INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Text Mining de Relatórios Clínicos

Ana Catarina Martins Gonçalves

Mestrado em Ciência de Dados

Orientadores:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Doutor Maurício Breternitz, Jr., Professor Auxiliar Convidado,
ISCTE - Instituto Universitário de Lisboa

Outubro, 2023

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

Text Mining de Relatórios Clínicos

Ana Catarina Martins Gonçalves

Mestrado em Ciência de Dados

Orientadores:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Doutor Maurício Breternitz, Jr., Professor Auxiliar Convidado,
ISCTE - Instituto Universitário de Lisboa

Outubro, 2023

Acknowledgment

This work was partially financed by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., within the scope of projects UIDB/04466/2020, UIDP/04466/2020 and DSAIPA/AI/0122/2020.

Resumo

No âmbito do projeto de investigação em Inteligência Artificial AIM Health, foram obtidos ficheiros de texto, em português europeu, com relatórios de procedimentos e exames médicos, para explorar a possibilidade de extrair variáveis para melhorar algoritmos de Aprendizagem Automática. Uma análise inicial revelou que os textos incluíam dados pessoais, como nomes de médicos e pacientes ou datas.

A recolha, tratamento e armazenamento de dados são estritamente regulamentados na Europa e, sem consentimento explícito, dados pessoais não podem ser partilhados.

A remoção de dados pessoais em grandes volumes de textos não é uma tarefa simples. Identificar os dados manualmente é uma solução onerosa e propensa a erros. Existem soluções automáticas para apoiar esta identificação, mas surgem inúmeras dúvidas ao avaliar o desempenho e a equidade destes mecanismos.

Este trabalho visa proporcionar uma melhor compreensão dos textos, dos possíveis dados pessoais neles contidos e dar apoio sobre como geri-los. O objetivo final e fornecer um solido ponto de partida para trabalhos futuros e promover o uso responsável dos dados.

Foram analisados cerca de 2.000 notas de admissão e relatórios de procedimentos e exames, e identificados quase 4.000 blocos de texto com eventual informação identificável, em 12 categorias distintas. Para apoiar a anotação manual, foi desenvolvida uma ferramenta customizada, e cerca de 12.000 abreviaturas registadas, resultando num dicionário auxiliar com 967 abreviaturas distintas, a sua forma completa e tipo semântico.

Finalmente, com base no relatório anterior, algumas experiências com identificação automática provaram que estes métodos, com supervisão responsável, podem ser um recurso valioso.

Palavras-chave: NLP; Texto Clínico; Anonimização de Texto.

Abstract

In the context of the Artificial Intelligence scientific research project AIM Health, text files, in European Portuguese, with reports of medical procedures and exams were made available, to explore the possibility of extracting features to improve Machine Learning algorithms. An initial analysis revealed that the texts included Personally Identifiable Information, such as full names of physicians and patients or dates.

Data collection, treatment, and storage are strictly regulated in Europe and without explicit consent, personal data cannot be shared. Removing Personally Identifiable Information from large amounts of text is not a simple endeavor. Manually identifying data is a very costly solution and prone to error. Automatic solutions can support the identification of sensitive data, but questions arise when assessing these mechanisms' performance and fairness.

This work aims to provide a better understanding of the texts, possible personal information in them, and support on how to govern them. The end goal is to provide a solid stepping stone for following works and promote responsible use of the data.

Around 2,000 admission notes and procedure reports were read and almost 4,000 possible Personally Identifiable Information were identified, in 12 distinct categories. To support manual annotation, a custom tool was developed, and nearly 12,000 abbreviations were registered, resulting in an auxiliary dictionary with 967 unique abbreviations, their complete form, and corresponding semantic types.

Finally, based on the previous report, some experiments with automatic identification proved that these methods, with responsible supervision, can be a valuable resource.

Keywords: NLP; Clinical Text; Text anonymization.

Contents

List of Figures

List of Tables

Acronyms

AIM Health: Artificial Intelligence based Mobile Applications for Public Health Response

Att-CNN: Attention Convolutional Neural Network

BERT: Bidirectional Encoder Representations from Transformers

Bi-LSTM: Bidirectional Long Short-Term Memory Networks

Bi-LSTM-CRF: Bidirectional Long-Short-Term-Memory Conditional Random Fields

BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

CNN: Convolutional Neural Networks

COVID-19: Coronavirus Disease 2019

CRF: Conditional Random Fields

CSV: Comma Separated Values

CT: Computed Tomography

cTAKES: Clinical Text Analysis and Knowledge Extraction System

CTC: Clinical Terminology Center (Centro de Terminologias Clínicas)

EHR: Electronic Health Records

ELMo: Embeddings from Language Model

EMR: Electronic Medical Records

FN: False Negative

FP: False Positive

GDPR: General Data Protection Regulation

GloVe: Global Vectors for Word Representation

GPT: Generative Pre-trained Transformer

GPT2-BioPT: Portuguese Biomedical GPT-2 small

HC: Health Care

HCW: Healthcare Workers

HIPAA: Health Insurance Portability and Accountability Act

i2b2: Informatics for Integrating Biology and the Bedside

ICD: International Statistical Classification of Diseases

JSON: JavaScript Object Notation

LOINC: Logical Observation Identifiers Names and Codes

LSTM: Long Short-Term Memory Networks

MedLEE: Medical Language Extraction and Encoding System

MeSH: Medical Subject Headings

MIMIC: Medical Information Mart for Intensive Care

ML: Machine Learning

MMLU: Massive Multitask Language Understanding

MRI: Magnetic Resonance Imaging

MT: Machine Translation

NER: Named Entity Recognition

NLM: United States National Library of Medicine

NLP: Natural Language Processing

NLTK: Natural Language Toolkit

PACS: Picture Archiving and Communication System

PHI: Protected Health Information

PII: Personally Identifiable Information

POS: Part-Of-Speech

RE: Relation Extraction

RNN: Recurrent Neural Networks

SDK: Software Development Kit

SNOMED CT: Systematized Nomenclature of Medicine: Clinical Terms

TP: True Positive

UMLS: Unified Medical Language System

UTI: Urinary Tract Infection

WHO: World Health Organization

CHAPTER 1

Introduction

1.1. Contextualization and Motivation

Artificial Intelligence based Mobile Applications for Public Health Response (AIM Health) is a scientific research project, funded by the Portuguese government, that aims to develop a mobile application and distributed service to enable the identification of Coronavirus Disease 2019 (COVID-19) patients and exposure risk in a preventive approach. These services have a potential wider impact by providing valuable public health information and other patient health assessment tools. In the scope of this project, among Hospital structured data, around 12,000 text files with reports of various medical exams and procedures, such as computed tomographies or angiograms, were made available.

