UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

MARCO ANTONIO SCHWERTNER

EXPLORING TEXT CLASSIFICATION METHODS IN ONCOLOGICAL MEDICAL NOTES USING MACHINE LEARNING AND DEEP LEARNING

São Leopoldo
2020

Marco Antonio Schwertner

EXPLORING TEXT CLASSIFICATION METHODS IN ONCOLOGICAL MEDICAL NOTES USING MACHINE LEARNING AND DEEP LEARNING

Dissertation presented as a partial requirement to obtain the Master's Degree from the Graduate Program in Applied Computation at the University of Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Dr. Sandro Rigo

São Leopoldo
2020

S415e Schwertner, Marco Antonio.
Exploring text classification methods in oncological medical notes using machine learning and deep learning / Marco Antonio Schwertner. – 2020.
80 f. : il.; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2020.
"Advisor: Prof. Dr. Sandro Rigo".

1. Inteligência artificial. 2. Aprendizado do computador. 3. Aprendizado profundo. 4. Oncologia. 5. Sistemas de Apoio a Decisões Clínicas. I. Título.

CDU 004.8

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecário: Flávio Nunes – CRB 10/1298)

(Esta folha serve somente para guardar o lugar da verdadeira folha de aprovação, que é obtida após a defesa do trabalho. Este item é obrigatório, exceto no caso de TCCs.)

To my wife, Claudia, that has always supported me and continues to motivate me through the tough moments.
To my parents Amelia and Clovis, who taught me that knowledge is the most significant heritage someone can leave to his or her children.

However bad life may seem, there is always something you can do and succeed at.
Where there's life, there's hope.
— STEPHEN HAWKING

ABSTRACT

With the preventive and personalized medicine advances, and technological improvements enabling better interaction from patients with their healthcare information, the volume of healthcare data gathered has increased. A relevant part of these data is recorded as an unstructured format in natural language free-text, making it harder for Clinical Decision Support Systems (CDSS) to process these data. Consequently, healthcare professionals get overwhelmed keeping themselves updated with the patient's healthcare information because they need more time to gather and analyze it manually. Furthermore, to define an oncology diagnosis and its treatment plan is a complex decision-making process because it is affected by a broad range of parameters.

This research's main objective is to apply several text classification methods in non-synthetic oncology clinical notes corpora to help with this decision-making process. First, the corpora were obtained from an Oncology EHR system from three different oncology clinics. Two corpora versions were created: the per-clinical-event version with each patient's medical note per record; and the per-patient version with one record per patient with his or her medical notes. Then, these corpora were preprocessed to leverage the performance of the classifiers. As the last step, several machine learning and one deep learning text classification methods were trained using these corpora with each patient's diagnosis as enriched data.

The following machine learning and deep learning classification methods were applied: Multilayer Perceptron (MLP) neural network, Logistic Regression, Decision Tree classifier, Random Forest classifier, K-nearest neighbors (KNN) classifier, and Long-Short Term Memory (LSTM). An additional experiment with an MLP classifier was performed to evaluate the preprocessing step's influence on the results, and it found that the classifier's mean accuracy was leveraged from 26.1% to 86.7% with the per-clinical-event corpus, and 93.9% with the per-patient corpus. The classifier that best performed was the MLP with 2 hidden layers (800 and 500 neurons), which achieved 93.90% accuracy, a Macro F1 score of 93.61%, and a Weighted F1 score of 93.99%. The experiments were performed in a dataset with 3,308 medical notes from a small oncology clinic.

Keywords: Artificial Intelligence. Deep Learning. Machine Learning. Healthcare. Oncology.

RESUMO

Com os avanços na medicina preventiva e personalizada, e as melhorias tecnológicas permitindo melhor interação do paciente com suas informações de saúde, o volume coletado de dados de saúde tem aumentado. Uma parte importante desses dados é armazenada em formato não estruturado em texto livre em linguagem natural, dificultando o processamento desses dados pelos Sistemas de Apoio à Decisão Clínica (SADC). Consequentemente, os profissionais de saúde ficam sobrecarregados tentando manter-se atualizados com as informações de saúde dos seus pacientes porque precisam de mais tempo para coletar e analisar esses dados manualmente. Definir um diagnóstico e tratamento oncológico é um processo de tomada de decisão complexo, pois é afetado por uma ampla gama de parâmetros.

Para ajudar neste processo de tomado de decisão, esta pesquisa possui como principal objetivo aplicar diversos métodos de classificação de textos em corpora com registros médicos não sintéticos, para aprender e sugerir o diagnóstico baseado no histórico clínico do paciente. Primeiro, os corpora foram obtidos de um S-RES (Sistema de Registro Eletrônico em Saúde) Oncológico de três diferentes clínicas de oncologia. Foram criadas duas versões dos corpora: a versão por-evento-clínico com um registro médico de paciente por registro; e a versão por-paciente com um registro por paciente com seus registros médicos. Então, os corpora foram pré-processados para alavancar o desempenho dos classificadores. Por fim, diversos métodos de classificação de texto de aprendizagem de máquina e aprendizagem profunda foram treinados utilizando os corpora junto com o diagnóstico de cada paciente como dados enriquecidos.

Diversos experimentos foram realizados, avaliando os seguintes métodos de classificação de textos de aprendizagem de máquina e de aprendizagem profunda: Multilayer Perceptron (MLP) neural network, Logistic Regression, Decision Tree classifier, Random Forest classifier, K-nearest neighbors (KNN) classifier, and Long-Short Term Memory (LSTM). Um experimento adicional com um classificador MLP foi realizado para avaliar a influência da etapa de pré-processamento nos resultados, e foi encontrado que a acurácia média do classificador foi alavancada de 26,1% para 86,7% com o uso do corpus por-evento-clínico, e 93,9% com o corpus por-paciente. O classificador com melhor desempenho foi o MLP com duas camadas ocultas (800 e 500 neurônios), que atingiu 93,90% de acurácia, um escore Macro F1 de 93,61%, e um escore Weighted F1 de 93,99%. Os experimentos foram realizados num conjunto de dados com 3.308 registros médicos de uma clínica de oncologia pequena.

Palavras-chave: Inteligência Artificial. Aprendizagem Profunda. Aprendizagem de Máquina. Saúde. Oncologia.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
BoW	Bag-of-Words
CDSS	Clinical Decision Support Systems
CNN	Convolutional Neural Network
ConvNet	Convolutional Neural Network
DB	Database
ER	Entity Relational
EHR	Electronic Health Records
ICD	International Classification of Diseases
KNN	K-nearest Neighbors Classifier
JSON	JavaScript Object Notation
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron Neural Network
MS	Microsoft
NLP	Natural Language Processing
PCA	Principal Component Analysis
PoS	Part-of-speech
SQL	Structured Query Language
VSM	Vector Space Model
XML	Extensible Markup Language

CONTENTS

1 INTRODUCTION

Current models of mass medicine will not be affordable beyond a few decades from now. A sustained model of large-scale healthcare provision will have to move away from the hospital being the center of routine care, replacing it with the person at the point where care is delivered (FERNANDES et al., 2009). Healthcare today is delivered the same way it was done a hundred years ago (MESKÓ, 2014). The healthcare system could be represented as a pyramid. The base features health insurers, governments, and pharmaceutical companies. In the middle are medical professionals, and patients in the smallest segment at the top of the pyramid. In this model, the medical professionals are bearing almost all the healthcare responsibility (MESKÓ, 2014).

As a new and revolutionary perspective in healthcare, precision medicine requires a personalized diagnosis and treatment plan. The world today is moving toward changing healthcare from reaction and hospital-centered to prevention and personalized treatment approaches (SABRA et al., 2018). To provide this type of care, physicians must be up to date with these new kinds of treatments and patient's health information. Furthermore, decision making is one of the most complex skills required for an oncologist. It is affected by a broad range of parameters, such as different information sources to define oncology diagnosis and a wide variety of treatment options with various outcomes (GLATZER et al., 2020). In this context, the information accumulated in a health institution's database can prove valuable to help physicians decide for the appropriate course of action in specific cases (REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015). Nevertheless, it is not feasible for the health professionals to evaluate and analyze these resources, due to the enormous effort it will demand, considering the size of these databases (ALEMZADEH; DEVARAKONDA, 2017), and the exponential growth in the amount of medical information (MESKÓ, 2014).

Advances in technology, combined with increased patient engagement in health care, have improved oncology care. These advancements include the use of information technology to support accessing health information, computerized order entry systems, electronic prescribing, and electronic health records (EHRs) (SCHULMEISTER, 2016), which generate a vast and heterogeneous amount of healthcare data.

Considering these challenges, beyond a healthcare system's restructuring, tools are needed to support healthcare professionals in their activities. Clinical Decision Support Systems (CDSS) provide additional assistance, synthesizing and integrating patient-specific information, performing complex evaluations, and presenting the results to clinicians in an adequate time (HUNT et al., 1998a), i.e., CDSS offers assistance to overcome the difficulties in dealing with this massive amount of information (BUCUR et al., 2013; POLPINIJ, 2011; WANG et al., 2018).

CDSS is defined as "any software designed to directly aid in clinical decision making in which characteristics of individual patients are matched to a computerized knowledge base to generate patient-specific assessments or recommendations that are then presented to clin-

icians for consideration" (HUNT et al., 1998b in REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015).

CDSS has three input types (REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015):

- Structured data as electronic health records;
- Semi-structured data as XML documents or two columns laboratory results;
- Unstructured data like narrative text, patients clinical observation, radiology reports, and operative notes.

Due to the rapid growth of electronic information, health information about patients is mostly found as unstructured data, i.e., narrative text or free text (REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015). The unstructured format of the enormous amount of medical data is the primary motive of recent research initiatives aiming to extract information from medical notes (KREIMEYER et al., 2017; CHEN et al., 2018; ZHANG et al., 2017; SHICKEL et al., 2017)).

Understanding the gap between CDSS and unstructured data, this work studies information extraction and text classification techniques. Furthermore, it focuses on different text classification methods and performs several experiments on non-synthetic oncological medical notes corpora. Finally, it evaluates the experiments' results and proposes future steps.

1.1 Research question and objectives

According to the introduction, a vast amount of unstructured data is generated in the healthcare area daily, and it is not feasible for healthcare professionals to keep up to date with the patient's clinical information.

The context above leads to the following research question: **"What are the text classification methods that better support the healthcare professional on diagnosis decisions, based on the patient's oncology clinical history?"**

According to related works, tasks such as text classification and information extraction are essential to support CDSS in dealing with structured information. This work focuses on the text classification task, covering the steps to create and process a clinical corpus for the experiments and also comparing evaluations with machine learning and deep learning approaches.

An oncology clinical notes corpus was created, preprocessed, and transformed to be used by the machine learning and deep learning methods. It is important to emphasize that this corpus was obtained from a real-world oncology clinic, de-identified to preserve the patient's and professional's identification, and is entirely in the Brazilian Portuguese language.

This work implemented several machine learning and deep learning methods on text classification, compared their performance, and evaluated their results.

Therefore, the main objective of this work consists of applying text classification approaches to support healthcare professional needs regarding diagnosis decisions.

In order to support this objective, the following specific objectives were defined:

- Produce of a de-identified corpus with non-synthetic medical notes from the oncology area, in the Brazilian Portuguese language;
- Perform experiments with text classification approaches;
- Compare the performance of machine learning and deep learning methods in text classification;
- Evaluate the application of text classification in an Oncology EHR system.

1.2 Methodology

This dissertation follows an exploratory research methodology. This work aims to provide more information about the problem that will be investigated, and afterward to build a hypothesis that will be confirmed through experiments.

This work can be characterized as an experimental work because it aims to explore different machine learning and deep learning methods on text classification tasks on oncological medical notes. It has the following steps:

- To perform a survey about text classification methods;
- To evaluate the related work;
- To obtain the data from the EHR system's database, de-identify them, and create the corpus;
- To preprocess, annotate and enrich the corpus;
- To develop experiments in text classification with machine learning and deep learning methods;
- To evaluate the experiments' performance and results;
- To document this research.

1.3 Text structure

This work is organized as follows. Chapter 2 talks about some background information about the main methods used in this work. In Chapter 3, related works are presented and discussed. Furthermore, the surveys performed to support the Related Work Chapter are described

in Appendixes A and B. The methodology and model of this work are detailed in Chapter 4, followed by the description of the performed experiments in Chapter 5. In the end, the conclusions are presented in Chapter 6, and this work's contributions are discussed.

2 BACKGROUND

This chapter reviews the essential concepts about clinical notes structure, data de-identification and pseudonymization, electronic health records, natural language processing, machine learning, and deep learning. It supports the work by providing insights into the objectives that text classification in medical notes should achieve and how machine learning and deep learning algorithms can build such systems.

2.1 Clinical Notes Structure

Clinical notes or medical records – whether electronic or handwritten – are essential for the continuity of patients' care. Adequate medical records enable us to reconstruct the essential parts of each patient contact without reference to memory. Therefore, they should be comprehensive enough to allow a health professional to carry on where a colleague left off (SOCIETY, 2018).

Writing clear and descriptive case notes is very different from most other types of writing. The objective in writing case notes is merely to create an accurate and informative record of treatment and patient progress (HODGES, 2011).

Medical records should summarize the key details of every patient contact. Clinical records should include:

- Relevant clinical findings;
- The decisions made and the actions agreed, and who is making the decisions and agreeing with the actions;
- The information that is given to patients;
- Any drugs prescribed or other investigation or treatment;
- Who is making the record and when.

On subsequent occasions, medical records should also note the patient's progress, findings on examination, monitoring and follow-up arrangements, details of telephone consultations, details about chaperones present, and any instance in which the patient has refused to be examined or comply with treatment.

Medical records can cover a wide range of material, including:

- Handwritten notes;
- Computerized records;
- Correspondence between health professionals;

- Laboratory reports;
- Imaging records, including x-rays;
- Photographs;
- Video and other recordings;
- Printouts from monitoring equipment;
- Text or email communication with patients.

When a health professional needs to present a patient to discuss at rounds, the team (including nurses, residents and attending physicians) will listen to the presentation to get an idea of what is going on with the patient. The information needs to be correct, well organized, and concise.

2.1.1 The SOAP note format

In order to facilitate a standard method for providing patient information, clinicians use the SOAP note format for both writing notes and presenting patients on rounds.

A SOAP note is information about the patient, which is written or presented in a specific order, including certain components. SOAP notes are used for admission notes, medical histories, and other documents in a patient's chart.

The purpose of a SOAP note is to have a standard format for organizing patient information. A SOAP note consists of four sections, including subjective, objective, assessment, and plan (HODGES, 2011):

- Subjective: refers to subjective observations that are verbally expressed by the patient, such as information about symptoms. This section should provide a narrative of the patient's feelings, concerns, problems expressed, goals, thoughts expressed, the intensity of problems expressed, and the problems' impact on significant relationships. The patient's perception of the issues should be clear to an outside reader of the record;
- Objective: this section is made up of what the physician can "see, hear, smell, count, or measure." It includes vital signs, such as pulse, respiration, and temperature. Information from a physical exam including color and any deformities felt should also be included. Results of diagnostic tests, such as laboratory work and x-rays, can also be reported in the SOAP notes' objective section;
- Assessment: is a summary of the physician's clinical impressions regarding the patient's problem or problems, i.e., the diagnosis or condition the patient has. It serves to synthesize and analyze the information expressed in the subjective and objective sections. In

some instances, there may be one precise diagnosis. In other cases, a patient may have several health issues. There may also be other times where a definitive diagnosis is not yet made, and more than one possible diagnosis is included in the assessment;

- Plan: the final section of the SOAP format is the plan. The plan is built on the information within the subjective, objective, and assessment sections. This section generally consists of two parts: the action plan and the treatment prognosis. It may involve ordering additional tests to rule out or confirm a diagnosis, the treatment that is prescribed, such as medication or surgery. The plan may also include information for self-care, including bed rest and days off work.

2.2 Data de-identification and pseudonymization

Frequently researchers need to work with data extracted from databases to work with the patient's clinical history. However, it is necessary to apply aggregation techniques without the patients' personal data disclosure to keep their privacy. This can be achieved with the use of pseudonymization (HEALTH INFORMATICS - PSEUDONYMIZATION - ISO 25237, 2020).

Anonymization is the process that removes the association between the identifying data set and the data subject. Pseudonymization is a particular type of anonymization that removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.

Pseudonymization is a de-identification type that removes the association of the data subject to its personal data, replacing it by an association with one or more pseudonyms. It is used when a longitudinal consistency must be kept. It is useful to keep all patient's registries associated with a single pseudonym.

Pseudonymization can be recoverable or unrecoverable, i.e., with or without the possibility to reidentify the data subject. A recoverable schema could be a secret query table that, when authorized, could be used to reidentify the data subject. Likewise, an unrecoverable schema could be a temporary table that is destroyed at the end of the process. However, it is essential to say that pseudonymized data will always have the risk of unauthorized reidentification.

The de-identification process contains three data types:

- Direct identifiers: are unique identifiers, such as social security number, that can be directly linked to an individual;

- Indirect or quasi-identifiers are identifiers such as gender, zip code, and so forth that are shared by more than one person, but if used with other identifiers, they could trace an individual;

- Non-identifying data: the remaining data.

The pseudonymization is generally used on direct identifiers, but it could also be used on indirect identifiers to reduce the risk while keeping the data longitudinal consistency.

