Xu H. Bert-based ranking for biomedical entity normalization. AMIA Summit Transl Sci Proc 2020;2020:269.

[30] Lopes É, Correa U, de Freitas LA. Exploring bert for aspect extraction in Portuguese language. In: The International FLAIRS Conference Proceedings. 34; 2021.

[31] Souza F, Nogueira R, Lotufo R. Bertimbau: pretrained bert models for brazilian Portuguese. In: Brazilian conference on intelligent systems. Springer; 2020. p. 403–17.

[32] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101..

[33] Xu G, Wang C, He X. Improving clinical named entity recognition with global neural attention. In: Web and big data: second international joint conference, APWeb-WAIM 2018. Macau, China: Springer; 2018. p. 264–79. July 23-25, 2018, Proceedings, Part II 2.

[34] Li S, Wang W, Lu B, Du X, Dong M, Zhang T, Bai Z. Long-term structural health monitoring for bridge based on back propagation neural network and long and short-term memory. Struct Health Monit 2023;22(4):2325–45.

[35] D. Man, Ontologies in computer science, DIDACTICA MATHEMATICA 31 (1)..

[36] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, Deeponto: a python package for ontology engineering with deep learning, arXiv preprint arXiv: 2307.03067..

[37] Niero LHP, Guilherme IR, Oliveira LE Se, de Araújo Filho GM. Psybertpt: a clinical entity recognition model for psychiatric narratives. In: 2023 IEEE 36th international symposium on computer-based medical systems (CBMS); 2023. p. 672–7. https://doi.org/10.1109/CBMS58004.2023.00298.

[38] Bettoni GN. Extração de informação em evoluções clínicas e integração com dados farmacogenômicos. Pontifícia Universidade Católica do Rio Grande do Sul; 2022. Master’s thesis.

[39] Sugimura V, Ishigaki VK. New web-usability evaluation method: scenario-based walkthrough. Fujitsu Sci Tech J 2005;41(1):105–14.