The original raw text files had not been explored or analyzed. The first inquiry that inspired this report was to try to understand if it was possible to extract features from these texts to enhance the performance of the deep learning models developed in the scope of AIM Health project.

The work started with the preparation and exploration of the given raw text files, resulting in a brief overview of all the data found. These files had the same data structure, including information about procedures and exams, patients, and the data extraction process. Afterwards, the two free text narratives found were analyzed with more detail: short admission notes, called observations, and longer technical reports with detailed descriptions of exams or procedures.

During this initial phase, in the context of the AIM Health research project, some inquiries were made to detect various pathologies or diseases' diagnostics in the texts, like thromboembolisms. This seemingly simple task made clear the difficulty, for a non-specialist, to extract clinical information from narratives. Most often the texts do not have a direct statement that allows a non-expert to assume a diagnostic.

Also in this initial exploration, one important feature of both text types was immediately obvious and had to be addressed if the goal was to use the texts to support research studies and Machine Learning (ML) tasks: the texts included Personally Identifiable Information (PII), such as full names of physicians and patients or exams dates. As they are, the texts cannot be shared, and their storage and manipulation have to be strictly monitored.

Data collection, subsequent treatment, and storage are strictly regulated in Europe. Any of those tasks may only be performed under previously given explicit and informed consent of the individuals. The alternative is to remove all identifiable personal data from the texts. The difficulty lyes in removing personal data from large amounts of text. Manually identifying data is a very costly solution and prone to error. Automatic processes can support identification, but questions arise when assessing these solutions, not only about their performance in warranting that identification is not possible but also about their fairness.

Additionally, there is a trade-off that must be taken into consideration: to remove data is to remove value. While the first objective is to guarantee that no identifiable information is disclosed, we need to ensure the minimum value loss.

1.2. Objectives and Research Questions

Knowing that these texts hold valuable data and in the face of the existence of personal information, this work aims to provide a better understanding of the texts, and the possible personal information in them and give some assistance on how to govern it. The end goal is to provide a solid stepping stone for following Natural Language Processing (NLP) tasks.

The challenge to remove identifiable information from free narratives begins with the question of what identifiable data is. The answer is not straightforward. NLP de-identification tasks usually start by classifying text using a pre-defined set of categories that can be considered as PII, such as names or telephone numbers.

In clinical text research, almost always these categories follow the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[1]. HIPAA guidelines, a United States regulation, are remarkable because they list and define types of Protected Health Information (PHI) that can be found in medical documents [**1**], being a very useful resource. At the same time, these categories can be a too narrow classification when trying to guarantee that all identifiable data is removed.

Some studies use a broader PII definition and include more information in the defined categories. For example, HIPAA guidelines only consider ages of 90 and above as PHI [**2**] but many studies identify all age references in texts. Moreover, there can be identifiable

[1] https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

information that is not possible to categorize [**3**,**4**]. For example, in the narratives studied in this work, it is mentioned an accident with an agricultural machine, makes it possible, arguably, due to the event rarity, to identify the patient. There is no PHI category in HIPAA guidelines to classify this information.

After identifying two possible text narratives to work with, a more in-depth analysis is required to understand what information the texts hold and if admission notes and reports on procedures or exams are similar or should be treated differently. The features to extract from different text types likely vary, and extraction strategies may have to adapt to the narrative type. Possibly, even the existing PII entities will differ, and the following de-identifying methods will require different approaches.

This text characterization aims to go beyond the commonly used PHI categorization and help support decisions on how to proceed in removing identifiable information.

The identified abbreviations were also examined and characterized, using their semantic type, given by a clinical domain-specific tool. This semantically categorized abbreviations dictionary can also be a useful resource for possible future NLP tasks using these texts or similar ones.

1.3. Methodology

The first proposed task was to detect and characterize all possible identifiable information, the texts had to be analyzed and annotated manually. Since no specialists were available, the task of annotating the texts was taken over by a non-specialist with no annotation experience, a questionable decision.

To support annotation, a custom annotation tool was developed. A customized tool allowed the adaptation and implementation of custom features more easily, not being limited to a standard annotation task. Also, this annotation mechanism did not require local installation or the exchange of data files, a significant benefit if, in the future, clinical experts could use it to review the annotated texts.

The annotation process was complicated by the omnipresence of abbreviations, many of them specific to the clinical domain, and difficult for the non-professional to fully understand, even when recurring to standardized medical terminologies. As the texts were being read in full to annotate possible sensitive data, a new feature was added to the above-mentioned annotation tool, to enable the annotation of abbreviations as well. The first idea was to give the annotator an easy-to-use abbreviation dictionary built on the abbreviations seen before. Knowing that this dictionary would be helpful to a human annotator, there was also the hope that it could be a resource for other following NLP tasks.

Around 1,000 observations and 1,000 reports were read resulting in the identification of almost 4,000 possible PII and nearly 12,000 abbreviations. This annotation process, with all its known faults, resulted in a more solid knowledge of the texts, but it was very time-consuming. One can argue that, given medical professionals' cost and the time needed to annotate large amounts of text, this non-expert annotation can be a building block for a future more robust annotation task. facilitating and speeding the specialists' work.

A detailed analysis of the found PII and abbreviations followed. Abbreviations' description and characterization were based on the Unified Medical Language System (UMLS) Metathesaurus Browser and Clinical Terminology Center (Centro de Terminologias Clínicas, CTC) Catalogues, but the most source of knowledge was the texts themselves.

Based on the previous assessments, the final task consisted of a simple approach to automatic personal data identification.

The first experiments used simple regular expressions to detect identification numbers, dates, and names. A Microsoft tool to anonymize text was tested. This tool allows language adaptation, in particular for the supporting NLP tasks like tokenization or Named Entity Recognition (NER), which was customized to use a spaCy's Portuguese pipeline. The tool also allows the adaptation of the PII categories recognizers, but these remained with the default settings. More two models were tested, based on complex neural networks, contextualized embeddings, and general and domain-specific language models, both trained to de-identify clinical text, but only one using the Portuguese language.

1.4. Dissertation Structure

The document starts with a brief overview of what text mining is and NLP and the more recent developments in the area. With a focus on the clinical domain and Portuguese language specificities, de-identification tasks are given special attention.