2.3 Electronic Health Records - EHR

A concise meaning or classification of Electronic Health Records (EHR) is difficult to articulate. The International Organization for Standardization ISO/TR 20514:2005 (HEALTH INFORMATICS - ELECTRONIC HEALTH RECORD - DEFINITION, SCOPE AND CONTEXT - ISO 20514, 2008) defined a "basic generic" EHR as simply "a repository of information regarding the health status of a subject of care, in computer processable form." This definition is intentionally concise and generic to ensure its application to existing and future users and systems.

The EHR classification adopted in this ISO technical report includes the Electronic Medical Records (EMR) and Electronic Clinical Records (ECR) often used in the hospital context with a medical/clinical focus; the EHR as an overall collection of a patient's health information from all sources; an Electronic Patient Record (EPR) as a patient's medical information from a single healthcare provider; and a Personal Health Record (PHR), containing information entered by the physician and the patient.

A more specific definition comes from Health et al. (2006); it reads:

> "The Electronic Health Record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. This information includes patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. The EHR automates and streamlines the clinician's workflow. The EHR has the ability to generate a complete record of a clinical patient encounter, as well as supporting other care-related activities directly or indirectly via interface—including evidence-based decision support, quality management, and outcomes reporting."

It is important to note that an EHR is generated and maintained within an institution, such as a hospital, integrated delivery network, clinic, or physician office. An EHR is not a longitudinal record of all care provided to the patient in all venues over time.

The first known medical record was developed by Hippocrates in the fifth century B.C. He prescribed two goals (HEALTH et al., 2006):

- A medical record should accurately reflect the course of the disease.
- A medical record should indicate the probable cause of disease.

These goals are still appropriate, but electronic health record systems can also provide additional functionality, such as interactive alerts to clinicians, interactive flow sheets, and tailored

order sets, all of which cannot be done by paper-based systems. The first EHRs began to appear in the 1960s.

Nguyen, Bellucci and Nguyen (2014) shows that EHR implementations are on-going to various levels of acceptance and success. EHRs are still predominantly used in hospitals, though significantly increasing in primary care. EHR for patient use at home is still at an early stage. Clinicians tend to have positive attitudes regarding EHR, but hold mixed levels of satisfaction and perceptions towards this system's quality and its impact on clinical practice and patient care.

2.4 Natural Language Processing - NLP

Natural language processing (NLP) investigates the use of computers to process or to understand human (i.e., natural) languages to perform useful tasks. NLP is an interdisciplinary field that combines computational linguistics, computing science, cognitive science, and artificial intelligence. Typical applications in NLP include speech recognition, spoken language understanding, dialogue systems, lexical analysis, parsing, machine translation, knowledge graph, information retrieval, question answering, sentiment analysis, social computing, natural language generation, and natural language summarization (DENG; LIU, 2018).

Natural language is a system constructed specifically to convey meaning or semantics and is by its fundamental nature a symbolic or discrete system. The surface or observable "physical" signal of natural language is called text, always in a symbolic form.

From a historical perspective, the development of the general methodology used to study NLP is a rich interdisciplinary field. The development of NLP can be described in terms of three major waves, each of which is elaborated below (DENG; LIU, 2018).

In the First wave, called Rationalism, it is possible to observe an approach based on the belief that knowledge of the language in the human mind is fixed in advance by generic inheritance, dominated most of NLP research between about 1960 and late 1980s. Rationalist approaches endeavored to design hand-crafted rules to incorporate knowledge and reasoning mechanisms into intelligent NLP systems.

Up until the 1980s, most notably successful NLP systems, such as ELIZA for simulating a Rogerian psychotherapist and MARGIE for structuring real-world information into concept ontologies, were based on complex sets of handwritten rules. This period coincided approximately with the early development of artificial intelligence, characterized by expert knowledge engineering, where domain experts devised computer programs according to the knowledge about the very narrow application domains they have. The experts designed these programs using symbolic logical rules based on accurate representations and engineering of such knowledge.

The second wave of NLP, called Empiricism, was characterized by the exploitation of data corpora and (shallow) machine learning, statistical or otherwise, to make use of such data. With the increasing availability of machine-readable data and a steady increase of computational

power, empirical approaches have dominated NLP since around 1990.

In contrast to rationalist approaches, empirical approaches assume that the human mind begins to work with general operations for the association, pattern recognition, and generalization. Rich sensory input is required to enable the human mind to learn the detailed structure of natural language. Early empirical approaches to NLP focused on developing generative models. Since the late 1990s, discriminative models have become the de facto approach in a variety of NLP tasks. Representative discriminative models and methods in NLP include the maximum entropy model, supporting vector machines, conditional random fields, maximum mutual information, minimum classification error, and perceptron.

The Empiricism in NLP and speech recognition in this second wave was based on data-intensive machine learning, which now is called "shallow" due to the general lack of abstractions constructed by many-layer or "deep" representations of data which would come in the third wave.

The third wave involves the use of Deep Learning. With a few exceptions, the (shallow) machine learning models for NLP often did not have the capacity sufficiently large to absorb the vast amounts of training data. Further, the learning algorithms, methods, and infrastructures were not powerful enough.

All this changed several years ago, giving rise to the third wave of NLP, propelled by the new paradigm of deep-structured machine learning or deep learning. In traditional machine learning, features are designed by humans, and feature engineering is a bottleneck, requiring significant human expertise. Deep learning breaks away the above difficulties by the use of deep, layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms.

2.5 Machine Learning

Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions. Here, experience refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis. This data could be in the form of digitized human-labeled training sets or other types of information obtained via interaction with the environment. In all cases, its quality and size are crucial to the success of the predictions made by the learner (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

Machine learning consists of designing efficient and accurate prediction algorithms. As in other computer science areas, some critical measures of the quality of these algorithms are their time and space complexity. However, in machine learning will also need a notion of sample complexity to evaluate the sample size required for the algorithm to learn a family of concepts. More generally, the guarantee of theoretical learning for an algorithm depends on the complexity of the concept classes considered and the size of the training sample.

Since the success of a learning algorithm depends on the data used, machine learning is inherently related to data analysis and statistics. More generally, learning techniques are data-driven methods combining fundamental concepts in computer science with ideas from statistics, probability, and optimization.

2.5.1 Multilayer Neural Networks

Multilayer neural networks contain more than one computational layer. The perceptron contains an input and an output layer, of which the output layer is the only computation-performing one. The input layer transmits the data to the output layer, and all computations are completely visible to the user. Multilayer neural networks contain multiple computational layers; the additional intermediate layers (between input and output) are referred to as hidden layers because the computations performed are not visible to the user (AGGARWAL et al., 2018).

The specific architecture of multilayer neural networks is referred to as feed-forward networks, because successive layers feed into one another in the forward direction from input to output. The default architecture of feed-forward networks assumes that all nodes in one layer are connected to those of the next layer. Therefore, the architecture of the neural network is almost fully defined, once the number of layers and the number/type of nodes in each layer have been defined. The only remaining detail is the loss function that is optimized in the output layer.

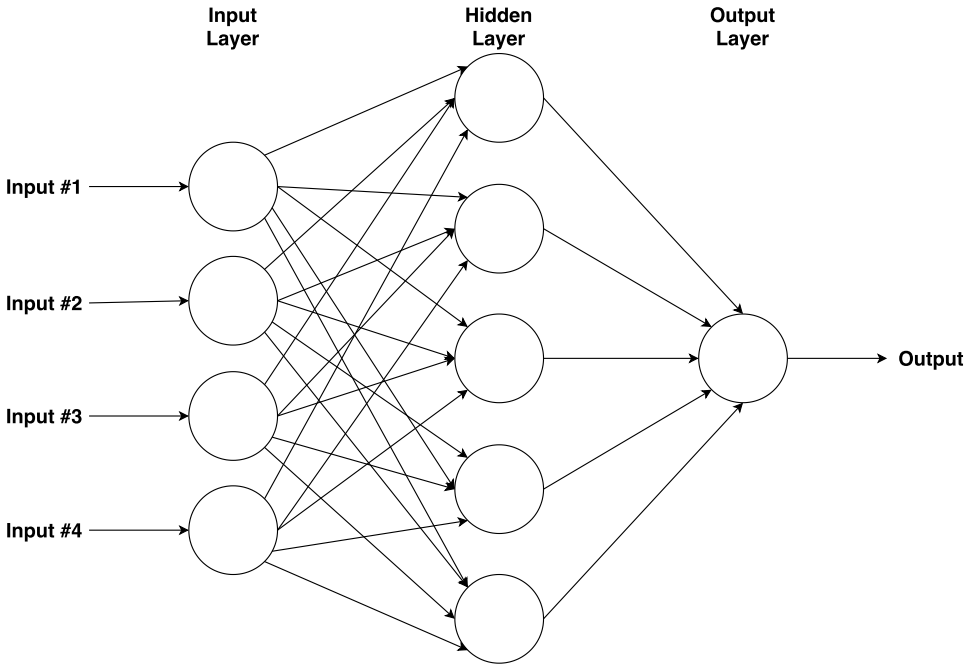

Figure 1 – A hypothetical example of Multilayer Perceptron Network.

2.6 Deep Learning

Conventional machine-learning methods were limited in their ability to process natural data in their raw form. For decades, constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved state-of-the-art speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech, and audio, whereas recurrent nets have shone a light on sequential data such as text and speech (LECUN; BENGIO; HINTON, 2015).

Deep learning is making significant advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is, therefore, applicable to many domains of science, business, and government. In addition to beating records in image recognition and speech recognition, it has surpassed other machine-learning methods at predicting the activity of potential drug molecules, analyzing particle accelerator data, reconstructing brain circuits, and predicting the effects of mutations in non-coding DNA on gene expression and disease. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding, particularly topic classification, sentiment analysis, question answering, and language translation.

2.6.1 Long Short-Term Memory - LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. RNNs have problems associated with vanishing and exploding gradients, i.e., that a neural network that uses only multiplicative updates is good only at learning over short sequences, and is therefore inherently endowed with good short-term memory but poor long-term memory. To address this problem, a solution is to change the recurrence equation for the hidden vector with the use of the LSTM. The operations of the LSTM are designed to have fine-grained control over the data written into this long-term memory (AGGARWAL et al., 2018). It can process not only single data points (such as images) but also entire sequences of data (such as speech or video).

LSTMs consist of multiple layers of nodes. Each node is also connected to adjacent nodes within the same layer (giving the network a sequence component). Furthermore, each node can remember previous information that persists through training steps, giving them a "memory" component.

Figure 2 shows an LSTM architecture with the three gates (input gate, forget gate and output gate), block input, activation functions, weights, a single cell, and the block output, which is repeatedly connected back to the input block and all gates.

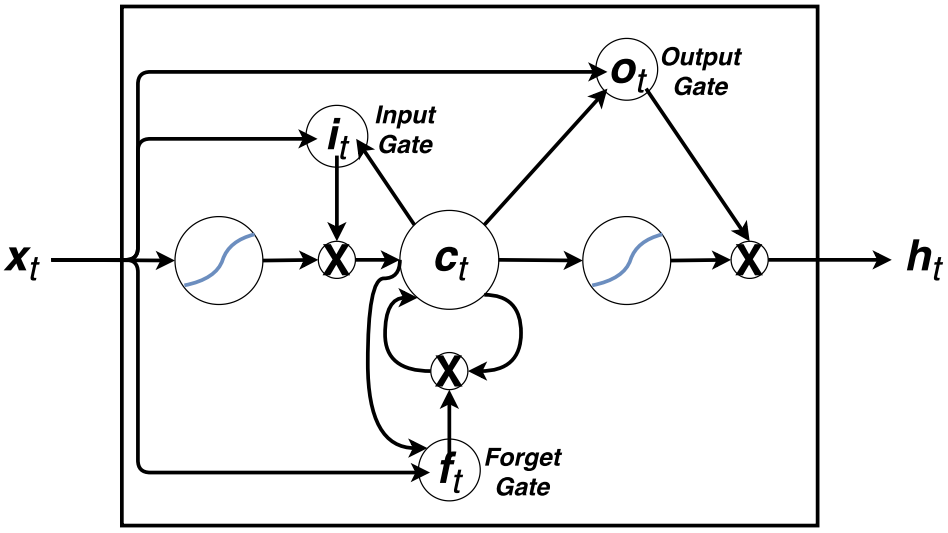

Figure 2 – A simple LSTM gate with only input, output, and forget gates.

3 RELATED WORKS

This chapter describes the effort developed in a literature review to support this research. A strategy composed of two distinct moments was adopted, dedicated to different objectives, being the first to obtain a general view of the research possibilities. This first study supported the decision on the main gaps and needs in the general field, providing ground to the choice of the main objective. The second study was aimed to collect in-depth information on the main focus of the developed work.

Therefore, a preliminary exploratory study was developed to support the definition of this work's objectives. This study was performed with two systematic review surveys. The first survey studied the Clinical Decision Support System (CDSS) on oncology patient diagnosis, and it is available in Appendix A. The survey provided insight in key aspects in the area. These insights are that patient's clinical data are often inputted as free-text format, and that there was a lack of structured data repositories on this topic. This topic was related as a significant problem in most studied papers considering that CDSS needs data in a structured format.

Considering this context, the second survey studied methods of text classification and information extraction in oncology clinical notes to understand how to better deal with the unstructured data. This survey is available in Appendix B. This second survey resulted in the selection of thirteen papers, listed in details in Table 1. The papers used one or a combination of natural language processing (NLP), machine learning, and deep learning methods. Table 1 presents the related works' main characteristics, which will be discussed below.

Considering the related works study conclusions, both text classification and information extraction are essential tasks to deal with unstructured data in the healthcare area. Hence, Table 1 presents papers that implemented text classification or information extraction tasks.

Paper (citation)	Publisher	Year	NLP preprocessing level	Embedding method	Corpora type	Corpora language	Machine learning methods	Deep learning methods
Banerjee et al. (2019)	Science Direct	2019	Advanced	word2vec	Domain-specific	English	Logistic Regression	None
Barash et al. (2020)	Springer	2020	Advanced	BoW, word2vec	Domain-specific	Hebrew	Logistic Regression	LSTM
Carrodeguas et al. (2019)	Science Direct	2019	Basic	BoW	Domain-specific	English	SVM, Logistic Regression	LSTM
Dev et al. (2017)	IEEExplore	2017	Not described	BoW, word2vec	Domain-specific	English	Logistic Regression, Random Forest	LSTM
Gupta, Banerjee and Rubin (2018)	Science Direct	2018	Advanced	BoW, word2vec	Domain-specific	English	k-means clustering	None
Li and Mao (2019)	Science Direct	2018	Basic	word2vec	Open challenge	English	None	CNN
Moen et al. (2020)	JAMIA	2020	Not described	BoW, word2vec	Domain-specific	Finnish	SVM, Random Forest	LSTM, CNN
Oliva et al. (2019)	Science Direct	2018	Advanced	None	Domain-specific	Brazilian Portuguese	None	None
Qiu et al. (2018)	IEEExplore	2018	Basic	VSM	Domain-specific and General-purpose	English	None	CNN
Shao et al. (2018)	IEEExplore	2018	Basic	BoW, word2vec, doc2vec	Domain-specific	English	SVM	None
Si et al. (2019)	arXiv	2019	Basic	word2vec, GloVe, ELMo, BERT	Open challenge and General-purpose	English	None	LSTM
Wang et al. (2019)	BMC	2019	Not described	word2vec	Domain-specific and Open challenge	English	SVM, Random Forest, Multilayer Perceptron	CNN
Wang et al. (2020)	EJBI	2020	Basic	word2vec, GloVe	Domain-specific	English	None	GRU

Table 1 – Related works list comparing its main characteristics.

The next items describe in detail the insight obtained with the study of the papers listed in table 1.

3.1 NLP preprocessing level

The preparation of the corpus to be transformed and used by machine learning and deep learning methods was performed in all papers. This step is called corpus preprocessing and it improves the performance of the algorithms.

According to the column "NLP preprocessing level" in Table 1, the preprocessing was classified into two levels:

- Basic: simple NLP tasks were performed, such as convert text to lowercase, stopwords and punctuation removal, duplicated elements removal, and tokenization;
- Advanced: more complex NLP tasks were performed, such as lemmatization, sentence splitting, part-of-speech (POS), stemming, dependency parsing, and low-frequency words removal, among others.

3.2 Embedding method

All papers that applied machine learning and deep learning algorithms used embeddings methods to extract and combine text features, as seen in the column "Embedding method" in Table 1. These methods can be applied to transform the text into a vector structure, or to include previous preprocessed information from a broader corpora.

The word embedding is the most frequently used component by works that apply deep learning methods, but the BoW was still used with consistent performance. In most papers a general-purpose word embedding was used. A few papers described the creation of its word embedding, but it requires a high computational effort to be built.

3.3 Corpora type

The corpora used in the related works could be classified in the following types (as seen in the column "Corpora type" in Table 1):

- Open challenge: corpora from open NLP challenges, such as i2b2[1] and SemEval[2], which provide a set of annotated and unannotated de-identified patient clinical data;
- Domain-specific: corpora obtained from a specific human knowledge domain. For this work all corpora of this type were of healthcare domain;
- General-purpose: corpora obtained from large databases containing several domains, such as Gigaword, Wikipedia, BooksCorpus+ English Wikipedia, Google News articles, and PubMed biomedical publications.

[1] https://www.i2b2.org/index.html
[2] http://alt.qcri.org/semeval2020/

The open NLP challenge corpus is useful to compare different models. Considering all works would have the same corpus, it is possible to compare the performance of the different models. However, the authors must perform their work on the domain the challenge corpus was created.

When created by the work, domain-specific corpora enable the opportunity to explore a specific domain that was not explored or still has space for improvements. Nevertheless, this option brings difficulties for comparisons among works and approaches.

3.4 Corpora language

As seen in the column "Corpora language" in Table 1, most selected papers deal with the English language, while some few examples, in Portuguese (1), Finnish (1), and Hebrew (1), are also noticed.

According to our study, the main techniques, both for preprocessing and for experiments, do not have significant differences due to the selected language in the work. Nevertheless, some linguistic resources and embeddings do present significant quality issues among different languages.

3.5 Machine Learning methods

The use of Artificial Intelligence approaches can be observed in almost all papers. Some used traditional machine learning algorithms, which presented similar results to newer deep learning algorithms, especially in the text classification task. It shows that some tasks do not need sophisticated techniques, or the experiments are not adequate in the quantity of data to adequately support the deep learning approaches.

The following machine learning methods were used in the assessed papers (as seen in the column "Machine learning methods" in Table 1: k-means clustering, Logistic Regression, Multi-Layer Perceptron, Random Forest, SVM.

3.6 Deep Learning methods

The use of deep learning models, with promising results reaching state-of-the-art performance or even improving it, is a trend in the last few years.

Deep learning methods are capable of considering large sections of text. Deep learning relies on stacked nodes and layers to automatically extract features that are believed to be analogous to human thinking. With deep learning, researchers started to construct models without complicated feature engineering and to minimize the reliance on NLP toolkits for feature acquisition.

According to the evaluated papers, the deep learning methods recently applied to NLP in-

formation extraction are CNN, GRU, and LSTM recurrent neural networks. CNN is a neural-based approach representing a feature function that is applied to constituting words or n-grams to extract higher-level features. LSTM-based models have been proposed for the sequence to sequence mapping (via encoder-decoder frameworks). GRU models are similar to LSTM but with fewer parameters.

From all papers that applied deep learning methods, the LSTM recurrent neural network was the most frequently used.

3.7 Limitations of the assessed related works

Several related work papers used domain-specific corpora (according to Table 1), but none of them used a corpus with non-synthetic medical notes from the oncology healthcare-specific domain in the Brazilian Portuguese language.

Furthermore, considering the text classification task, most of the papers applied machine learning or deep learning methods. Just a few of them presented a comparison between several machine learning and deep learning methods.

4 MATERIAL AND METHODS

This chapter presents the methodological aspects adopted in this work. This work aims to evaluate text classification techniques to support health professional needs regarding diagnosis decisions. Therefore, text classification experiments were conducted using machine learning and deep learning approaches. These techniques were applied in Brazilian Portuguese clinical notes corpora obtained from a system in the oncology domain. The corpora creation is an essential element in this research and constitutes one of the contributions of this work, because it was created from non-synthetic data with specific needs of preprocessing. Other contributions of this work are the corpus de-identification and enrichment process, described in the following sections. The evaluation and performance comparison of several machine learning and deep learning classifiers also constitutes a contribution of this work.

The following sections present details of this approach. Section 4.1 presents an overview of this work's main elements. In Section 4.2, the Oncology EHR system used as the basis for the corpora generation is presented, and the system's components used by the healthcare professional users to input the data are explained. In Subsection 4.2.1, the system's database is detailed and the different clinics' databases used in this work are described.

Section 4.3 presents an overview of the corpus' creation, and its steps are detailed in its subsections. Subsection 4.3.1 details the data anonymization process. Subsection 4.3.2 features how the data were loaded to be processed and exported. Subsection 4.3.3 describes how the corpora were annotated, and, in Subsection 4.3.4, how they were enriched with structured data. In the last Subsection (4.3.5), how the corpora were exported to be used by the text classification algorithms is described.

The Section 4.4 describes the model used by the general format of the experiments performed in this work. The experiments performed and its results, as well their evaluations, are described in the next Chapter 5.

4.1 Overview of the proposed approach

In this section, an overview of the proposed approach is described. Figure 3 shows the general view containing the elements of the proposed approach for this work. This context is derived from real-world observation, considering real medical clinics and expresses health professional needs.

The overall process starts with the creation of a record of a medical note by the healthcare professional. In the real world observed cases, this situation generates textual medical records and structured clinical information. In this step, the healthcare professionals use the Oncology EHR system (INTERPROCESS, 2019) to input their observations about the patient, which are recorded in the system's database (as described in Section 4.2). These observations can be composed of free-text and structured data, and both data types are used together to achieve

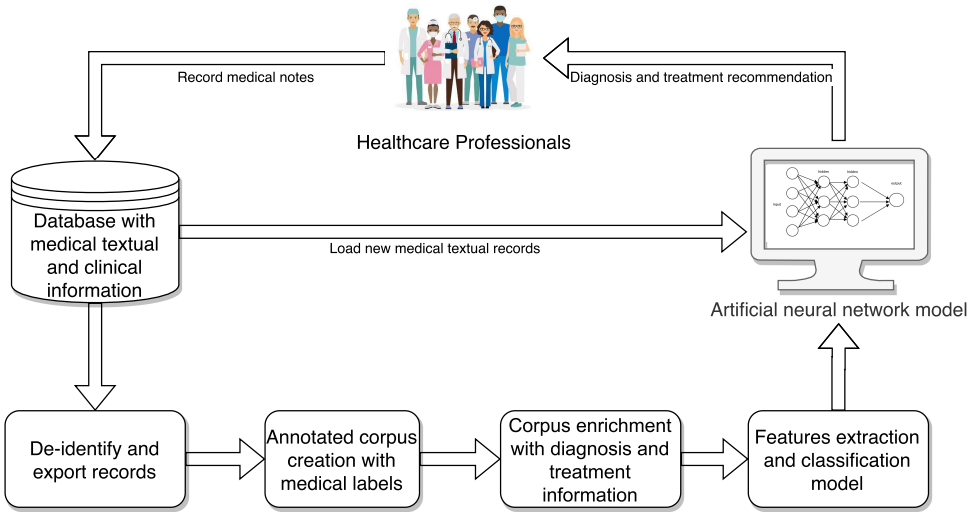

Figure 3 – General view of the approach

better results.

The upper part of Figure 3 describes the primary steps considered for exploring possible answers to the research question in this work. The development of a flow of operations that start with the assistance of the health professional was evaluated, as identified in the flow "Load new medical textual records" that will be used together with a support service, applying the classification models studied. The application of the models will generate support for determining a response with suggestions for framing and similar contexts, as represented by the "Diagnosis and treatment recommendation" flow that will be used by the physician. The generation of new information about each patient's clinical event is stored in the Oncology EHR system's database, as indicated in the "Record medical notes" flow.

This initial flow consists of a vision of the future use of the system by physicians and healthcare professionals. To support this flow, the other items in Figure 3 were studied and experimented within this proposal.

The first necessary step, to generate the corpora and to use the proposed approach in the future, was to anonymize and export these records. A de-identification process was applied to anonymize the data, preventing the personal identification of each patient and professional. After the de-identification process, the records were exported, creating the corpus used in text classification approaches (as described in Section 4.3). In this step, the textual data was exported along with some structured clinical information.

The corpus was annotated (Section 4.3.3), taking advantage of free-text data labels generated by the medical staff when using the EHR system. Therefore, this corpus was able to be considered for some of the classification approaches tasks, such as the training tasks.

After the corpus creation and annotation, it was enriched with some structured clinical information (as described in Section 4.3.4). As mentioned previously, the EHR system stores the data as a free-text type or structured types. Some of these structured data are the diagnosis and

treatment information used to enrich the corpus.

The final step involves all the preprocessing steps necessary for training the evaluated classification algorithms. Before training the artificial neural network model, the corpus was preprocessed as described in Section 4.4.1. In this step, the corpus' annotations and enriched data were assessed. When necessary, textual features were extracted. After that, the corpus and its features were used as input to train the following Artificial Neural Network (ANN) models: Multilayer Perceptron (MLP) and Long short-term memory (LSTM). Part of the corpus was reserved for testing to evaluate each ANN model's performance (as described in Section 4.4).

Various experiments were developed, combining different approaches such as corpus format, new corpus with updated information, document or paragraph processing levels, and different machine learning and deep learning classifiers (as described in Chapter 5). The experiments' results are evaluated in Section 5.3.

As a result of the work overviewed above, the creation of a corpus with non-synthetic oncological medical notes and the implementation of a de-identification and enrichment process of the corpus are highlighted. In addition, the evaluation and performance comparison of a machine learning and a deep learning classifier, along with the evaluation of its application in the Oncology EHR system, are also highlighted contributions of this work.

In the following sections, each component or step involved in the general approach of Figure 3 is detailed.

4.2 Oncology EHR system

The corpus used by this work is a set of medical notes obtained from an Oncology Medical EHR system (INTERPROCESS, 2019). These medical notes contain clinical events recorded by physicians, nurses, and other health professionals, with clinical information about the patient.

To record this information, the health professional uses a customizable form from the EHR system for each clinical event. These forms contain one or more questions, with a variety of types of fields like numbers, dates and time, lists, and free-text inputs. The free-texts inputs allow health professionals to insert text in natural language. For more information about the system's database, please refer to Section 4.2.1.

In Figure 4, it is possible to observe a customizable form in insert mode, with a free-text field highlighted inside the red box. Below this field, there is another free-text field and a structured field.

The customizable forms enable healthcare professionals to create personalized options of inserting data, which are used to take medical notes. During the form creation process, the healthcare professionals (in this case, "EHR system users," or just "users") can create their questions to apply and answer them during a clinical event, such as a consultation. Each question has a label and some properties, which describe the type of answer expected. These are the available answer's types:

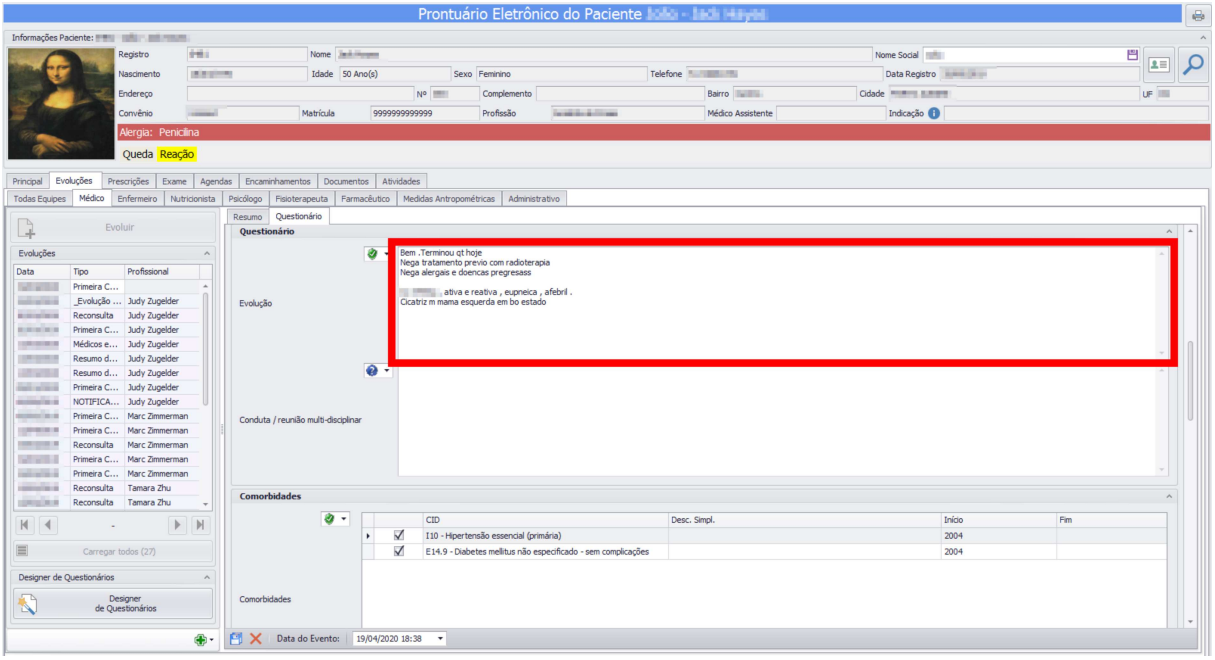

Figure 4 – Customizable form in insert mode with a free-text field highlighted in the red box

- Text: this type has three sub-types: text, number, and date-time. The type text is the most commonly used, and enable the users to insert free-text data into the medical note;

- List: define that the answer is an item (could be more than one) from a designated list of values;

- Component: this type represents the use of predefined components built for specific cases. When used, it is necessary to define which component should be applied. Examples of components are ICD search box, patient's measurements (i.e., height, weight, body surface area), patient's main diagnosis, and so forth.

The label of each question is defined by the user and describes the question itself or its category. These labels will be used to annotate the corpus later in the corpus annotation step (Section 4.3.3).

The "component" type enables the user to input the information through specific purpose components, like the main diagnosis based on ICD code, patient's measurements, allergies, medications, and so forth. Furthermore, the user may use the electronic prescribing feature to define the patient's treatment based on chemotherapy protocols. These features enable the EHR system to get the data and store it in a structured format. Some of these structured data will be used later in the corpus enrichment step (Section 4.3.4).

4.2.1 About the EHR system's database

The Oncology EHR system's database (DB) stores all information about the patients' health, such as medical notes, electronic prescriptions, and patients' appointments history.

	Small	**Medium**	**Large**
Physicians	1,653 (2)	23,119 (18)	248,639 (210)
Nurses	1,269 (1)	16,076 (5)	185,518 (43)
Pharmacists	386 (1)	1,133 (4)	36,072 (12)
Psychologists	-	1,441 (4)	13,134 (17)
Nutritionists	-	1,145 (2)	6,964 (9)
Physiotherapists	-	51 (1)	3,031 (5)
Dentists	-	-	165 (4)
Social service professionals	-	-	84 (2)
Receptionists	-	75 (2)	-
Other professionals	-	-	166 (7)
Medical notes' total	**3,308**	**43,040**	**493,773**
Distinct Patient's total	**397**	**5,407**	**40,335**
Distinct Professional's total	**4**	**36**	**309**

Table 2 – Comparison of different Oncology EHR system's database sizes, with the number of medical notes per professional category and, in parentheses, the number of individuals that registered the notes.

For each oncology clinic that uses the system, an instance of the database is created to store its information separately. Three databases from different clinics were used in this work. The clinics are from different sizes in terms of the number of medical appointments and procedures performed monthly.

In Table 2, a comparison between the three databases is presented. This table presents the number of medical notes registered per professional category and in parentheses the number of professionals who registered these notes. In all three databases, physicians and nurses are the two professionals categories that have recorded medical notes the most, and this is due to the higher demand for patient's consultations and care with these professionals.

The database stores not only clinical information, but also information about finances, health insurance billings, pharmacy controls, and management reports. Only the clinical information was loaded, and in Figure 5, the entity relational (ER) diagram of this information is shown. This ER diagram can be segmented in five sections, as follows:

- Personal identification contains the patient's and professional's identification. It includes the following tables: Patient, Professional;

- Customizable forms structure contains the information about the customized form, its groups of questions, and its individual questions. It includes the following tables: CustomForm, CustomFormGroup, CustomFormQuestion;

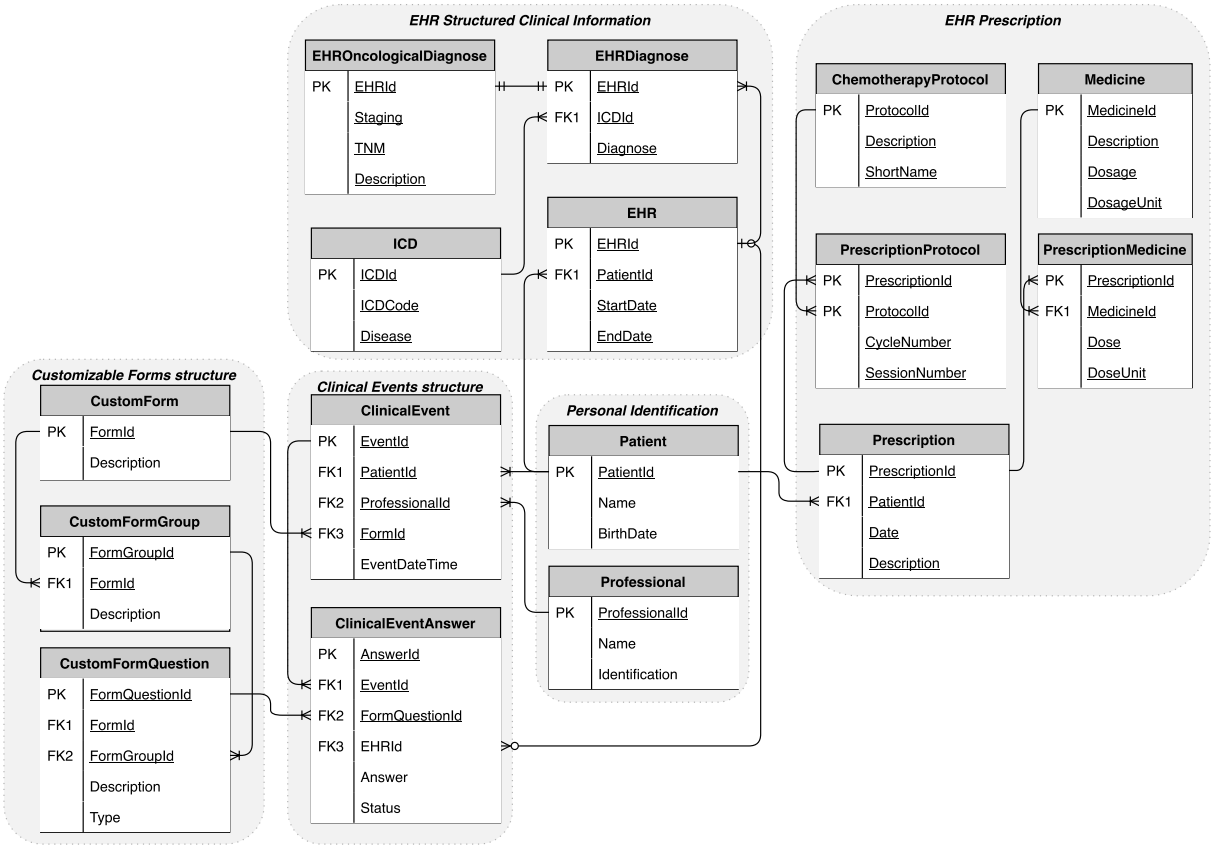

Figure 5 – The clinical events and medical notes' entity relational (ER) diagram

- EHR structured clinical information is a group of tables that holds the patient's clinical information in a structured format. Only the tables used in this work were considered. It includes the following tables: EHR, EHRDiagnose, EHROncologicalDiagnose, ICD;

- EHR prescription is a group of tables with information about the patient's chemotherapy prescription. Each prescription is generated by one or more protocols, which has the medicine schemas to generate the patient's chemotherapy treatment. It includes the following tables: Prescription, PrescriptionProtocol, PrescriptionMedicine, ChemotherapyProtocol, Medicine;

- Clinical events structure contains the information about consultations with the patient, or any information about an event related to the patient. It includes the following tables: ClinicalEvent, ClinicalEventAnswer.