The methodology part of this document describes in detail data preparation, text annotation, and analysis tasks, and the use and evaluation of automatic processes to identify PII. For each task, a step-by-step explanation is given, justifying the choices made along the process.

The outputs from the tasks described in the methodology are presented in the results chapter. From both narrative types perspectives, identified PII findings are examined category by category, and found abbreviations description was organized using their semantic type.

To avoid, as much as possible, extensive and repetitive text listing numeric data, some results are presented in tables that aim to be as simple as possible. Also, some figures are included to better illustrate relevant findings.

The results from the automatic PII identification methods tested are presented by method, with a final overview comparing them, once again the analysis is done for each category separately.

The document closes with a succinct report on the most relevant conclusions, including the work contributions, limitations, and future work perspectives.

CHAPTER 2

Literature Review

2.1. Text Mining and Natural Language Processing

Text is our most common medium for the formal exchange of information, but often can lack the needed structure to allow traditional automatic processing. Text Mining is the process of extracting structured data from text [**5**, pp.515]. NLP is part of the methods and techniques used in Text Mining. Witten et al. define it as “the process or ability of a machine or program to understand natural (or human) text or speech” [**6**, pp. 283], clarifying that tools for information retrieval, like web scrapping, or the use of dictionaries, are within the scope of Text Mining but not NLP. NLP makes use of several techniques to achieve its goals, NER and Relation Extraction (RE) are very common and omnipresent techniques [**7**]: NER aims to identify, classify and categorize named identities and it is commonly used in clinical NLP to de-identify text, but also to discover symptoms, diseases, drugs or body parts [**8**]. RE identifies the relation between these entities. Another valuable NLP technique is Part-Of-Speech (POS) tagging, which classifies words in text according to their morphosyntactic value.

ML techniques are classified depending on the process complexity: 1) classical machine learning obtains results directly from a given dataset; 2) representational learning makes use of some intermediate data mapping prior to results, a classic example is using a nearest neighbour algorithm to a dataset where, based on the original features, principal components have been identified and summarized; and 3) deep learning relies on multiple steps, each additional step provides an architecture “depth” [**5**]. In this context, we will consider classical and representation machine learning as traditional machine learning or ML, as opposed to deep learning.

Deep learning builds on multiple abstract levels, mainly neural network-based algorithms that make use of advanced optimization techniques like Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) or transformers [**9**]. Deep learning has revolutionized NLP, but it also presents some new technical challenges: these algorithms are computationally demanding and hard to parallelize; CNNs can have difficulty identifying long-distance relations present in texts and RNNs, that explicitly model sequential relations are limited by text length; some variations have been used to address these issues, but with limited success [**10**]. Many of these problems have been addressed with the introduction of attention mechanism and transformer architectures [**9**], like Bidirectional Encoder Representations from Transformers (BERT) or Generative Pre-trained Transformer (GPT).

Text, or words, are the input for these methods. Embeddings are representations of words, that, ideally, retain as much as possible the word meaning in the given context, or its semantic, in a low-dimensional continuous space and make it easier to do ML on large inputs [**11**]. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. An embedding can be learned and reused across models.

Deep learning methods can receive as input embeddings. We can distinguish embedding methods in static and contextual. Word-level vector representation methods, or static word embeddings, assign a constant vector to each word, a single global representation [**12**]. Examples of word-level vector representation available methods are Word2vec[1], a word-level representation method that improves continuous Skip-gram model [**13**], Global Vectors for Word Representation (GloVe)[2] [**14**] and FastText [3] [**15**]. More recently, contextual representation methods, like Embeddings from Language Model (ELMo) [**16**], BERT [**17**] and GPT [**18**], took embeddings to a new level when considering word context in their vector representation by assigning different vectors to the same word, depending on their context, and replaced the use of static word embeddings with significant improvements on many NLP tasks [**11**, **19**]. The term "contextual representation" can be misleading because static representations use context to generate the representation, but the representations themselves are not context-dependable. Contextual embedding models are pre-trained on large unlabelled corpus to build context-sensitive embeddings; but whereas ELMo takes these embeddings as input features for the downstream task (is feature based) BERT and GPT-2 integrate the entire model downstream, a fine-tuning approach; additionally, GPT-2 takes advantage of a deep uni-directional transformer model and BERT of a bidirectional one that can improve long-distance context comprehension [**11**, **19**]. Also, BERT uses as input the entire given context at once while GPT-2 is

[1] https://www.tensorflow.org/text/tutorials/word2vec

[2] https://nlp.stanford.edu/projects/glove/

[3] https://fasttext.cc/

auto-regressive and incorporates its previous output to the sequence of inputs in the next step [**12**].

Flair embeddings, or contextual string embeddings, are, considering that these representations vary depending on text context, and contextual representation methods, but operate in a character sequence level. These methods maintain the ability to be trained in large unlabeled corpora, and also, due to the character sequence approach, handle well rare, misspelt and subwords, like suffixes and prefixes [**20**].

In March 2023, OpenAI reported the development of GPT-4. Their technical report describes better performances than other models in English and other 26 languages, after translating the Massive Multitask Language Understanding (MMLU) benchmark. Portuguese is not one of the tested languages but Spanish, French and Italian, similar Latin languages, were among the top six performers, with better results than English GPT-3.5 [**21**, fig.5].

2.2. Clinical Text Mining

There is a considerable amount of research focused on texts with clinical content, some patient-authored, some created by experts, or others like tweets, online forums or medical literature, all of which have different linguist profiles [**22**]. Text is still our most natural and meaningful way to record clinical events and Electronic Medical Records (EMR) hold valuable information about patients in free-text clinical descriptions. EMR are digital reports of patients' assessments, predominantly created by clinical professionals and administrators [**22**]. It is worth mentioning that Garets and Davis distinguish the terms Electronic Health Records (EHR) and EMR depending on their environment and scope [**23**], but in this context EMR and EHR will assume the same meaning, taking the broader definition of both.

Even when only contemplating clinical notes from EMR, there is a vast variety of audiences, authors' profiles, styles, tasks, requirements and outcomes to consider. For example, radiology reports and discharge summaries differ in style, structure and vocabulary. Furthermore, even in the scope of a specific task and text, like disease phenotyping, there are distinct attributes when considering different medical specialities. As an example, though metabolic diseases are more prevalent in the general population, there is more clinical NLP research in diseases of the circulatory system than metabolic diseases, maybe because metabolic diseases diagnostics rely more on structured data [**24**].

As these clinical texts are taken from discharge summaries from patient's EHR, they contain some known issues as grammatical errors, use of acronyms, and lack of a formal structure

Clinical text differs from a standard text in many aspects: misspellings, and domain-specific terminology and abbreviations are prevalent; clinical texts contain incomplete sentences, frequent negations (e.g. when excluding symptoms), and vague, uncertain or speculative expressions; and there are differences between text types and medical specialities [**8**,**25**]. For example, discharge summaries and autopsy reports, written for a broader audience, are usually better structured than other clinical narratives which explains their prevalent use in research [**26**]. Clinical NLP is NLP accommodating for clinical text specificities, needs and requirements.

Clinical concept extraction, the most popular clinical NLP task [**19**,**22**], can be defined as the automatic process of identifying clinical concepts from unstructured text. It consists of concept detection, generally NER, and concept encoding, which allocates standard or pre-defined terminologies to the identified entities. The term BioNER is also used to identify the various tasks that aim to identify biomedical entities [**27**]. Another common clinical NLP task is phenotyping, which consists of the identification of a patient's health condition and relation extraction [**22**].

Clinical concepts in NLP tasks are controlled vocabularies like the Systematized Nomenclature of Medicine: Clinical Terms (SNOMED CT) or standardized terminologies or ontologies like the Medical Subject Headings (MeSH)[4] thesaurus, an index for health-related concepts, or the International Statistical Classification of Diseases (ICD)[5], a database with 80,000 entries characterizing diseases and syndromes, maintained by the World Health Organization (WHO), intended to offer global accurate and comparable statistics on causes of mortality and morbidity or health-related phenomena [**28**, **29**]. Developed by the United States National Library of Medicine (NLM), UMLS[6] is omnipresent in Clinical NLP. It comprises three knowledge sources: the Metathesaurus, the Semantic Network, which provides high-level categories for the concepts in the Metathesaurus, and a syntactic lexicon, SPECIALIST Lexicon, and related tools [**23**]. The UMLS Metathesaurus is updated quarterly and integrates biomedical terms from various other sources, including SNOMED CT, MeSH, ICD, Logical Observation Identifiers Names and

[4] https://www.nlm.nih.gov/mesh

[5] https://www.who.int/standards/classifications/classification-of-diseases

[6] https://www.nlm.nih.gov/research/umls

Codes (LOINC)[7], a catalogue on health measurements, observations, and documents, and RxNorm[8], a normalized naming system for clinical drugs. Although there are several domain-specific resources, Bay et al. note that some patient-related documents have a low rate of medical ontological terms and highlight the importance of non-medical language in patient-related document characterization proposing the construction of a specific body of terms from both medical and everyday language [**30**].

Since 2018, and despite its "black box" status, there seems to be a growing acceptance of deep learning by the medical community. However, although deep learning has largely outperformed traditional methods, there is an understandable lag in the adoption by the medical field [**22**]. Wu et al. show that, at least until 2019, RNNs have dominated clinical deep learning architectures research in all types of tasks, except text classification where CNN appeared in half of the analyzed documents [**22**]. At the same time, they highlight the increasing popularity of attention mechanisms, often used with other methods, and predict the future growing research in attention mechanisms.

Embedding pre-training can be an alternative to collecting annotated data, a difficult task, especially in the area of clinical data [**18**]. Developing an annotated corpus in a clinical context can be very difficult due to sensitive and private information, text specificity, and the need for medical experts' input [**25**].

Another hurdle when working with these new embedding methods is that they are computationally demanding and can struggle to process long texts. Ethayarajh suggests extracting static representations from contextualizing models, claiming that these can often perform better than traditional static embeddings [**11**]. Another alternative is BERT-based models pre-trained with public biomedical texts or clinical texts, like Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) [**31**], ClinicalBERT [**32**] or EhrBERT [**12**].

Domain adaptation is key in clinical narrative analysis. Transfer learning, for example using the pre-trained BERT models with an extremely large set of domain-specific narratives to produce a language representation model, is one possible strategy to leverage the domain specificity challenge. Another is multi-concept learning (or multitask learning), which means making use of features from multiple datasets to improve model performance [**10**].

[7] https://loinc.org/

[8] https://www.nlm.nih.gov/research/umls/rxnorm/index.html

A comparative study of various embedding methods for clinical concept extraction found that: i) contextual embeddings perform better than traditional embeddings, and can provide additional semantic information; ii) it is beneficial to pre-train on a clinical domain corpus (domain-specific embedding versus open-domain) in both traditional and contextual embeddings; iii) pre-trained deep language models with a large corpus, followed by a fine-tuning approach (BERT) seem to outperform all other methods [**19**].

One of the challenges of clinical NLP is misspelling. Tissot & Dobson [**33**] note that generic drug name spelling errors occur in up to one out of six entries in information systems, and that these errors are not innocuous and can lead to adverse drug events. To address this issue, they combined string and phonetic similarity to correct drug misspellings. The hurdle is that phonetic representation (based on the sound of the spoken word) is highly language-dependent, demanding a language-specific tailored solution. Unfortunately, they only explore a hybrid ruled-based solution, suggesting the use of machine learning methods in possible future work.

2.3. Portuguese Clinical Text

Another area where transfer training can be useful, is language adaptation. Wu and al. observed that a great majority of published clinical NLP studies used English datasets, followed by Chinese (almost 20%), and works with datasets in other languages were less than 3% [**22**]. In their study, there is no mention of Portuguese datasets. Also, essential standardized medical ontologies, clinical thesauri, or controlled vocabularies are much more mature for the English language, which is predominant in the biomedical field, when compared to the Portuguese language [**34**], and the majority of clinical corpora for research are in English [**8**].

In European Portuguese, the CTC[9], publishes several semantic health catalogues, mapped with internationally recognized classifications and terminologies like SNOMED CT and LOINC, to be used in Portuguese Health Information Systems.

Historically, a common approach to tackle the Portuguese language handicap has been making use of Machine Translation (MT) to translate the texts to and from English. In 2007, Coutinho et al, translated Portuguese X-rays reports to English and then used the Medical Language Extraction and Encoding System (MedLEE), a NLP Service, to binary classify 165 texts considering two questions: radiography normality and presence of devices [**35**]. More recently, Lamy et al. intended to use the open-source Clinical

[9] https://www.ctc.min-saude.pt/

Text Analysis and Knowledge Extraction System (cTAKES) to detect a variety of clinical information: diseases, medications, symptoms, signs, anatomical regions and procedures. To take advantage of both cTAKES and SNOMED CT, not available in Portuguese, the texts were translated to English with Google Translator [**36**]. The common use of clinical acronyms and abbreviations was addressed by replacing them with the full word prior to translation, but these were manually identified. As expected, results were below cTAKES' standards. In 2021, Schneider et al. trained a GPT-2-based model with titles and abstracts from scientific papers that were translated to Portuguese using MT [**12**].