The tables in the ER diagram (Figure 5) were queried to obtain the information, which will be processed as described in the further sections.

4.3 Corpus

The corpus used by this work was created from the Oncology EHR system's free-text medical records and some structured clinical information, as described in Section 4.2.

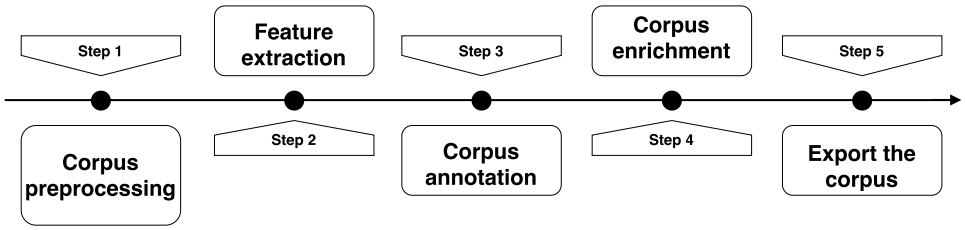

Figure 6 – Pipeline built to create the corpus

A pipeline was built to generate the corpus, as seen in Figure 6. The following steps compose the pipeline:

- De-identify the information in the EHR system's database (Section 4.3.1);
- Load the data from the database (Section 4.3.2);
- Annotate them with the labels from the EHR system (Section 4.3.3);
- Enrich the corpus with structured clinical information (Section 4.3.4);
- Export the corpus in the appropriate format to be used by the ANN algorithms (Section 4.3.5).

The pipeline was built using scripts SQL (Structured Query Language) and Python programming language. The former was executed in the Microsoft SQL Server Management Studio to perform the pseudonymization step, and the latter was a Python console application that ran the load, annotation, and enrichment steps.

In the first step, a set SQL script (queries and procedures) was executed on the database to de-identify its information, detailed in Section 4.3.1. This process was applied to avoid patients' and professionals' unique identification before this information was processed and exported.

After the anonymization step, in the Load-EHR records step (detailed in Section 4.3.2), the application connects to the EHR system's database to query the records with the relevant fields. While the records were iterating, the following two steps were applied:

1. The corpus annotation was prepared, as described in Section 4.3.3;
2. The corpus was enriched with structured clinical information, described in Section 4.3.4.

At the pipeline's final step, all records that were anonymized, loaded, annotated, and enriched are then exported to a format that suits the machine learning and deep learning classifiers needs. In Section 4.3.5, this step is detailed.

This pipeline was executed with different instances of the Oncology EHR system's database, where each instance represents an oncology clinic. For each database instance, a corpus was created. In Chapter 5, the performed experiments are described with the created corpora.

4.3.1 Data de-identification

As the first step of the corpus creation, a de-identification (or pseudonymization) process was applied. It was necessary to avoid the disclosure and unique identification of patients and healthcare professionals. As explained in Section 2.2, there are different levels of data anonymization. Pseudonymization is a de-identification sub-category, where the person's data are kept together without disclosing his or her identity.

In this step, an unrecoverable pseudonymization process was applied on direct identifiers, i.e., without the possibility of reversing this process and revealing the original patient's and professional's identity. Furthermore, patients' or professionals' demographic data were not exported, making it harder to reverse the pseudonymization process.

Several pipeline versions were built to generate different corpus formats. For all versions, the same pseudonymized process was applied. This process consists of changing the direct identifiers by pseudonyms, in this work randomly created data was applied. The following fields were changed in the patients' and professionals' database tables: name, Brazilian's identity card numbers (RG, CPF), phone numbers, and address. This information was not used in the corpus creation, but it was pseudonymized just in case. The pseudonymized data consists of:

- A list of fictitious names (first and last names) randomly combined;
- Fictitious identity card numbers;
- Fictitious phone numbers;
- Fictitious addresses.

The fields listed above were updated with these new data. If the next steps of the pipeline query them, they will get pseudonymized data.

Pseudonymizing the free-text fields, a process to remove proper names, was applied. The proper-name fields was used from the following personal identification tables: Patient, Professional, Suppliers. A list with first and last names was created with the names from these tables. This list was used to find the first and last names in the texts and when found replace them by a pseudonymized name.

In Figure 7, a sample of the SQL script that creates a base table with the fictitious names to be randomly applied in the pseudonymized process is shown. These names were randomly combined, creating a male and a female full name table, as presented in Figure 8. Finally, Figure 9 shows how these tables were used to update the name on the patient's table.

```
CREATE TABLE [dbo].[BaseNome](
    [Nome] varchar(50) NOT NULL,
   [Tipo] char(1) NOT NULL)--Types: 'M': male first name
                            --       'F': female first name
                            --       'L': unisex last name
 GO
  --Male first names, Type=M
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Aaron', 'M')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Abbot', 'M')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Abdul', 'M')

--Female first names, Type=F
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Zoe', 'F')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Zorita', 'F')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Jacqueline', 'F')

--Unisex last names, Type=L
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Abbott', 'L')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Acevedo', 'L')
INSERT [dbo].[BaseNome] ([Nome], [Tipo]) VALUES (N'Acosta', 'L')
```

Figure 7 – Sample of SQL script to create the names source table

```
-- MALE: randomly join the male's first and last names - creates the male's names table
select    top(select max(idpaciente) from Paciente) PrimeiroNome.Nome + ' ' + Sobrenome.Nome as [Nome],
        id = IDENTITY(INT,1,1)
into #baseMasculino
from BaseNome PrimeiroNome
full outer join BaseNome Sobrenome on (PrimeiroNome.Nome <> Sobrenome.Nome)
where PrimeiroNome.Tipo = 'M'
and Sobrenome.Tipo = 'L'
order by newid()

-- FEMALE: randomly join the female's first and last names - creates the female's names table
select    top(select max(idpaciente) from Paciente) PrimeiroNome.Nome + ' ' + Sobrenome.Nome as [Nome],
        id = IDENTITY(INT,1,1)
into #baseFeminino
from BaseNome PrimeiroNome
full outer join BaseNome Sobrenome on (PrimeiroNome.Nome <> Sobrenome.Nome)
where PrimeiroNome.Tipo = 'F'
and Sobrenome.Tipo = 'L'
order by newid()
```

Figure 8 – SQL script that randomly creates male and female full name tables, with the joint of first and last names

```
--Update male names on the patient's table
Update paciente
    set Paciente.nome = #baseMasculino.Nome,
        Paciente.Pesquisa = dbo.NomePesquisa(#baseMasculino.Nome),
        Paciente.CPF = '99999999999',
        Paciente.Fone = '99 3333-3333',
        Paciente.FoneR1 = '',
        Paciente.FoneR2 = '',
        Paciente.FoneCelular = '99 999999999'
from Paciente
    inner join #baseMasculino on #baseMasculino.id = paciente.idpaciente
    where paciente.Sexo = 'M'

--Update female names on the patient's table
Update paciente
    set Paciente.nome = #baseFeminino.Nome,
        Paciente.Pesquisa = dbo.NomePesquisa(#baseFeminino.Nome),
        Paciente.CPF = '99999999999',
        Paciente.Fone = '99 3333-3333',
        Paciente.FoneR1 = '',
        Paciente.FoneR2 = '',
        Paciente.FoneCelular = '99 999999999'
from Paciente
    inner join #baseFeminino on #baseFeminino.id = paciente.idpaciente
    where paciente.Sexo = 'F'
```

Figure 9 – SQL script to update male and female names on the patient's table

4.3.2 Loading the data from the Oncology EHR system's database

After the pseudonymization process, the data were loaded from the database to be iterated and processed. This process was done by a Python console application that was developed with basically the following steps:

- Loading the data querying the Microsoft (MS) SQL Server database;
- Iterating the data and build the corpus;
- Annotating and enriching the corpus (described in Sections 4.3.3 and 4.3.4).

A SQL query was built and executed on the MS SQL Server database, according to the ER diagram. In Figure 5 it is possible to observe the clinical events' entity relational (ER) model with the patient, professional, customizable form, structured clinical information, and prescription table. A SQL query sample can be observed in Figure 10.

Each record represents an event that a healthcare professional had with a patient, like a consultation. These events may contain one or more structured clinical information and free-text medical notes, registered by the healthcare professional. The following fields were obtained for each record: event's date and time; patient's name, mother's name and identity card number; healthcare professional's name and identification; customizable form's name; and the answer (the data registered in the event).

```
Select Paciente.IdPaciente, Paciente.Nome as PacienteNome, Paciente.Documento,
        Paciente.CPF, Paciente.NomeMae as PacienteNomeMae,
        PronQuestionario.IdProfissional, ChProfissional.Descricao as ProfissionalNome,
        Profissional.Registro as ProfissionalRegistro, PronQuestionarioXML.IdQuestResposta,
        PronQuestionarioXML.Conteudo, PronQuestionario.DataEvento,
        Questionario.Descricao as QuestionarioNome
    from PronQuestionarioXML
        inner join PronQuestionario
            on PronQuestionarioXML.IdQuestResposta = PronQuestionario.IdQuestResposta
        inner join Questionario
            on PronQuestionario.IdQuestionario = Questionario.IdQuestionario
        inner join Paciente
            on PronQuestionario.IdPaciente = Paciente.IdPaciente
        inner join Chave ChProfissional
            on PronQuestionario.IdProfissional = ChProfissional.IdChave
        inner join Profissional
            on ChProfissional.IdChave = Profissional.IdChProfissional
    order by Paciente.IdPaciente, PronQuestionario.DataEvento
```

Figure 10 – SQL query used to query the database

Regarding the answer field, it contains the clinical information about the patient in the corresponding event and is stored as an XML document. Each XML document may contain one or more records, with the following tags: the answer, the answer's label (or the question description), the answer status, and the type. The type tag is essential to define if the answer's content is structured or free-text information. The string-type record contains data in the free-text format with content in natural language, which will be used to create the corpus of this work. In this case, the answer's label will be used to annotate the corresponding text. The remaining records contain structured information, and some of them will be used to enrich the corpus.

The records' iteration was performed per clinical events or patients. In the per-clinical-event iteration, each record represented a clinical event and was processed and exported individually. Each record was annotated (as described in Section 4.3.3) and enriched with structured information (as described in Section 4.3.4).

In the per-patient iteration, all patient's clinical events were processed together. In this case, all clinical events' free-text answers were joined into a large text, creating a single record per patient. Furthermore, each record was enriched with structured information (as described in Section 4.3.4).

After each iteration, the processed data was stored in an in-memory structure. It allowed a standard algorithm to process the data without concern about file types and formats in the corpus export (described in Section 4.3.5).

4.3.3 Corpus annotation

The corpus annotation process benefits from the question's label of the customizable form (more information in Section 4.2). The system's user may personalize a customizable form by

adding questions. Each question has a label that describes its contents, and this label was used in the annotation process described in this Section.

While iterating the records (according to Section 4.3.2), when a free-text data was found, it was stored in an in-memory structure with its corresponding question's label. In the corpus export step, this in-memory structure was used to export the processed data to the desired file type and format.

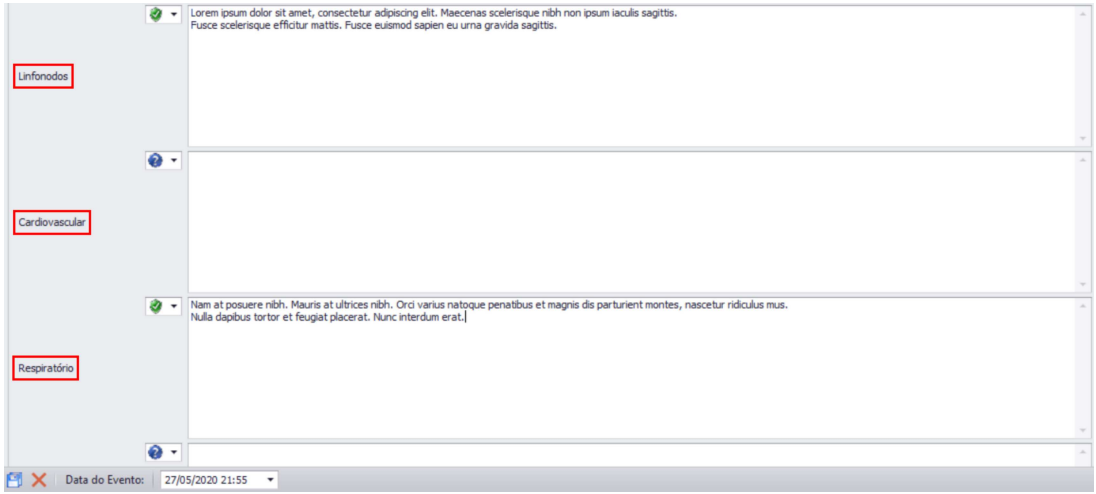

Figure 11 – A customizable form with three free-text questions, with its labels highlighted

	A	B	C	D	E	F
1		SENTENCE_ID	SENTENCE	SETENCE_CATEGORY	ARGUMENT_2	ARGUMENT_2_CATEGORY
20	18	18	Exame físico normal	Ectoscopia / Pele e anexos	C50.9 - Mama	Diagnostico
21	19	19	Continua Anastrozol, Eligard e Zometa Solicito Ex lab e TCs de reestadiamento.	Conduta	C50.9 - Mama	Diagnostico
22	20	20	# Carcinoma ductal invasivo de mama D / Triplo negativo / Pré menopausa # 4x AC + 12 T até 23/09/2015 # RXT 25/11/2015 # Seguimento 5 meses > Recidiva óssea (CO : lesão em oitavo arco costal D) # 17 x Xeloda e Zometa # SVCS pelo PC - TEP - Stent com Dr. Luis Otávio em GYN	Histórico Oncológico	C50.9 - Mama	Diagnostico
23	21	21	Melhora dos sintomas de SVCS após retirada de PC e stent HB 13,8 L 3600 Plaq 120 mil Cr 0,9	Sintomas	C50.9 - Mama	Diagnostico
24	22	22	Boa tolerância ao Tratamento	Toxicidade	C50.9 - Mama	Diagnostico
25	23	23	Hematoma em tórax a E	Ectoscopia / Pele, anexos/ Mamas	C50.9 - Mama	Diagnostico
26	24	24	Impalpáveis	Linfonodos	C50.9 - Mama	Diagnostico
27	25	25	Ritmo cardíaco regular em dois tempos, BNF s/sopro	Cardiovascular	C50.9 - Mama	Diagnostico
28	26	26	MV + ARA	Respiratório	C50.9 - Mama	Diagnostico
29	27	27	Flácido, indolor, sem visceromegalias	Abdominal	C50.9 - Mama	Diagnostico
30	28	28	sem edema	Membros	C50.9 - Mama	Diagnostico
31	29	29	Ca de mama E T2N1M0 Triplo negativo EIIA R Osso em tratamento de primeira linha	Impressão	C50.9 - Mama	Diagnostico
32	30	30	Libero ciclo 18 Xeloda e Zometa	Conduta	C50.9 - Mama	Diagnostico
33	31	31	# Colectomia E por obstrução intestinal _ Adenocarcinoma G2, invasão perineural, T3N0M0 # 2x Roswell Park	Histórico Oncológico	C18.9 - Cólon	Diagnostico
34	32	32	Insônia	Sintomas	C18.9 - Cólon	Diagnostico
35	33	33	Boa tolerância a Qt	Toxicidade	C18.9 - Cólon	Diagnostico

Figure 12 – A sample of a per-clinical-event corpus

The annotation process was applied only when the iteration was performed per clinical events. As mentioned previously, a clinical event contains one or more questions, according to its customizable form (Section 4.2). Therefore, in the per-patient iteration, considering that all patient's clinical events were processed together, it would result in several question labels.

In Figure 11, a customizable form is shown with three free-text questions, and its labels highlighted. In this case, the corresponding label was used to annotate it. In the corpus export

(Section 4.3.5) step, the label was exported along with its text and later used to train the classifiers. Figure 12 shows a per-clinical-event corpus sample, with some texts ("SENTENCE" column) and its annotation label ("SENTENCE_CATEGORY" column).