Even though BERT and similar models are conceptually simple and promise great results, training these models requires large amounts of data that can be difficult to access, particularly when working with non-English languages [**37**].

Acknowledging that i) neural networks are the state-of-the-art POS tagging architecture approach, in particular Bidirectional Long Short-Term Memory Networks (Bi-LSTM); ii) the scarcity of studies exploiting this method for Brazilian Portuguese clinical texts; and iii) the difficulties in building clinically annotated corpora; de Oliveira et al. propose a POS tagging neural architecture for these texts making use of a corpus that results from the combination of four different biomedical and journalistic corpora and "achieved comparable results to other state-of-the-art studies in journalistic contexts", with an accuracy of 92.39% for Brazilian Portuguese clinical texts [**25**].

Motivated by the scarcity of clinical NLP research in Portuguese, Lopes et al. extended their previous Conditional Random Fields (CRF) study [**38**], and compared the performance of different methods for NER in Portuguese clinical data using CRF and two Long Short-Term Memory Networks (LSTM) models with different word embeddings pre-training approaches using fastText: a general-knowledge and an in-domain model [**39**]. The general-knowledge model was built on billions of tokens from Wikipedia and common crawl, and the in-domain model was trained with 3.377 clinical texts from a journal (a total of 686.762 tokens). The LSTM models achieved the best performance and, in line with what has been found in other languages, the in-domain embeddings outperformed the general knowledge, even if trained with much fewer tokens. More interestingly, the results suggest that it is possible to train a model with publicly available data (for training purposes, the research only used Portuguese clinical texts from journals). However, it had a very small test dataset of 20 clinical notes from the Neurology service of a Portuguese hospital, including admission notes, diagnostic test reports and patient discharge letters.

Duarte et al., also using deep learning with word embeddings, show accuracy scores over 75,9% when assigning ICD-10 full-codes (a large number of classes sparsely used) for underlying cause of death to Portuguese death certificates, clinical bulletins, and autopsy reports [**40**]. Unfortunately, this study does not reflect on the specifics and challenges of Portuguese or the domain language.

Coutinho and Martins reproduce the task by Duarte et al. [**40**] using a transformer-based model BERT model and outperform the previous models with an accuracy score of over 80% when assigning full-codes with a model that was pre-trained with Portuguese clinical data [**37**].

Another study also assigns ICD-10 codes to Brazilian-Portuguese clinical notes using logistic regression, a CNN, a CNN and Attention Convolutional Neural Network (Att-CNN) [**41**]. All neural networks used word2vec embeddings and they worked with different subsets of note types (discharge summaries, clinical developments and physical exams). The Att-CNN model, with all types of notes, was the one with the best result with a F1 score of 0.485.

Aiming to compare clinical word-embedding models with different granularities (coarse versus fine-grained), Silva e Oliveira et al., set to identify Urinary Tract Infection (UTI) cases in clinical narratives from hospitals using a neural network [**42**]. They trained three word2vec models with three different datasets: 1) a coarse-grained, with 745,731 documents from 3 different hospitals, multiple types of notes and several medical specialities; 2) a subset of this coarse-grained with only the narratives annotated with UTI related codes and 3) also narratives annotated with UTI related codes but selected from another hospital. They also trained a GloVe model with the third dataset to train the neural network. The best results were very similar for all models, allowing to conclude "a tradeoff between corpus size and similarity when it comes to word-embeddings". To evaluate the coarse word embeddings model (an evaluation that does not depend upon the downstream tasks), experts translated and adapted to Brazilian Portuguese the Bio-SimLex database, a clinical nouns dataset where word-pairs similarity is scored to support intrinsic evaluation of word representation models [**43**]. Although with some challenges in the translation (ambiguous terms and pairs with no obvious translation), they conclude that the results are comparable to the state-of-the-art models.

Portuguese Biomedical GPT-2 small (GPT2-BioPT) was made available by Schneider et al. in 2021 [**12**]. GPT2-BioPT is a GPorTuguese-2 model trained with titles and abstracts from Portuguese scientific papers; whereas GPorTuguese-2 is an English GPT-2 model pre-trained with roughly 1GB of data from Portuguese Wikipedia [10]. The model was evaluated on a classification task with the Flair API, where given an annotated dataset with progress notes the goal is to identify fall incident reports, a classification task. The FLAIR framework[11], developed using the PyTorch[12] Python library, provides a simple interface to facilitate the easy use of different embeddings. [**44**], and an exhaustive assessment of various static embeddings (Word2Vec and FastText) and contextual embeddings models (ELMo, BERT and GPT) with and without domain (generic, biomedical and clinical) and language (English and Portuguese) specific training was carried out. GPT2-BioPT outperformed all other models with an F1-score of 0.90. The second-best result was obtained by BioBERTpt [13], a deep model developed to support Portuguese clinical and biomedical NER, based on a multilingual-BERT model pre-trained with a corpus of clinical texts and biomedical papers in Brazilian Portuguese [**26**]. The results strongly suggest a gain when applying contextual versus static embeddings and the benefits of domain and language-specific pre-training.

2.4. Text De-identification

Data privacy regulations have a strong impact on all data-driven medical research and production due to the large volume of PHI in EMR [**45**]. Since 2018, the General Data Protection Regulation (GDPR) has enforced strict constrains regarding data collection, treatment, and storage in Europe. Without explicit and informed consent of the individuals, personal data can only be used for the purpose used when collecting it [**46**]. De-identification is the process of removing data that can be associated with a specific person [**47**]. Although some authors make a distinction between de-identification and anonymization, for example, Kayaalp writes that de-identication is a well-defined process and anonymization a goal [**47**], the difference is not very relevant in this context and the terms will be used interchangeably.

De-identification methods are the combination of two very different tasks: the identification of sensitive data and the concealment of the identified data. Different suppression

[10] https://huggingface.co/pierreguillou/gpt2-small-portuguese

[11] https://github.com/flairNLP/flair

[12] https://pytorch.org/

[13] https://huggingface.co/pucpr/biobertpt-all

strategies result in different levels of information disclosure. Removing the identified data and all its context, or the phrase containing it, is the more radical approach. But it is also possible to mask the information with a generic mask, like ‘XXXX’, or replace it with a specific class or category, like NAME, giving more information about the hidden data. Another option is to use a surrogate, or pseudo, like ’JOHN DOE’ [**48**].

With the exponential growth of the available data and the cost of manually identifying sensitive information, there is an increasing interest in automatic anonymization methods using NLP techniques [**49**]. It is an example of the success of NLP requiring the use of NLP.

Automatic de-identification tolls, like Presidio[14] and Amnesia[15] are becoming popular, but this automatic anonymisation poses new challenges balancing data protection and keeping the research value of anonymised datasets [**50**].

Another hurdle of automatic identification of sensitive data is the limitation to predefined categories, which may ignore texts that do not fit the defined labels but could still identify individuals [**3**].

Jain et al., in one of the first studies using spaCy[16] for clinical text de-identification, a Python open-source library for NLP processing, compare different deep architectures on the Informatics for Integrating Biology and the Bedside (i2b2) data set, including Bi-LSTM-CNN and conclude that, although Bi-LSTM-CNN shows a better F1 score, the model based on spaCy (version 2.1) outperforms all other models, with the second best F1 score and a better time cost [**51**]. On a later work, Pearson et al. also used spaCy and OpenNLP[17] to identify five categories of PHI on medical examination reports, and spaCy showed a better F1 score of 0.9075 and better recall scores.

More recently, using a multi-institutional dataset of 6,193 documents, with identified PHI, chest X-ray, Computed Tomography (CT) reports and medical notes, Chambon et al.developed a transformer-based deidentification pipeline, that also replaces the identified PHI with surrogates, and report to outperform all human labelers on i2b2 2014 data [**52**].

On the privacy issues, literature addressing specifically the Portuguese case is very scarce [**53**]. Santos et al. believe to carry, in 2021, the first study on the detection of patient names in Portuguese texts, with contextualized embeddings. The downstream tasks were based on a Bidirectional Long-Short-Term-Memory Conditional Random Fields

[14] https://microsoft.github.io/presidio/

[15] https://amnesia.openaire.eu/

[16] https://spacy.io/

[17] https://opennlp.apache.org/

(Bi-LSTM-CRF) neural architecture, supported by Flair framework: a Bi-LSTM model, trained for traditional sequence labeling, and a CRF token labeler. They used several language models generated from a corpus of both Brazilian and European Portuguese: WE-NILC[18], generated from 17 different sources and from various domains with almost 1.4 billion tokens, and trained using word embeddings (Word2vec, FastText, Wang2vec and Glove) [**54**]; WE-EHR-Notes[19]: also trained using word embeddings, it was generated from 603 million tokens from hospital progress notes [**55**]; FlairBBP: a pre-trained language model with 4.9 billion Portuguese words corpus, trained using with Flair embeddings [**56**]; and $FlairBBP_{FnTg}$: a FlairBBP refinement using the EHR-Names Corpus as input, a corpus built on 2,500 clinical notes from Brazilian hospitals manually annotated with 4,999 identified entities [**57**].

NLP classification methods evaluation usually employ precision, recall and F1 scores, where F1 is the harmonic mean of precision and recall. These measures compare the number of classified items in a result set to a previously defined classified set, the target set. But traditional metrics are being disputed. Hendrycks et al. reflect on NLP recent evolution with the advent of transformer models and how the model assessment will have to adapt to new standards, especially domain-specific trained models [**58**].

For text anonymization tasks, performance is highly associated with recall scores since recall assesses the significance of the terms that should have been identified and were not [**48**]. But, in de-identification, different terms and different masking techniques carry diverse information significance, they are not all equal. Recall, and other similar measures, do not this take into consideration and assign the same weight to all terms. An alternative is to evaluate re-identification risk, a much more complex metric [**4**].

Beyond pure performance evaluation, another important aspect to consider when employing automatic techniques is fairness, Xiao et al. have recently shown, studying name de-identification in clinical records, that existing de-identification systems can be biased [**59**].

2.5. Chapter Conclusions

The recent developments in NLP methods, supported by innovative pre-processing strategies, like flair embeddings, coupled with complex neural architectures, like transformers, have exhibited remarkable results.

[18] http://nilc.icmc.usp.br/embeddings

[19] https://github.com/nlp-pucrs/fall-detection

The rapid success of these very complex methods gives rise to new questions on model assessment and fairness.

Even among this success, there are lingering challenges when addressing domain-specific texts and languages, like in the case of clinical texts in European Portuguese.

Besides the models themselves, increasing attention is being given to the content of the texts used to train them. De-identifying text has been an active research area, due to strict regulation and the growing need for large amounts of text.

Traditional PII categorization and the following methods' evaluation present several constraints that call for new approaches, like de-identification risk.

To guarantee that these amazing tools deliver an accurate, responsible, and fair output, we face a seemingly unreasonable task: it is important to have a greater knowledge of the increasingly larger texts being used.

This work aims to provide a better understanding of two types of clinical narratives that will possibly be used in the development of ML algorithms, with particular attention to the existing PII and how to treat it. The following chapter will describe the tasks undertaken to achieve this, afterward, the results of the described activities will be reported and discussed.

CHAPTER 3

Methodology

This chapter describes in detail all the steps that were taken during this work in order to mine the text data: starting with the raw data preparation and exploration and ending with the performance evaluation of the automatic identification process. In between, the options taken when manually annotating, and the reasoning behind them, are presented, as are the descriptions of the validation, classification, and analysis tasks that have followed.

3.1. Data Preparation

The original data set comes from 12,651 text files, structured in JavaScript Object Notation (JSON) format with three properties: the extraction timestamp, an object "Patient" with an identification number and birth date, and an object *Report*. The *Report* includes an identification number (accession number), the exam's type, a validation timestamp, and two texts in rich text format: *Observation* and *Report*.

After transforming all rich text into a simple text format, using striprtf[1], this data was saved in JSON files (35MB) and loaded to a pandas dataframe, adding the original filename as a column for future reference. Ignoring filename and the extraction timestamp, dropped duplicates, in a total of 217 duplicated records. The resulting dataframe, with 12,434 distinct records and 8 columns was saved to a Comma Separated Values (CSV) file (27.4MB).