4.3.4 Corpus enrichment

A corpus enrichment process was developed to leverage the use of structured data available in the EHR system. As described in Section 4.2, in the customizable forms one or more questions of free-text or structured types are possible. The structured types contain information such as ICD oncology diagnosis, chemotherapy protocols, patient's body measurements, allergies, prescribed medicines, and so forth.

The structured and free-text information were used to train the text classifiers models, and later to suggest the category of new texts using trained classifiers. The following patient's structured information was selected: the patient's ICD diagnosis, and the chemotherapy protocol used in the patient's treatment.

	A	B	C	D	E	F
1		SENTENCE_ID	SENTENCE	SETENCE_CATEGORY	ARGUMENT_2	ARGUMENT_2_CATEGORY
20	18	18	Exame físico normal	Ectoscopia / Pele e anexos	C50.9 - Mama	Diagnostico
21	19	19	Continua Anastrozol, Eligard e Zometa Solicito Ex lab e TCs de reestadiamento.	Conduta	C50.9 - Mama	Diagnostico
22	20	20	# Carcinoma ductal invasivo de mama D / Triplo negativo / Pré menopausa # 4x AC + 12 T até 23/09/2015 # RXT 25/11/2015 # Seguimento 5 meses > Recidiva óssea (CO : lesão em oitavo arco costal D) # 17 x Xeloda e Zometa	Histórico Oncológico	C50.9 - Mama	Diagnostico
23	21	21	Melhora dos sintomas de SVCS após retirada de PC e stent HB 13,8 L 3600 Plaq 120 mil Cr 0,9	Sintomas	C50.9 - Mama	Diagnostico
24	22	22	Boa tolerância ao Tratamento	Toxicidade	C50.9 - Mama	Diagnostico
25	23	23	Hematoma em tórax a E	Ectoscopia / Pele, anexos/ Mamas	C50.9 - Mama	Diagnostico
26	24	24	Impalpáveis	Linfonodos	C50.9 - Mama	Diagnostico
27	25	25	Ritmo cardíaco regular em dois tempos, BNF s/sopro	Cardiovascular	C50.9 - Mama	Diagnostico
28	26	26	MV + ARA	Respiratório	C50.9 - Mama	Diagnostico
29	27	27	Flácido, indolor, sem visceromegalias	Abdominal	C50.9 - Mama	Diagnostico
30	28	28	sem edema	Membros	C50.9 - Mama	Diagnostico
31	29	29	Ca de mama E T2N1M0 Triplo negativo EIIA R Osso em tratamento de primeira linha	Impressão	C50.9 - Mama	Diagnostico
32	30	30	Libero ciclo 18 Xeloda e Zometa	Conduta	C50.9 - Mama	Diagnostico
33	31	31	# Colectomia E por obstrução intestinal _ Adenocarcinoma G2, invasão perineural, T3N0M0 # 2x Roswell Park	Histórico Oncológico	C18.9 - Cólon	Diagnostico
34	32	32	Insônia	Sintomas	C18.9 - Cólon	Diagnostico
35	33	33	Boa tolerância a Qt	Toxicidade	C18.9 - Cólon	Diagnostico

Figure 13 – An example of an Excel file with a medical note and its corresponding enriched data highlighted

Section 4.3.2 describes how the information was loaded and iterated. For each record iteration, the enriched data were loaded using two other SQL queries. These queries loaded the patient's diagnosis and protocols according to the loading type, as follows:

- Per-clinical-event loading: the enriched data that was informed to the EHR system up to the event date was loaded, that is, the enriched data that existed at the moment the healthcare professional was registering the clinical event;
- Per-patient loading: in this case, all patient's clinical events were joined in a large text; therefore, the current patient's enriched data was loaded.

This structured information enriched the corpora generated by both loading types (according to Section 4.3.2). In the per-clinical-event loading, each clinical event record was enriched with

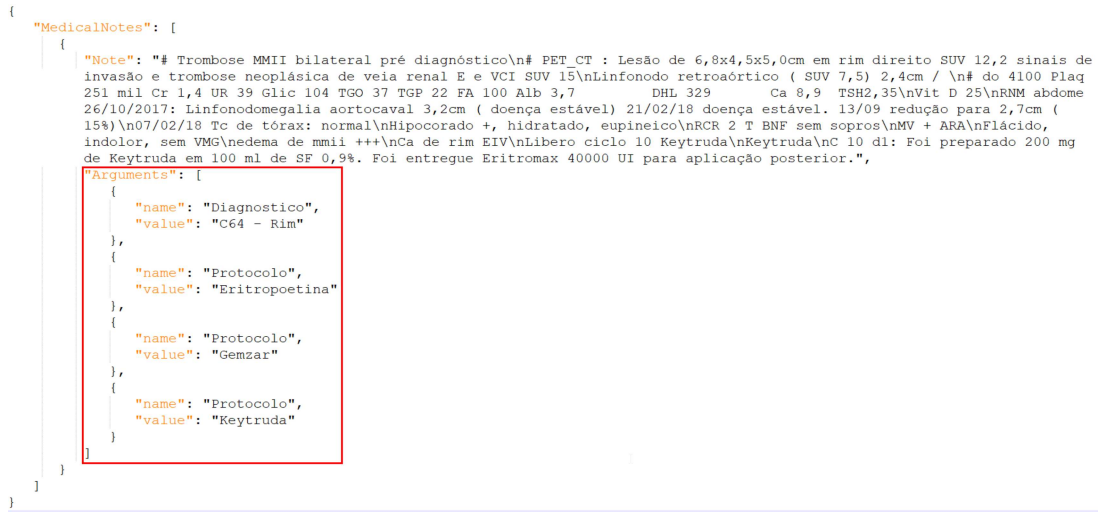

Figure 14 – An example of a JSON file with a medical note and its corresponding enriched data highlighted

the patient's diagnosis and protocols. In the per-patient loading, each patient record with all his or her clinical events was also enriched with the patient's diagnosis and protocols. Figures 13 and 14 respectively present an Excel and a JSON file with a medical note sample with its corresponding enriched data (highlighted with a red box).

4.3.5 Corpus export

In the last step of the corpus creation, all loaded, iterated, and processed data was exported. Two file formats were used to support the experiments performed: Microsoft Excel and JSON. The former was used in the per-clinical-event loading, and the latter was used in the per-patient loading.

The per-clinical-event loading generated a single MS Excel file with a row for each clinical event, as seen in Figure 15. Each row contains the free-text medical note ("SENTENCE" column), the annotation label ("SETENCE_CATEGORY" column), and the enriched data ("ARGUMENT_2" and "ARGUMENT_2_CATEGORY" columns). The question's label was used to annotate the text, as described in Section 4.3.3. In the case of the enriched data, if the corresponding clinical event had more than one diagnosis or protocol the text and its annotation label were repeated for each one.

Three different MS Excel files with the corpus were generated, one for each clinic (more information about the clinics and its databases in Section 4.2.1). Therefore, each file represented a different clinic corpus enabling their processing individually.

The per-patient loading generated a JSON file for each patient, with all clinical events free-text joined and its enriched data, as seen in Figure 16. Each JSON file contains a large text with all the patient's clinical events (tag "Note") and a list of its enriched data (tag "Arguments"). In this case, the corresponding question's label was not used to annotate to text, because it could

	A	B	C	D	E
1	#	SENTENCE	SETENCE_CATEGORY	ARGUMENT_2	ARGUMENT_2_CATEGORY
9	7	1- ADENOCARCINOMA DE PROSTATA EC-IV- OSSO, LINFONODOS ABDOMINAIS tratamento prévio 23/06/16- PROSTATECTOMIA ROBOTICA. PSA PRE OP=4,9, RECEBEU ZOLADEZ EM MAIO 2016 28/12/17 A 04/01/18, RADIOTERAPIA PALIATIVA EM COXO- FEMURAL ESQUERDA- DOSE DE 20GY nov. 2017- LECTRUM =18/11/17- PSA=0,61 = 27/12/17- PSA=0,74 =05.04.18- PSA=0,03 =30.04.18- psa=0,02 2- EM USO DE ABIRATERONA 1000MG/D; PREDNISONA 5MG 12/12HS; (dia 10/01/18) LECTRUM (FEZ NOV. 2017)	Plano Terapêutico atual	C61 - Próstata	Diagnostico
10	8	1- ADENOCARCINOMA DE PROSTATA EC-IV- OSSO, LINFONODOS ABDOMINAIS tratamento prévio 23/06/16- PROSTATECTOMIA ROBOTICA. PSA PRE OP=4,9, RECEBEU ZOLADEZ EM MAIO 2016 28/12/17 A 04/01/18, RADIOTERAPIA PALIATIVA EM COXO- FEMURAL ESQUERDA- DOSE DE 20GY nov. 2017- LECTRUM =18/11/17- PSA=0,61 = 27/12/17- PSA=0,74 =05.04.18- PSA=0,03 =30.04.18- psa=0,02 2- EM USO DE ABIRATERONA 1000MG/D; PREDNISONA 5MG 12/12HS; (dia 10/01/18) LECTRUM (FEZ NOV. 2017)	Plano Terapêutico atual	DENOSUMAB	Protocolo
11	9	1- ADENOCARCINOMA DE PROSTATA EC-IV- OSSO, LINFONODOS ABDOMINAIS tratamento prévio 23/06/16- PROSTATECTOMIA ROBOTICA. PSA PRE OP=4,9, RECEBEU ZOLADEZ EM MAIO 2016 28/12/17 A 04/01/18, RADIOTERAPIA PALIATIVA EM COXO- FEMURAL ESQUERDA- DOSE DE 20GY nov. 2017- LECTRUM =18/11/17- PSA=0,61 = 27/12/17- PSA=0,74 =05.04.18- PSA=0,03 =30.04.18- psa=0,02 2- EM USO DE ABIRATERONA 1000MG/D; PREDNISONA 5MG 12/12HS; (dia 10/01/18) LECTRUM (FEZ NOV. 2017)	Plano Terapêutico atual	ZOMETA	Protocolo

Figure 15 – An Excel file sample with corpus generated by the per-clinical-event loading


```
{
  "MedicalNotes": [
    {
      "Note": "# Trombose MMII bilateral pré diagnóstico\n# PET_CT : Lesão de 6,8x4,5x5,0cm em rim direito SUV 12,2 sinais de
      invasão e trombose neoplásica de veia renal E e VCI SUV 15\nLinfonodo retroaórtico ( SUV 7,5) 2,4cm / \n# Toxicidade
      limitante >\n# 9 x  Keytruda\nBoa tolerância a Qt\nMelhora do estado geral.\nHb 12,20 Leuco 4100 Plaq 251 mil Cr 1,4 UR
      39 Glic 104 TGO 37 TGP 22 FA 100 Alb 3,7         DHL 329       Ca 8,9  TSH2,35\nVit D 25\nRNM abdome 26/10/2017:
      Linfonodomegalia aortocaval 3,2cm ( doença estável) 21/02/18 doença estável. 13/09 redução para 2,7cm ( 15%)\n07/02/18
      Tc de tórax: normal\nHipocorado +, hidratado, eupineico\nRCR 2 T BNF sem sopros\nMV + ARA\nFlácido, indolor, sem
      VMG\nedema de mmii +++\nCa de rim EIV\nLibero ciclo 10 Keytruda\nKeytruda\nC 10 d1: Foi preparado 200 mg de Keytruda em
      100 ml de SF 0,9%. Foi entregue Eritromax 40000 UI para aplicação posterior.",
      "Arguments": [
        {
          "name": "Diagnostico",
          "value": "C64 - Rim"
        },
        {
          "name": "Protocolo",
          "value": "Eritropoetina"
        },
        {
          "name": "Protocolo",
          "value": "Gemzar"
        },
        {
          "name": "Protocolo",
          "value": "Keytruda"
        }
      ]
    }
  ]
}
```


Figure 16 – A JSON file sample with corpus generated by the per-patient loading

have more than one label.

The number of JSON files varies according to the number of patients each database had. The small size database generated 397 files, the medium size did 5,398 files, and the large generated 39,326 JSON files. The reason for using JSON files instead of Excel files is that the latter has a limit of 32,767 characters per cell, and the large text with all patient's clinical events frequently exceeds that limit.

4.4 Model

In this section, the details of the model used for the experiments are presented. The Machine Learning and Deep Learning architectures will be described, along with some aspects of the hyperparametrization, some complementary information of the feature engineering, and the general view of the set of experiments planned. In Figure 17 a model's overview is presented.

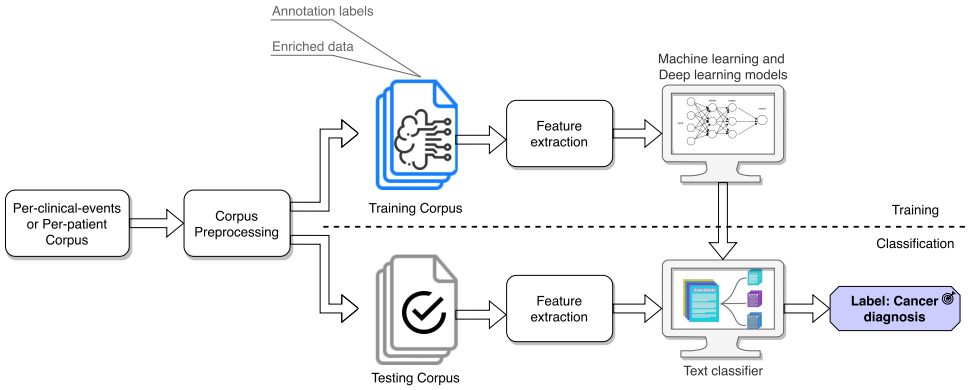

Figure 17 – Overview of the model applied in this research

The additional corpora preprocessing, the feature extraction details, and the model specifications are described in the following subsections.

4.4.1 Corpus preprocessing

After the corpora creation process, it was necessary to preprocess them to normalize the text. The same preprocessing was performed for both per-clinical-event and per-patient corpora type. The text analysis and preprocessing techniques applied are described below.

Before the text preprocessing, a diagnosis histogram was generated as seen in Figure 18. It was possible to observe that a small group of diagnoses concentrates the most frequent occurrences. Hence the diagnoses with less than 50 occurrences were joined into a single group called "Outros" ("Others"). Furthermore, to evaluate the neural network's performance according to the dataset sparsity, a new version of the dataset was created with the 12 most occurring diagnoses, as seen in Figure 19.

The following tasks were performed:

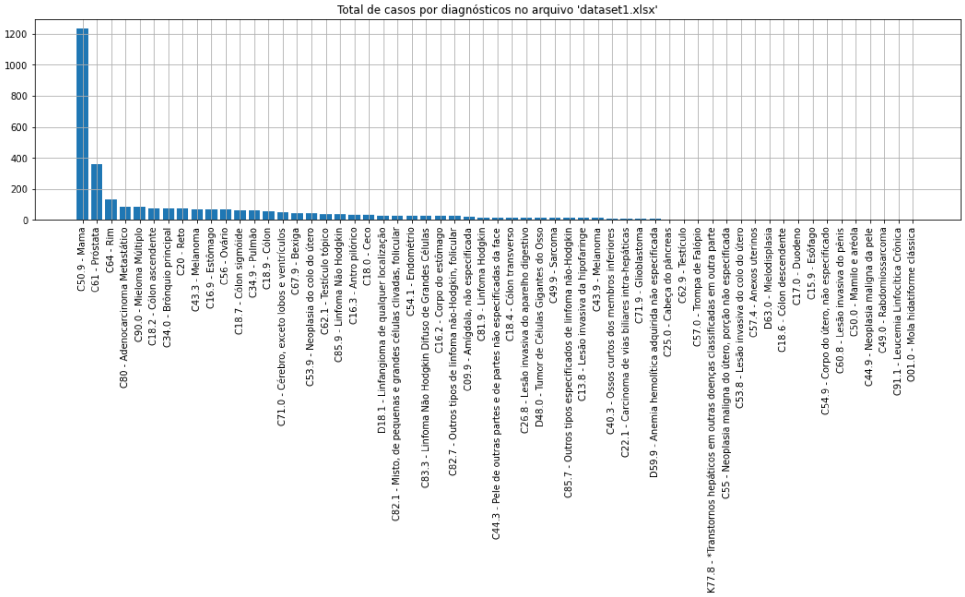

Figure 18 – Diagnosis histogram example with the total diagnoses occurrences by ICD code

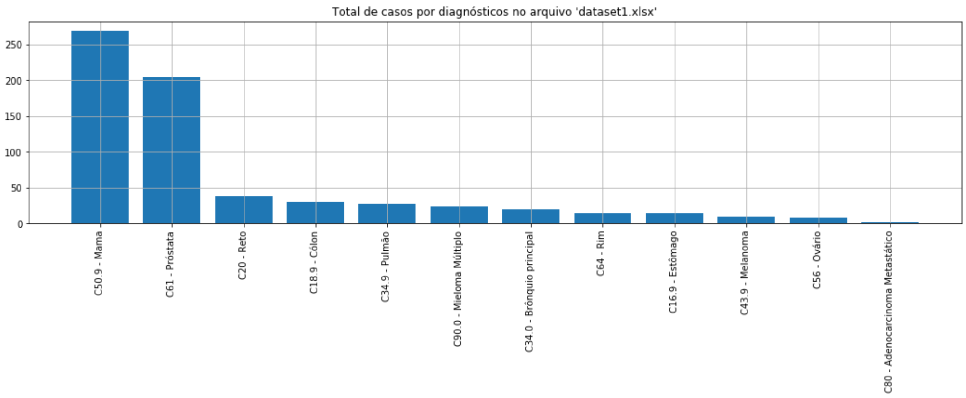

Figure 19 – Most occurring diagnoses.