3.2. Data Exploration

All records were exported in February 2023 and include procedures or exams ranging from November 2020 to January 2023. From these, 85.33% (10,610) date from 2021, 10.56% from 2022, 0.01% from 2020 and 2023 and the remaining 510 had no date. In total, there are 1,541 distinct patients. There are only 7,302 distinct accession numbers and, for the same accession number, there can be different *exam types*, *reports*, and *validation timestamps*, but always for the same patient and presenting the same *observation* text, although the same *observation* can show different accession numbers. The fact that there

[1] https://pypi.org/project/striprtf/

are more records than unique values for several variables seems to indicate that these were exported from a relational database where an admission (*accession number*) is related to one patient and one *observation*, and more than one *exam type* and respective *report* text. Also, the same *report* can be found in different records with different *accession numbers*, *observations*, and *exam types*. For example, the most frequent *report* appears in seven records with seven distinct *accession numbers*, four distinct *observation* texts, and distinct *exam types*, all CTs of different anatomic parts.
Table 3.1 depicts a summary of the variables in the dataset.

TABLE 3.1. Variables description.

Variable	Type	Values count	Null values	Unique	Max. Freq.
Extraction timestamp	date and time	12,434	-	12,434	1
Patient id	integer	12,434	-	1,541	68
Patient birthdate	date	12,434	-	1,476	68
Accession number	integer	12,434	-	7,302	8
Observation	text	10,943	1,491	5,173	158
Exam type	text	12,433	1	254	1,865
Report	text	12,434	-	6,807	7
Validation date	date and time	11,924	510	6,303	9

Table 3.2 aims to give a brief description of the text variables, taking into consideration, for each variable, non-null unique values. Text length is the direct result of the *len()* Python function, which returns the number of characters in a string, and the token count counts the length of the result of the function *word_ tokenize()* from the Natural Language Toolkit (NLTK) Tokenizer Package[2], for Portuguese language default parameters.

Observations are text narratives that contain, predominately, exams or procedure requests with a brief clinical history. These are, sometimes, included in full in the *report* text labeled as clinical information. Concerning the text size frequency, three text lengths occur most frequently and were analyzed in more detail: 325 texts with 107 characters and

[2] https://www.nltk.org/api/nltk.tokenize.html

TABLE 3.2. Unique non null text variables description

Variable	Values count	Text length			Token count		
		Min.	Average	Max.	Min.	Average	Max.
Observation	5,173	2	242.1	1,025	1	41.0	216
Exam type	254	3	32.8	64	1	5.3	16
Report	6,807	2	1,493.5	10,546	-	235.4	1,687

192 with 62 were found to be scheduling notes, all with the same format, with date and location for an appointment or exam. Another 112 texts have exactly 1,025 characters, which happens to be the maximum size for all of the texts. These seem to be truncated texts, ending abruptly in the middle of a sentence or word. Some limitations on the exporting process may have truncated these texts.

For the annotation task, all 339 *observations* beginning with *Agendamento* (Booking) were removed after manual inspection to confirm there was no other relevant information. The resulting data, with 4,834 unique texts and the original filename for future reference, were shuffled and saved.

Exam types are texts that are shorter than *observations*, stating a type of exam or medical procedure and, sometimes, the corresponding part of the body to be examined. Presenting only 254 unique values and 68 types that only appear once, it is possible that there was a list of exams to choose from, extended with the possibility to add a new type with free text. The most frequent *exam type* is the CT, with 63% of all records referencing *TC* (which stands for CT in Portuguese) in the *exam type* text. The top-5 most common exams are CTs: chest CT (1,865 texts); upper abdomen CT (1,305); pelvic CT (1,258); additional angiography CT (1,097) and skull CT (875). Other common exam types are ultrasounds, 13% of the records, and 6% refer *RM*, which means Magnetic Resonance Imaging (MRI) in Portuguese.

Similar to the *observation* variable, but this time only showing unique values, an analysis of text length frequency showed that texts were also probably truncated in the exportation process, with 22 texts presenting exactly 64 characters, which is the maximum value. When reading these 22 texts, it becomes obvious that once more these are incomplete texts.

The *exam types* that have been found do not entirely match with the terms in the CTC catalogues for Radiology [**60**], Gastroenterology [**61**] and Cardiology [**62**], with the given preferred terms and synonyms. Of the 254 distinct exam types, 97 were found in the Radiology Catalogue, five in the Gastroenterology, and only one in the Cardiology catalogue.

For the text not found in any of the catalogues, another matching approach was performed: (i) considering all cases when the text for the exam type was included in the catalogue term; (ii) ignoring the last character in all unmatched texts. The rationale lies in the fact that Portuguese words can be gendered and that gender can alter the last letter in a word, e.g. there is an *exam type TC pélvic**o*** and a CTC catalogue term *TC pélvic**a*** and both mean 'pelvic computed tomography'; (iii) ignoring all text after including, the characters *(*, - and the word *excluí* (Portuguese for exclude).

For each entry, the catalogues display a preferred term and a synonym in Portuguese. These searches were always performed using the preferred term. If no match was found, a new search was performed using the synonym value. As an example, with the approximate search, the exam type *TC dos membros inferiores - Perna (ESQ)* (CT of the lower limbs - Leg (LEF)), will now match *TC dos membros inferiores (cada segmento anatómico)* (CT of the lower limbs (each anatomical segment)), a reasonable match. Using the approximate search it was possible to categorize a total of 136 records, 33 more than when using the exact match.

Counting all records, we have 10,273 *exam types*, 83% obtained with an approximate match. However, with this approximation, four cases matched terms in more than one catalogue. The most frequent *exam type* (665 records) with no corresponding term found in any of the CTC catalogues, is *Imag Geral - TC, suplemento de contraste endovenoso* (Imag General - CT, intravenous contrast supplement). In the Radiology catalogue, there is a synonym term *TC, suplemento de contraste endovenoso* (CT, intravenous contrast supplement) that could label this *exam type*, so maybe a more robust and complex matching system could yield even better results.

Reports are longer texts with detailed image descriptions and technical details, many of them ending with the name and clinical speciality of one or more physicians. These *reports* include data that may be redundant because it can be found in other fields or calculated based on other variables, like the patient's age. Patient and clinician names and identification numbers are frequent, and, for 385 unique *reports* (697 in total) the corresponding *observation* text is included at the beginning of the *report* described as *Informação clínica* (Clinical information), maybe intending to contextualize the following data. It can be assumed that the information system makes it difficult for the clinic to visualize all relevant information at once. Also common are references to previous exams or procedures, using their date as reference. Similar to what was found in *observations*, text length analysis revealed that there are texts that are not medical narratives but booking details, but no truncated texts were found. Although all texts have characters, 6 *reports* have a zero in the token count because they only consist of newline characters `\n`, and these are all ignored.