- Tokenization: split the text into tokens that correspond to words;
- Stop-words filtering: removal of most common words in the Brazilian Portuguese language, punctuation, and special characters;
- Case folding: conversion of all words to lowercase.

In the per-clinical-events corpora, an additional manual analysis was done on the corpus. The text was assessed to understand how it could be transformed to improve ANN algorithms. Repeated medical notes were removed in several annotation labels. These notes could weaken the representation strength of the corpus.

An additional experiment was performed to evaluate how this step leveraged the results of the classifiers. As described in Section 5.1, a significant improvement was achieved with the application of this step.

4.4.2 Feature extraction

The text from medical notes must be transformed into a structure that could be used by the classifiers. For that reason, the Bag-of-Words (BoW) method was used, which is a representation that turns arbitrary text into fixed-length vectors by counting how many times each word appears. This representation is useful to be used by the classifiers algorithms.

This work used the medical notes bag-of-words (BoW) to extract the features to be used in the machine learning and deep learning training. In the per-clinical-event corpora, the BoW was generated for each medical note. Likewise, in the per-patient corpora, it was generated for each patient with all their medical notes.

Before creating the medical notes BoW, the text was normalized as described in the preprocessing Section 4.4.1. This preprocessing step aims to reduce the number of useless words, special characters, and punctuations, which would not make a difference in the classifiers models training. It also helped to reduce the computational effort to create the BoW.

The BoW applied to the medical notes resulted in a sparse representation, i.e., the vector sequence of numbers representing each word contained too many zeros. The Principal Component Analysis (PCA) technique was applied to reduce the data sparsity. The PCA technique converts a set of observations of possibly correlated features into a set of values of linearly uncorrelated features. The PCA with 500 features was used.

4.4.3 Machine learning and deep learning architectures

For the text classification task, this work applied machine learning and deep learning methods, comparing their results. Several machine learning classification algorithms were applied to evaluate which had the best performance. Furthermore, an LSTM deep learning algorithm was also applied to compare between traditional machine learning and a deep learning recurrent neural network.

The following Machine Learning algorithms were evaluated: Multilayer Perceptron (MLP) neural network; Logistic Regression; Decision Tree classifier; Random Forest classifier; Extra Trees classifier; K-nearest Neighbors (KNN) classifier. Furthermore, an Long-short Term Memory (LSTM) deep learning experiment was also performed.

The datasets were divided into two groups to train and test the neural networks: one with 80% of the data to train the models, and the other with 20% of the data to test the models. The data were shuffled to keep the categories' proportion, and then they were divided as aforementioned.

The machine learning algorithms were implemented using scikit-learn[1], and the deep learning LSTM was performed using Keras library[2], both were implemented in Python. In the first

[1] `https://scikit-learn.org/`
[2] `https://keras.io/`

set of experiments, seven tests were performed, with the following architecture details. An MLP with one hidden layer with 500 neurons; an MLP with two hidden layers with 800 and 500 neurons; a Logistic Regression classifier; a Decision Tree with a maximum of twenty levels and three samples by leaf; a Random Forest with a maximum of twenty levels and three samples by leaf; an Extra Trees with a maximum of twenty levels and three samples by leaf; a KNN classifier with a unitary K.

The second set of experiments was performed, this time with the 12 most occurring diagnoses in the dataset, as seen in Figure 19. For this experiment, the machine learning that best performed was selected to compare with an LSTM deep learning recurrent neural network. The machine learning algorithm had the following architecture: an MLP with two hidden layers with 800 and 500 neurons. The deep learning algorithm had the following architecture: an LSTM with the library Keras built on top of Tensorflow in a python implementation, in which the parametrization used was composed of Batch size 128, Dropout rate of 0.2, validation split of 0.2, Optimizer with adam, Loss measure was categorical cross entropy. Also, to prevent overfitting, EarlyStopping was used. In these experiments, since the main focus was dedicated to the general application, standard values for the parameters were used, as described in the literature. All the described models were evaluated using standard metrics, indicated in the literature, such as accuracy and both macro and weighted F1 score.

5 EXPERIMENTS AND RESULTS

In this chapter, the results of the experiments outlined in the previous chapter are described and evaluated.

Several machine learning classifiers and a deep learning recurrent neural network were applied in this work's experiments. The main objective of these experiments was to address the main research question and identify a possible workflow to use the dataset and text classification algorithms to identify potential support for healthcare professionals.

The dataset used was composed of an arrangement of the available options, using both the per-clinical-event and the per-patient versions. In the first step, the complete dataset with several machine learning text classification algorithms was used. In this experiment, both datasets (per-clinical-event and per-patient) were used. In the second step a new experiment was carried out involving the per-patient dataset and the algorithms MLP and LSTM. The per-patient dataset was chosen in the second step because all patient's clinical notes were joined into a single record, which would perform better considering the LSTM ability to process entire sequences of data.

Therefore, two main experiments were performed: a) Machine learning - several machine learning classifiers were experimented and their performance compared (described in Section 5.1); b) Deep learning - an experiment with a deep learning recurrent neural network was performed (described in Section 5.2).

5.1 Machine learning experiments

The per-clinical-event corpus (described in Section 4.4.1) of the smallest clinic in the preprocessed and raw versions were used to perform the classifiers. This is the clinic database described in the column "Small" in Table 2 and contains 3.308 clinical notes, and 397 distinct patients. The preprocessed dataset was used first with the machine learning classifiers, described in Section 4.4.3.

Method	Mean Accuracy	Macro F1 score	Weighted F1 score
MLP 1 (1 hidden layer, 500 neurons)	84.89%	84.21%	84.99%
MLP 2 (2 hidden layers, 800 and 500 neurons)	87.62%	87.44%	87.70%
Logistic regression	84.89%	82.75%	84.75%
Decision tree	71.86%	63.95%	71.98%
Random forest	80.23%	76.09%	79.53%
Extra trees	78.46%	76.71%	78.03%
KNN classifier	85.05%	83.93%	85.20%

Table 3 – Machine learning classifiers' experiments results.

To perform the experiments, the dataset was randomly divided into two parts: 80% for training and 20% for testing. A shuffle method was used to generate these two parts, and for

each time it was performed, a different set of training and testing datasets were created. Hence different classifiers' metrics were obtained, but it always kept the performance order.

The mean accuracy, Macro F1 score, and Weighted F1 score of each classifier are presented in Table 3. These experiments were performed to evaluate which machine learning classifier had the best performance. According to Table 3, the MLP 2 classifier achieved the best accuracy, Macro F1, and Weighted F1 scores. These results are evaluated in Section 5.3.

An additional experiment was performed to evaluate how the dataset's structure and preprocessing step (Section 4.4.1) leveraged the classifiers' performance. This experiment used the same dataset from the Clinic identified as "Small" in Table 2. The preprocessed and raw versions of the per-clinical-event dataset, plus the preprocessed per-patient dataset, were used with the best performance classifier. According to Table 3, the MLP 2 classifier had the best performance, and it was used in this experiment.

Table 4 presents the mean accuracy of the MLP 2 classifier with preprocessed and raw versions of the per-clinical-event dataset, plus the preprocessed per-patient dataset.

Dataset	Mean accuracy
Per-clinical-event raw dataset	26.1%
Per-clinical-event preprocessed dataset	86.7%
Per-patient preprocessed dataset	93.9%

Table 4 – Comparison of the MLP 2 classifier's performance with the per-clinical-event dataset in raw and preprocessed versions, plus the per-patient preprocessed dataset.

These results are evaluated in Section 5.3, exploring the improvements obtained with the integration of the clinical events in a more complete view of the patient history. As it can be observed in Table 4, the mean accuracy improves with more complete patient data.

5.2 Deep learning experiments

In this set of experiments, the following machine learning and deep learning classifiers were tested:

- The machine learning classifier that best performed (the MLP 2), according to Section 5.1;
- An LSTM (Long-short term memory) deep learning recurrent neural network.

The per-patient corpus (described in Section 4.4.1) of the smallest clinic in the preprocessed version was used to perform the classifiers. The per-patient corpus was chosen because all patient's clinical notes were joined into a single registry, which would perform better considering the LSTM ability to process entire sequences of data (according to Section 2.6.1).

Table 5 presents the mean accuracy, Macro F1 score and Weighted F1 score of the MLP 2 and the LSTM classifiers. These results are evaluated in Section 5.3.

Method	Mean Accuracy	Macro F1 score	Weighted F1 score
MLP 2 (2 hidden layers, 800 and 500 neurons)	93.90%	93.61%	93.99%
LSTM	84.81%	84.57%	84.93%

Table 5 – MLP 2 and LSTM classifiers performed with its performance.

5.3 Results evaluation

Several experiments were performed to understand the behavior of the selected machine learning and deep learning classifiers with the corpora created (Section 4.3) and preprocessed (Section 4.4.1).

First, a set of experiments with seven machine learning classifiers were performed with the per-clinical-event corpus, according to Section 5.1. Considering the mean accuracy, Macro F1, and Weighted F1 scores, the classifier that best performed was the MLP 2, as seen in Figure 20.

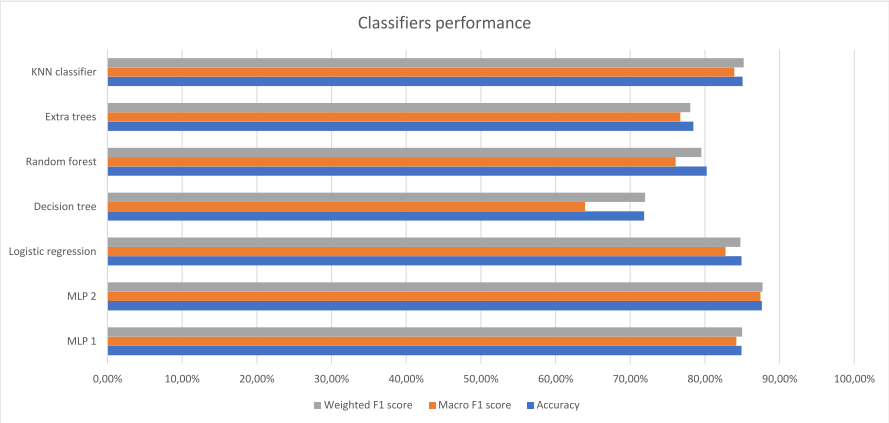

Figure 20 – Machine learning classifiers' performance chart.

The preparation of the corpus to be used in machine learning and deep learning classifiers was an important step. An experiment was performed with the MLP 2 classifier, the preprocessed and raw versions of the per-clinical-event dataset, and the preprocessed per-patient dataset (as described in Section 4.4.1). The performance of each is presented in Table 4. This table shows a significant improvement of the MLP 2 classifier performance with the preprocessing of the dataset (Section 4.4.1), in both per-clinical-event and per-patient datasets. Furthermore, the per-patient corpus performed better than the per-clinical-events corpus. For that reason, the next experiments used the per-patient corpus.

After the evaluation of the machine learning classifier that best performed, the MLP 2 classifier was selected. A new experiment was performed, to compare the MLP 2 classifier with an LSTM recurrence neural network, using the preprocessed per-patient corpus (as described in Section 5.2). As seen in Figure 21, the MLP 2 performed better than the LSTM classifier, even though the latter is a more recent neural network. The reason for this result can be associated with the fact that this experiment used the smallest per-patient corpus, and deep learning

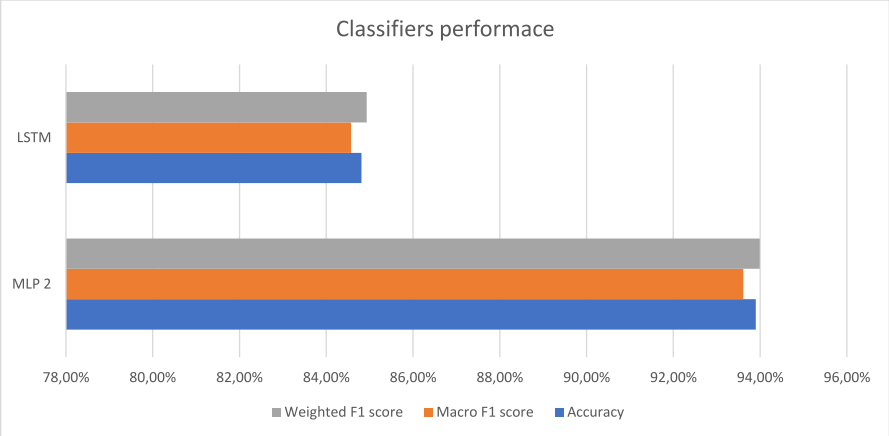

Figure 21 – The performance chart of the machine learning classifier and deep learning classifier.

algorithms perform better on large datasets.

6 CONCLUSION

With a large amount of information generated every day in healthcare, it is not suitable for humans to process it manually. Furthermore, a big part of this information is recorded as unstructured data, transforming it into a hard-working task. It is necessary to develop tools to help healthcare professionals to deal with it, automate the classification and extraction of information from medical notes, and enable this information to be processed by machines. This context motivates this research, mainly dedicated to the study and analysis of the real-world problems, data, and processes to be addressed.

The research question proposed by this work was "What are the text classification methods that better support the healthcare professional on diagnosis decisions, based on the patient's oncology clinical history?". To address the answer for this question, this work surveyed text classification methods and applied several machine learning and deep learning classifiers experiments on an oncological medical notes corpus. The corpus was created by this work, based on non-synthetic oncological medical notes from an Oncology EHR system database. This corpus creation process allowed the identification of a variety of preprocessing steps (as described in Section 4.4.1), and specific term treatment to improve overall results. The experiments achieved good accuracy, especially an MLP machine learning and an LSTM deep learning methods, showing possibilities of using these resources for medical notes text classification. Therefore, this set of tasks provided strong support indicating resources and processes to text classification in the specific context of health professional support.

The first objective of this work is related to the production of a de-identified corpus with non-synthetic medical notes from the oncology area, in the Brazilian Portuguese language. This is important because, according to the evaluation of the related works (Chapter 3), most of the research on NLP in the healthcare area is performed using corpora in the English language. Considering this specific objective, this work created its corpora in the Brazilian Portuguese language, based on thousands of medical notes from a real-world oncology clinic database. They were preprocessed to improve the classifiers' performance. As evaluated in Section 5.1, the best machine learning's performance was considerably improved due to the preprocessing step. However, other NLP techniques could also be applied and leverage this outcome.

The second and third objectives aimed to perform experiments with text classification approaches and to compare these approaches. To accomplish these objectives, this work implemented and compared six different machine learning classifiers and an LSTM deep learning classifier. The machine learning and deep learning classifiers achieved similar accuracy to the same classifiers in the related works, and it could be leveraged with tuning and applying improved versions of the classifiers. Although the best machine learning classifier performed better than the deep learning classifier (according to Section 5.3), the latter should perform better with a larger dataset and with an improved LSTM version. Therefore, this classifiers comparison shows that some traditional machine learning classifiers perform with similar results to

modern methods in some cases. It shows that it is possible to use simpler algorithms, especially when a small corpus is used.

The last specific objective aimed to evaluate the application of text classification in an Oncology EHR system. This objective was presented in this research's overall conduction, considering the necessity of in-depth and detailed analysis of an EHR system, with all the related context, involving from dataset and forms to healthcare professional routines. Therefore, an in-depth observation of healthcare professional needs and unstructured data details was produced. One of the published articles from this work (SCHWERTNER et al., 2019) described some of the insights about these possibilities of using text classification (and other complementary approaches) to foster the use of medical notes in an oncology EHR system.

Artificial Intelligence, in special machine learning and deep learning algorithms, has been widely applied in several industries, sometimes surpassing the human accuracy. In the healthcare industry, several processes can be improved by AI, leveraging healthcare professionals. In the oncology area, the diagnosis and treatment decision-making is one of these complex processes that can be aided by AI algorithms. Furthermore, it is also essential to apply this research in real-world applications to support healthcare professionals in their daily routine. The application of AI can help healthcare professionals care more for people and less for machines, which can improve these professionals' assistance to their patients.

Some limitations of this work are considered, which can be addressed in the future. One of these is regarding word embeddings. An essential step that almost all related works performed was the text transformation to a vector representation. According to Section 4.4.2, this work applied the Bag-of-Words method, but improved versions or different embedding methods could improve the results, such as the TF-IDF (Term Frequency-Inverse Document Frequency) with the Bag-of-Words or the word embeddings. According to the related works in Chapter 3, word embeddings were widely used with deep learning methods. However, in most cases, a general-purpose word embedding was used. A few papers described the creation of its word embedding, but it requires a high computational effort to be built. The limited set of experiments that do not consider all the available datasets at this moment could also be referred as another limitation of this work, restricting the analysis to the smaller dataset.

Considering this work's development, the following list of future steps is suggested:

- Create larger corpus with the medical notes from several oncology clinic databases and conduct new experiments;

- Enhance the corpus preprocessing step by removing low-frequency words, spell checker, replacing acronyms and abbreviations by its standard word;

- Create a domain-specific word embedding from the corpora and apply it with the classifiers. Another option is to use general-purpose word embeddings and fine-tune it with the corpora;

- Improve the enrichment process with more structured data available from the Oncology EHR system, such as prescribed medications, patient's problems, and allergies;
- Tune the implemented classifiers and try different versions, such as a Bidirectional LSTM (Bi-LSTM);
- Integrate the implemented classifiers with the Oncology EHR system to obtain feedback from the healthcare professionals about their accuracy, and suggest the diagnosis based on the patient's clinical history.

6.1 Contributions

In the first part of this work, the author published and presented two papers related to the main focus of this research which served as the basis for this work. The first paper was a systematic review of clinical decision support systems and information extraction in oncology (SCHWERTNER; RIGO, 2018), presented as a full paper in the CBIS 2018 - XVI Congresso Brasileiro de Informática em Saúde, on October 1st to 4th of 2018, in Fortaleza, Brasil. The paper was published in the proceedings of the congress.

The second paper was an experiment developed in collaboration with a doctoral degree student describing the overall vision of the health professional support (SCHWERTNER et al., 2019). This experiment integrated three main resources: an information extraction procedure, a knowledge base, and a question answering system. It was submitted and accepted in the 32nd IEEE CBMS International Symposium on Computer-Based Medical Systems and was presented by the author in Cordoba, Spain, on June 5th to 7th of 2019.

Besides the published work, the following contributions considered in this research so far are highlighted:

- Use of a corpus with non-synthetic medical notes from the oncology area in the Brazilian Portuguese language. The corpus was created from the InterProcess Gemed Oncology EHR system's database (INTERPROCESS, 2019), which contains daily medical notes from several healthcare professionals, such as physicians, nurses, pharmacists, psychologists, nutritionists, and physiotherapists, dentists, social service professionals, receptionists, and others. These were real data from clinical events (like consultations) registered by these professionals;
- Implementation of a de-identified approach that made the use of a non-synthetic dataset possible, based on an unrecoverable pseudonymization process, making it harder to identify the patient or professional directly and at the same time keeping the clinical history longitudinal consistency;
- Description and study of a data enrichment process allowing the application of knowledge from the health field and the EHR system to provide necessary annotation attributed for

the dataset in an automated approach. This enrichment process takes advantage of the knowledge inputted by the system users, who are experts in the healthcare field;

- Evaluation of several approaches of text classification with machine learning and deep learning algorithms;
- Comparison of the performance of machine learning and deep learning methods in text classification;
- Evaluation of the possibilities of application of text classification in an Oncology EHR system, such as getting feedback from expert users about the diagnosis suggestions accuracy and helping them in the oncology diagnosis decision-making process.

This work is also a partnership between Unisinos University and InterProcess company, contributing to technology transferring between university and industry, and providing the means to apply technology developed in a real-world case. This partnership is to be continued in a Doctoral work, already being developed, using funds from the "Doutorado para Inovação" initiative from CNPQ.

REFERENCES

AGGARWAL, C. C. et al. **Neural networks and deep learning**. [S.l.]: Springer, 2018.

ALEMZADEH, H.; DEVARAKONDA, M. An nlp-based cognitive system for disease status identification in electronic health records. In: IEEE EMBS INTERNATIONAL CONFERENCE ON BIOMEDICAL HEALTH INFORMATICS (BHI), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p. 89–92.

BANERJEE, I. et al. Automatic inference of bi-rads final assessment categories from narrative mammography report findings. **Journal of Biomedical Informatics**, [S.l.], v. 92, p. 103137, 2019.

BARASH, Y. et al. Comparison of deep learning models for natural language processing-based classification of non-english head ct reports. **Neuroradiology**, [S.l.], p. 1–10, 2020.

BUCUR, A. et al. Clinical decision support framework for validation of multiscale models and personalization of treatment in oncology. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING, 13., 2013. **Anais...** [S.l.: s.n.], 2013. p. 1–4.

CARRODEGUAS, E. et al. Use of machine learning to identify follow-up recommendations in radiology reports. **Journal of the American College of Radiology**, [S.l.], v. 16, n. 3, p. 336 – 343, 2019.

CHEN, X. et al. A bibliometric analysis of natural language processing in medical research. **BMC Medical Informatics and Decision Making**, [S.l.], v. 18, n. 1, p. 14, Mar 2018.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support? **Journal of biomedical informatics**, [S.l.], v. 42, n. 5, p. 760–772, 2009.

DENG, L.; LIU, Y. **Deep learning in natural language processing**. [S.l.]: Springer, 2018.

DEV, S. et al. Automated classification of adverse events in pharmacovigilance. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p. 905–909.

FERNANDES, A. S. et al. p-health in breast oncology: a framework for predictive and participatory e-systems. In: SECOND INTERNATIONAL CONFERENCE ON DEVELOPMENTS IN ESYSTEMS ENGINEERING, 2009., 2009. **Anais...** [S.l.: s.n.], 2009. p. 123–129.

GLATZER, M. et al. Decision making criteria in oncology. **Oncology**, [S.l.], v. 98, n. 6, p. 39–47, 2020.

GUPTA, A.; BANERJEE, I.; RUBIN, D. L. Automatic information extraction from unstructured mammography reports using distributed semantics. **Journal of Biomedical Informatics**, [S.l.], v. 78, p. 78 – 86, 2018.

Health informatics - electronic health record - definition, scope and context - iso 20514. Rio de Janeiro, BR: Associação Brasileira de Normas Técnicas (ABNT) - International Organization for Standardization (ISO), 2008. Standard.

Health informatics - pseudonymization - iso 25237. Rio de Janeiro, BR: Associação Brasileira de Normas Técnicas (ABNT) - International Organization for Standardization (ISO), 2020. Standard.

HEALTH, N. I. of et al. Electronic health records overview. **National Center for Research Resources. National Institutes of Health, Bethesda**, [S.l.], 2006.

HODGES, S. **The counseling practicum and internship manual**: a resource for graduate counseling programs. [S.l.]: New York, NY: Springer Publishing Company, 2011.

HUNT, D. et al. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. **JAMA: Journal of the American Medical Association**, [S.l.], v. 280, n. 15, p. 1339 – 1346, 1998.

HUNT, D. L. et al. Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient OutcomesA Systematic Review. **JAMA**, [S.l.], v. 280, n. 15, p. 1339–1346, 10 1998.

INTERPROCESS. **Interprocess gemed oncology - oncological management system**. Accessed: 2020-04-05, `http://www.interprocess.com.br/en/gemed-oncology/`.

KREIMEYER, K. et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. **Journal of Biomedical Informatics**, [S.l.], v. 73, p. 14 – 29, 2017.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, [S.l.], v. 521, n. 7553, p. 436, 2015.

LI, P.; MAO, K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. **Expert Systems with Applications**, [S.l.], v. 115, p. 512 – 523, 2019.

MESKÓ, B. **The guide to the future of medicine**: technology and the human touch. [S.l.]: Webicina Kft., 2014.

MOEN, H. et al. Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. **Journal of the American Medical Informatics Association**, [S.l.], v. 27, n. 1, p. 81–88, 2020.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2018.

NEIVA, F.; SILVA, R. Revisão sistemática da literatura em ciência da computação - um guia prático. , [S.l.], 06 2016.

NGUYEN, L.; BELLUCCI, E.; NGUYEN, L. T. Electronic health records implementation: an evaluation of information system impact and contingency factors. **International Journal of Medical Informatics**, [S.l.], v. 83, n. 11, p. 779 – 796, 2014.

OLIVA, J. T. et al. A computational system based on ontologies to automate the mapping process of medical reports into structured databases. **Expert Systems with Applications**, [S.l.], v. 115, p. 37 – 56, 2019.

POLPINIJ, J. The cancerology ontology: designed to support the search of evidence-based oncology from biomedical literatures. In: INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS (CBMS), 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p. 1–6.

Qiu, J. X. et al. Deep learning for automated extraction of primary sites from cancer pathology reports. **IEEE Journal of Biomedical and Health Informatics**, [S.l.], v. 22, n. 1, p. 244–251, Jan 2018.

REYES-ORTIZ, J. A.; GONZÁLEZ-BELTRÁN, B. A.; GALLARDO-LÓPEZ, L. Clinical decision support systems: a survey of nlp-based approaches from unstructured data. In: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS (DEXA), 2015., 2015. **Anais...** [S.l.: s.n.], 2015. p. 163–167.

SABRA, S. et al. Performance evaluation for semantic-based risk factors extraction from clinical narratives. In: IEEE 8TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE (CCWC), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 695–701.

SCHULMEISTER, L. Technology and the transformation of oncology care. **Seminars in Oncology Nursing**, [S.l.], v. 32, n. 2, p. 99 – 109, 2016. The Transformation of Health Care: Nurses as Partners in Change.

SCHWERTNER, M. A. et al. Fostering natural language question answering over knowledge bases in oncology ehr. **32nd IEEE CBMS International Symposium on Computer-Based Medical Systems**, [S.l.], 2019. Available at `http://www.cbms2019.org/`.

SCHWERTNER, M. A.; RIGO, S. J. Sistemas de apoio à decisão clínica e extração de informação na oncologia. **XVI Congresso Brasileiro de Informática em Saúde - CBIS 2018**, [S.l.], 2018. Available at `http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/issue/view/86`.

SHAO, Y. et al. Clinical text classification with word embedding features vs. bag-of-words features. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 2874–2878.

SHICKEL, B. et al. Deep ehr: a survey of recent advances on deep learning techniques for electronic health record (ehr) analysis. **CoRR**, [S.l.], v. abs/1706.03446, 2017.

SI, Y. et al. Enhancing clinical concept extraction with contextual embedding. **CoRR**, [S.l.], v. abs/1902.08691, 2019.

SOCIETY, M. P. **Medical records**. Available at `https://www.medicalprotection.org/uk/articles/eng-medical-records`.

WANG, S.-m. et al. Using deep learning for automatic icd-10 classification from free-text data. **European Journal of Biomedical Informatics**, [S.l.], v. 16, n. 1, 2020.

WANG, Y. et al. Clinical information extraction applications: a literature review. **Journal of Biomedical Informatics**, [S.l.], v. 77, p. 34 – 49, 2018.

WANG, Y. et al. A clinical text classification paradigm using weak supervision and deep representation. **BMC Medical Informatics and Decision Making**, [S.l.], v. 19, n. 1, p. 1, Jan 2019.

ZHANG, R. et al. Automatic methods to extract new york heart association classification from clinical notes. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p. 1296–1299.

APPENDIX A – CDSS AND INFORMATION EXTRACTION SURVEY

To better understand the recent scenario of Clinical Decision Support Systems (CDSS), a survey was conducted in a systematic review format (SCHWERTNER; RIGO, 2018). The objective of this survey was to understand which artificial intelligence techniques are used to support CDSS in oncology is presented. CDSS need access to patient's health data, which are provided in the following formats (REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015):

- Structured data as electronic health records;
- Semi-structured data as XML documents or two columns of laboratory results;
- Unstructured data like narrative text, patient's clinical observations, radiology reports, and operative notes.

Due to this variety of input data format, specifically unstructured data in free text format, this survey was extended to discover what techniques have been used to extract information from clinical notes with natural language processing (NLP) techniques.

A.1 Systematic review methodology

This work followed the "Revisão Sistemática da Literatura em Ciência da Computação" (NEIVA; SILVA, 2016) guide for the performance of a systematic review. The main topics were Clinical Decision Support Systems on the patient's oncology diagnosis, and also information extraction from clinical notes with natural language processing.

The following steps were performed: searching questions definition, keywords definition, researched databases query, papers classification and selection, and answers to the defined questions.

Regarding the main objective of this survey, the following main question was asked: **What are the artificial intelligence techniques that have been used to support the decision of the patient's diagnosis on oncology?**

While performing the survey and classifying the selected papers, the same problem was found in many papers: how hard it was to obtain the data in a structured format to be processed by CDSS (REYES-ORTIZ; GONZÁLEZ-BELTRÁN; GALLARDO-LÓPEZ, 2015; DEMNER-FUSHMAN; CHAPMAN; MCDONALD, 2009; SABRA et al., 2018; ALEMZADEH; DEVARAKONDA, 2017). Furthermore, the use of information extraction methods was also identified to obtain structured data from clinical notes. Therefore, a secondary objective was added by asking the following question: **What techniques have been used to extract information from clinical notes?**

The keywords were defined based on these questions. For the main question, the following keywords were defined: survey, clinical decision support system, artificial intelligence, oncology. And for the secondary question: survey, medical, oncology diagnosis, non-structured data, natural language processing, NLP.

Based on the above questions and keywords, the following search strings were defined:

- Main question:
  - (survey) AND (clinical decisions support system) AND (oncology)
  - (oncology) AND (artificial intelligence)
- Secondary question:
  - (survey) AND (nlp) AND (medical)
  - (oncology diagnosis) AND (non-structured data) AND (natural language processing)
  - (healthcare) AND (oncology diagnosis) AND (non-structured data) AND (natural language processing)

After the definition of the search string, the following research databases were queried:

- IEEE Xplore: `https://ieeexplore.ieee.org`
- Science Direct: `http://www.sciencedirect.com`
- Google Scholar: `https://scholar.google.com.br`

Google Scholar returned papers from other publishers. Therefore the papers obtained from the above-researched databases were published by the following publishers: ACM, arXiv.org, BMC, IEEEXplore, iJarcet, JAMIA, JMIR, PubMed, ResearchGate, SAGE, Science Direct, Semantic Scholar, Springer.

The next step was the papers' classification and selection. Two filters were applied to select the papers, with the following constraints:

- First filter:
  - Language: English;
  - Publication year: from 2009 to 2018;
  - The document should be in academic paper format;
- Second filter:
  - Main area of interest: healthcare;
  - Secondary area of interest: at least one of bellow:

* CDSS: if the paper uses clinical decision support systems to support the patient's diagnosis;
* NLP: if the paper uses natural language processing to extract information from clinical notes.

Seven searches were performed in the related databases, resulting in 369 papers. After applying the first filter, 53 papers were selected. Moreover, after the second filter applied, 39 papers were selected. A spreadsheet was created to help evaluate the papers. Its main characteristics were: publisher, publication year, how the paper answers the main and the secondary questions.

A.2 Evaluation of the papers

Most of the 39 papers were published in 2018 (7 papers), 2017 (7 papers), 2015 (5 papers), and 2013 (5 papers), according to Figure 22.

Figure 22 – Chart of the amount of papers per year

Furthermore, the main publishers were IEEEXplore (17 papers), Science Direct (5 papers), and JAMIA (4 papers) according to Table 6.

Regarding the main question, 18% (7 papers) answered it, as seen in Figure 23, and regarding the secondary question, 92% (36 papers) answered it, as seen in Figure 24. Some papers answered both the main and secondary questions. The high number of papers that answer the second question is due to the fact that many articles were about CDSS and the use of NLP to extract information in a structured format.

Regarding the area of interest, all selected papers were from the healthcare area, according to the applied constraints. However, only 23% (9 papers) were from the oncology area, as seen in Figure 25.

Amount of papers	Publication year										
Publisher	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
arXiv.org									1		1
BMC										3	3
IEEEXplore	1	3	1		1	1	3	1	5	1	17
iJarcet							1				1
JAMIA					4						4
JMIR										1	1
PubMed	1	1									2
ResearchGate								1			1
SAGE				1							1
Science Direct						1		1	1	2	5
Semantic Scholar						1	1				2
Springer						1					1
Total	2	4	1	1	5	4	5	3	7	7	39

Table 6 – Amount of paper distribution by year and publisher

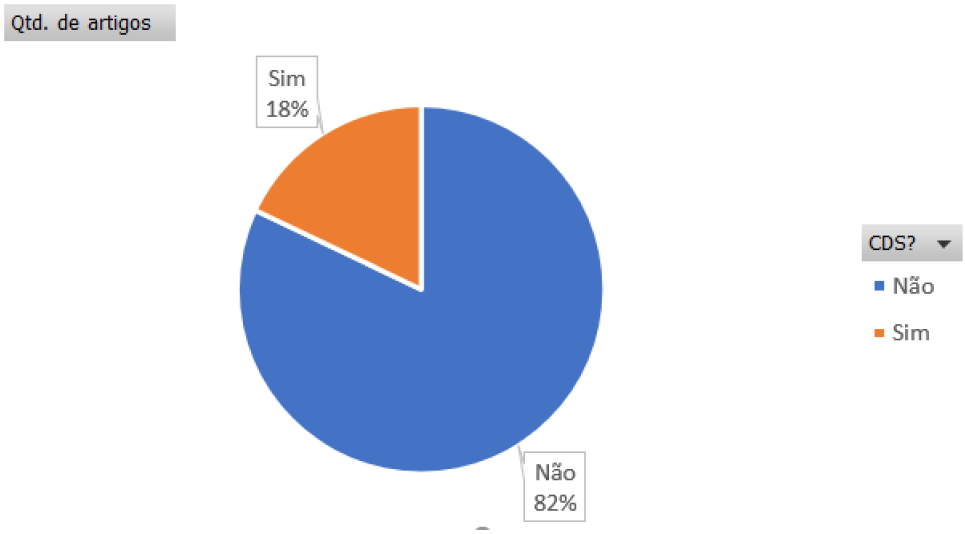

Figure 23 – Amount of papers that answer the main question

Regarding only the oncology area papers, 33% of them (3 papers) answered the main question (according to Figure 26), and 67% of them (6 papers) answered the secondary question (according to Figure 27).

A.3 Evaluation of the main question

Regarding the main question (*What are the artificial intelligence techniques that have been used to support the decision of the patient's diagnosis of oncology?*), seven papers provided answers, 5 of which presented CDSS' results and 2 of which presented CDSS' surveys.

The importance of obtaining the structured data to be used by CDSS in almost all papers was observed. Moreover, it was also observed the use of NLP techniques in these papers.

Among the papers which showed CDSS' results, the techniques applied were rule-based

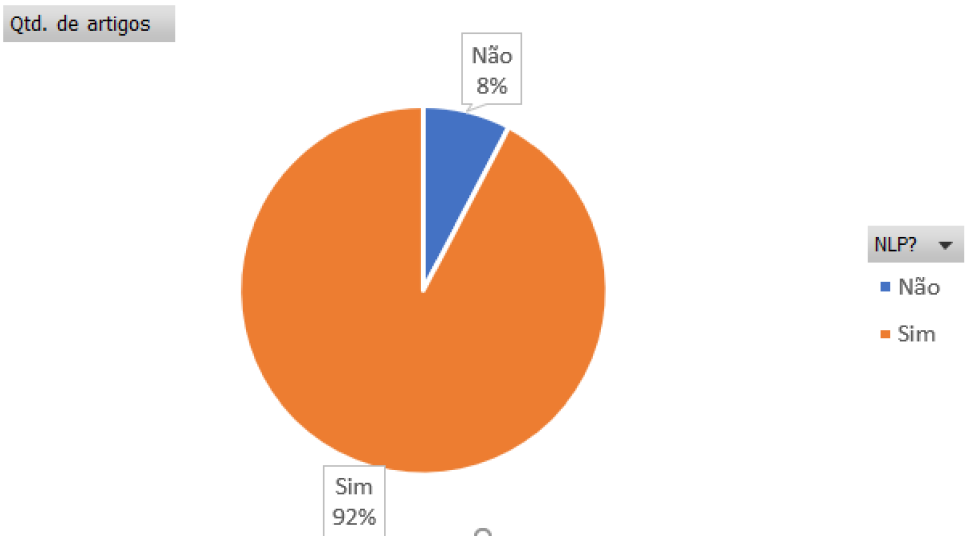

Figure 24 – Amount of papers that answer the secondary question

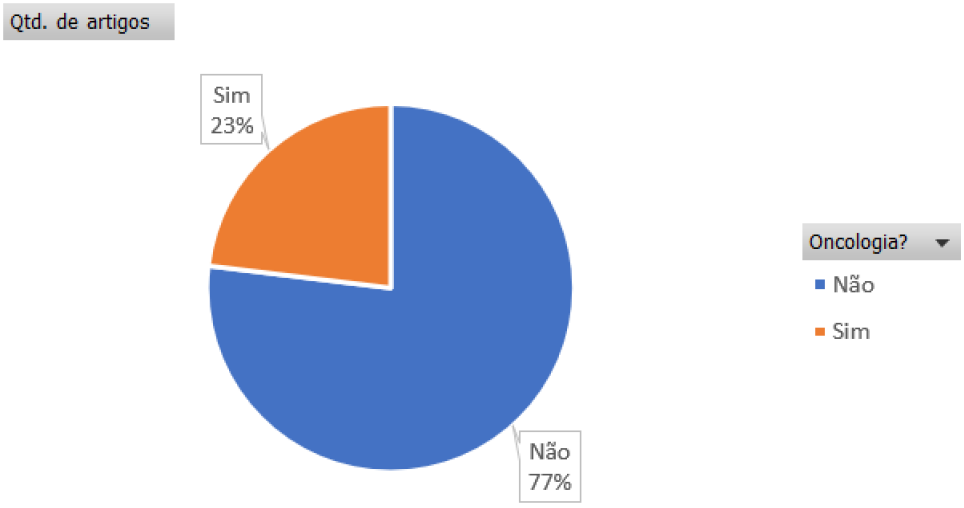

Figure 25 – Amount of papers regarding the oncology area

systems and artificial neural networks. Some papers that used a rule-based system presented its implementation, while the others used available artificial intelligence (AI) platforms.

Two papers presented a survey about the use of NLP in CDSS, one focusing on obtaining clinical data and the other on processing and support of the clinical decision.

A.4 Evaluation of the secondary question

Regarding the secondary question (*What techniques have been used to extract information from clinical notes?*), 36 papers answered it, 27 of which presented information extraction applications and 9 presented NLP's surveys.

Among the 27 papers which presented information extraction applications, various techniques of NLP and AI were applied: machine learning (SVM, Naive Bayes), syntax and lexicon

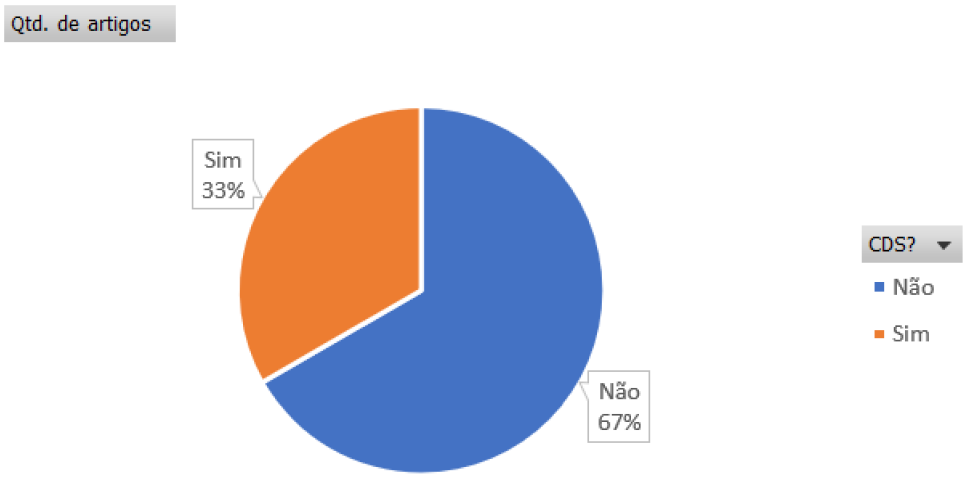

Figure 26 – Amount of papers regarding the oncology area which answered the main question

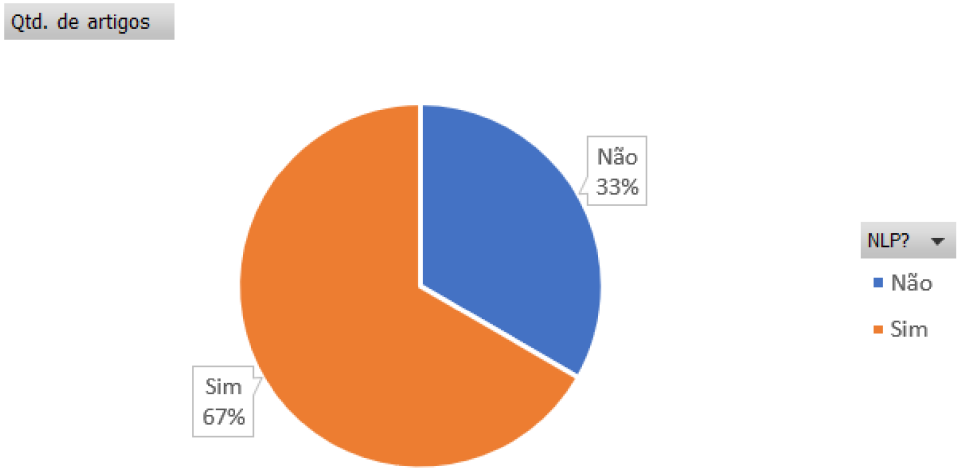

Figure 27 – Amount of papers regarding the oncology area which answered the secondary question

features (text tokenization, text normalization, part-of-speech, n-Gram), artificial neural networks, available frameworks, and platforms.

The other 9 papers presented NLP's surveys, with the following approaches: text mining, NLP applied to CDSS, NLP in general, and information extraction.

A.5 Conclusion

This survey presented a systematic review, with the main objective to review the use of AI techniques to support the decision of the patient's diagnosis of oncology. Its second objective was to learn what techniques have been used to extract information from clinical notes were approached.

Regarding the main objective, the main techniques applied were rule-based systems and artificial neural networks. Through the evaluation of these papers, almost all of them mentioned

how hard it was to obtain structured data. It was said that much data were stored as free-text clinical notes in an unstructured format. To address it, information extraction techniques were applied to these clinical notes.

With the challenge of obtaining structured data in mind, a second objective was defined. NLP techniques were observed in almost all papers: machine learning, syntax and lexicon features, artificial neural networks, available frameworks and platforms, new methods, and hybrid approaches. The use of these techniques changed according to the kind of information to be extracted, the language, and the research area.

This survey found that most papers were published in recent years, showing that there is still much research to be done in this area. Therefore, there are many opportunities for new research in information extraction applied to healthcare clinical notes.

APPENDIX B – INFORMATION EXTRACTION AND TEXT CLASSIFICATION SURVEY

The survey performed in Appendix A concluded the importance of obtaining structured data to work with CDSS. Another survey was performed with a focus on text classification and information extraction methods. This new survey had the objective to understand what methods have been recently applied to text classification and extract information from clinical notes using NLP, machine learning, or deep learning.

As a result of this survey, thirteen papers that focus on NLP, machine learning, and deep learning on text classification or information extraction were selected and studied. The following sections describe the survey, evaluate the papers, and conclude this survey.

B.1 Systematic review methodology

This work followed the "Revisão Sistemática da Literatura em Ciência da Computação" (NEIVA; SILVA, 2016) guide for the performance of a systematic review. The main topics were text classification and information extraction on the patient's oncology clinical notes with natural language processing. The following selection criteria were applied to select the relevant papers published for these topics:

- The following research repositories were selected: Science Direct, IEEE Xplore and Google Scholar;
- Only recent papers (with publication year from 2017 to 2020) were selected;
- The repositories were queried with the following search strings:
  - ("healthcare" or "oncology") and ("machine learning" or "deep learning") and ("text classification");
  - ("information extraction") and (("natural language processing" or "nlp") or ("machine learning") or ("deep learning"));
- Duplicated entries and papers written in languages other than English were removed. Some papers which described the same model and were published by the same authors were removed.

The above selection resulted in thirteen papers published by arXiv, BMC, EJBI, IEEExplore, JAMIA, ScienceDirect, Springer. The complete list of these papers is available in Table 7, and the bibliographic citation in "Citation" column will be used further in this work to reference each paper.

After the selection described above, thirteen papers were selected, all of which will be evaluated in the next section.

Title	Citation	Publisher	Year
Automatic inference of BI-RADS final assessment categories from narrative mammography report findings	Banerjee et al. (2019)	Science Direct	2019
Comparison of deep learning models for natural language processing-based classification of non-English head CT reports	Barash et al. (2020)	Springer	2020
Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports	Carrodeguas et al. (2019)	Science Direct	2019
Automated Classification of Adverse Events in Pharmacovigilance	Dev et al. (2017)	IEEExplore	2017
Automatic information extraction from unstructured mammography reports using distributed semantics	Gupta, Banerjee and Rubin (2018)	Science Direct	2018
Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts	Li and Mao (2019)	Science Direct	2018
Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods	Moen et al. (2020)	JAMIA	2020
A computational system based on ontologies to automate the mapping process of medical reports into structured databases	Oliva et al. (2019)	Science Direct	2018
Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports	Qiu et al. (2018)	IEEExplore	2018
Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features	Shao et al. (2018)	IEEExplore	2018
Enhancing Clinical Concept Extraction with Contextual Embedding	Si et al. (2019)	arXiv	2019
A clinical text classification paradigm using weak supervision and deep representation	Wang et al. (2019)	BMC	2019
Using Deep Learning for Automatic Icd-10 Classification from Free-Text Data	Wang et al. (2020)	EJBI	2020

Table 7 – List of the thirteen related work papers, their titles, citations, publishers, and publication years.

B.2 Evaluation of the papers

Different approaches have been used to classify or extract information from unstructured data, in most cases, from a free-text written in natural language. Some research used only traditional NLP methods along with medical experts (OLIVA et al., 2019); some used machine learning (BANERJEE et al., 2019; GUPTA; BANERJEE; RUBIN, 2018; SHAO et al., 2018); some used deep learning (LI; MAO, 2019; Qiu et al., 2018; SI et al., 2019; WANG et al., 2020); and others used both machine learning and deep learning approaches (CARRODEGUAS et al., 2019; BARASH et al., 2020; DEV et al., 2017; MOEN et al., 2020; WANG et al., 2019)). Although deep learning is the more recent and promising approach, traditional machine learning methods are still used with excellent results.

All selected papers preprocessed their corpora by using standard NLP tasks. Just three papers didn't describe the preprocessing applied (DEV et al., 2017; MOEN et al., 2020; WANG et al., 2019). Some used just a few basic methods, such as convert text to lowercase, stopwords and punctuation removal, duplicated elements removal, and tokenization (CARRODEGUAS et al., 2019; LI; MAO, 2019; Qiu et al., 2018; SHAO et al., 2018; SI et al., 2019; WANG et al., 2020). Other papers also used advanced methods like lemmatization, sentence splitting, part-of-

speech (POS), stemming, dependency parsing, and low-frequency words removal (BANERJEE et al., 2019; BARASH et al., 2020; GUPTA; BANERJEE; RUBIN, 2018; OLIVA et al., 2019). These tasks are essential to reduce the number of words to be processed and improve the performance of machine learning and deep learning methods.

Regarding the corpora, all papers used domain-specific corpora. In most cases, corpora were obtained from clinical reports or clinical notes stored in healthcare institutions (BANERJEE et al., 2019; BARASH et al., 2020; CARRODEGUAS et al., 2019; GUPTA; BANERJEE; RUBIN, 2018; MOEN et al., 2020; Qiu et al., 2018; SHAO et al., 2018; WANG et al., 2019, 2020). Some reports were de-identified because they were a result of patient care. De-identifying the reports and clinical notes used in the studies was a common concern in all papers, providing patient privacy, and avoiding patient identification from resulted data. Only one paper (OLIVA et al., 2019) used a corpus created by medical experts specifically to the study. Three papers (LI; MAO, 2019; SI et al., 2019; WANG et al., 2019) used corpora from open NLP challenges like i2b2 and SemEval, which provide a set of annotated and unannotated de-identified patient clinical data.

Furthermore, two papers (SI et al., 2019; Qiu et al., 2018) also used general-purpose corpora, providing a comparison between both approaches, i.e., which kind of corpora provided better performance. These corpora were obtained from general-purpose datasets like Gigaword 5, Wikipedia 2014 and 2017, BooksCorpus+ English Wikipedia, Google News articles, and PubMed biomedical publications. These papers applied the extraction information methods in both domain-specific and general-purpose corpora, comparing both performance results. Through the recent evolution of deep learning and unsupervised methods, they found that domain-specific corpora had better results than general-purpose corpora. Only one paper (OLIVA et al., 2019) used a corpus in the Brazilian Portuguese language. Most of the other papers used corpora in the English language, one in Finnish language and one in Hebrew language.

The corpora must be transformed to be used by machine learning and deep learning algorithms. They must be transformed from text to a vector structure. Seven papers (BARASH et al., 2020; CARRODEGUAS et al., 2019; DEV et al., 2017; GUPTA; BANERJEE; RUBIN, 2018; MOEN et al., 2020; Qiu et al., 2018; SHAO et al., 2018) used the Bag-of-Words (BoW) or VSM (Vector Space Model) embedding method to transform the corpora text into fixed-length vectors. All papers except two (BANERJEE et al., 2019; BARASH et al., 2020; CARRODEGUAS et al., 2019; DEV et al., 2017; GUPTA; BANERJEE; RUBIN, 2018; LI; MAO, 2019; MOEN et al., 2020; SHAO et al., 2018; SI et al., 2019; WANG et al., 2019, 2020)) used word embedding methods such as word2vec, doc2vec, GloVe, ELMo, and BERT. These embedding methods have dramatically advanced NLP tasks, especially when combined with deep learning-based models. Only one paper (OLIVA et al., 2019) did not use embedding methods because it did not use machine learning or deep learning methods.

Traditional machine learning or deep learning-based models were applied in almost all

papers, and sometimes both models were applied in the same paper. Some papers applied only machine learning-based models (BANERJEE et al., 2019; BARASH et al., 2020; CARRODEGUAS et al., 2019; DEV et al., 2017; GUPTA; BANERJEE; RUBIN, 2018; MOEN et al., 2020; SHAO et al., 2018; WANG et al., 2019), such as k-means clustering, Logistic Regression, Multi-Layer Perceptron, Random Forest, SVM. Other papers applied deep learning-based models (BARASH et al., 2020; CARRODEGUAS et al., 2019; DEV et al., 2017; LI; MAO, 2019; MOEN et al., 2020; Qiu et al., 2018; SI et al., 2019; WANG et al., 2019, 2020), using LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and CNN (Convolutional Neural Network) architecture. Most of the previously cited papers compared machine learning to deep learning methods, hence applying both methods. Only on paper (OLIVA et al., 2019) did not use machine learning or deep learning methods, using only traditional NLP techniques.

B.3 Conclusion

This survey provided an understanding of the methods that have been recently applied to text classification and extract information from clinical notes using NLP, machine learning, or deep learning. It was essential to support the development of experiments on text classification or information extraction.

As a result of this survey, thirteen papers that focus on NLP, machine learning, and deep learning on text classification or information extraction were selected and studied. It was possible to observe that machine learning classifiers are still in use providing similar results to deep learning classifiers, especially with small datasets. Furthermore, while evaluating the selected papers, it has been notice that there is space to improve the deep learning classifiers, which could leverage its performance.

This survey was used as the basis for the chapter of the related work of this work.