For the annotation task, all 36 *reports* beginning with *Agendamento* (Booking) were ignored, after confirming that they did not contain more information. 6,771 unique texts, and the corresponding original filename for future reference, were shuffled and saved.

A brief note about the three CTC catalogues used in this work: the Portuguese Catalogue of Radiology v1.0, the Portuguese Catalogue of Gastroenterology v1.1, and the Portuguese Catalogue of Cardiology v1.0, all issued in 2021. These had to be requested from the CTC and the data was provided in Excel files, but they had to be treated to maintain the information in a tabular form.

Table 3.3. CTC catalogues record number per level

Catalogue	Area	Sub-area	Sub-sub-area	Exam or procedure
Radiology	7	16	37	409
Gastroenterology	3	6	8	104
Cardiology	8	-	-	143
Total	18	22	45	656

Each catalogue has a list of procedures or exams with a hierarchical categorization. The difference between a category row and a procedure row is that the rows for the categories are a single merged cell, and the rows for procedures have various data. For the categories, different types of text indicate different levels: *Areas* are in uppercase text; *sub-areas* in normal text and *sub-sub-areas* in underlined text. Each procedure category is determined only by the row order in the file. Table 3.3 relates the number of categories and procedures in each catalogue.

To ensure that the defined hierarchical categorization did not rely on record order or text type, three new variables were added: i) the ID (a sequential number assigned to all rows); ii) the level and iii) the parent. The categories description was assigned to the English Preferred Term column. *Area*, a category in bold text and upper case, e.g. ***COMPUTED TOMOGRAPHY (CT)*** was assigned level 1, while level 2, the *sub-area*, was assigned to entries with descriptions in lower case, such as *Head and Neck*. Underlined descriptions, such as *Abdomen and Pelvis (Non-Vascular)*, are assigned as level 3, i.e, the *sub-sub-area*. All other entries, procedures or exams, were registered as level 4. The parent variable identifies the ID of the category to which the respective record belongs.

All the catalogues were reviewed manually, added to pandas dataframes and saved as CSV files.

3.3. Manual Data Annotation

3.3.1. Sensitive Data Identification

The first proposed task was to identify and characterize all data that could associate an individual with a specific record or event.

Identified data was labeled following HIPAA and several previous annotation methods [**1**, **63**–**67**], with a similar classification to the i2b2[3] and Medical Information Mart for Intensive Care (MIMIC)[4] datasets [**63**]. Using these guidelines, three main categories were considered: general data, which includes location and date references; patient (or related individuals, like family) and clinician or staff data.

Dates were considered only if more detailed than a year and less detailed than a day. When only the day of the week or month were mentioned these were also not identified.

All locations were annotated, including countries, cities, institutions and services (or wards). Floor and bed numbers were also considered locations and labeled as services.

Patients' and clinicians' personal data found in the text was annotated, such as sex, age, name, identification number or contact. Although HIPAA only considers PHI ages above 89 [**2**], all age references were annotated. For the sex category only explicit references were taken into account. Because Portuguese words are gendered it can be easy to infer the patient or clinician's gender, for example, "o doente" is a male patient and "a doente" a female male, or, due to the medical nature of the *reports*, anatomic or clinical description sometimes reveal the patient sex.

[3] https://www.i2b2.org/

[4] https://mimic.mit.edu/

A category “other” was also created to include all cases where the previous categories were not suitable.

Each string in the defined categories, named entity, was added to a list of identified entities and stored in JSON format detailing the position of the first and last characters of the string in the given text and the given category. For example:

```
{
    "ents": [
        {
            "start": 218,
            "end": 235,
            "label": "LOCATION"
        },
        {
            "start": 293,
            "end": 298,
            "label": "DATE"
        }
    ]
}
```

3.3.2. Abbreviations

The second proposed task was to detect whenever a word or phrase appeared in a truncated form, not in a complete version, namely abbreviations and acronyms. The use of symbols, when used with other meaning than the most mathematical meaning; for example / was identified when used as an or statement (not when it was used as a quotient). The term annotated abbreviation will be referring to both of these situations from now on.

Some rules were defined for consistency:

- Whenever possible, longer abbreviations were considered for annotation, for example, if *TC* (computed tomography, tomografia computorizada in Portuguese) was preceded by Angio, *Angio TC* was annotated as a single abbreviation with the description *ngiografia tomografia computorizada* (Angiography Computed Tomography). Exam names were described, whenever possible, using the preferred term in the CTC catalogues. In the previous example, *Angiografia Tomografia Computorizada* occurs as a preferred term. There were two exceptions to this rule:

    - When the abbreviation included texts in different languages these were annotated separately, e.g. *LMA NPM1* was annotated as *LMA*, *Leucemia Mieloide Aguda* (Acute Myeloid Leukemia) and *NPM1* as *Nucleophosmin 1.*
    - Exam names and anatomical terms were also annotated separately. For example, *TC TAP*, meaning *Tomografia Computorizada de Tórax, Abdómen e Pélvis* (Computed Tomography of the Chest, Abdomen and Pelvis) required two records: *TC* (Computed Tomography) and *TAP*(Computed Tomography)).
- To facilitate the following analyses, the description assumed the singular form, and when the abbreviation had no specific gender the description used the masculine form.
- Abbreviations immediately preceded by their full form were not annotated, but all other occurrences in the same text were.
- The given descriptions always followed the language used in the abbreviation. If an abbreviation could refer both to Portuguese and English languages, the Portuguese language was used. For example, when *CoV* was found, it was described as *Coronavírus* and not *Coronavirus.*
- If the abbreviation includes a symbol, like a period in the end, this was included as part of the abbreviation text.
- Obvious abbreviations where it was not possible to find the corresponding description were annotated as a single question mark.
- Dubious cases were not identified. As an example, many specific clinical tool names are upper case and can be, or not, abbreviations.

Identified abbreviations were stored in JSON format with the position of the first and last characters in the text and the corresponding description:

```
"abbrs": [
        {
            "start": 9,
            "end": 17,
            "desc": "Angiografia Tomografia Computorizada"
        },
        {
            "start": 76,
            "end": 78,
            "desc": "Endovenoso"
        }
    ]
```

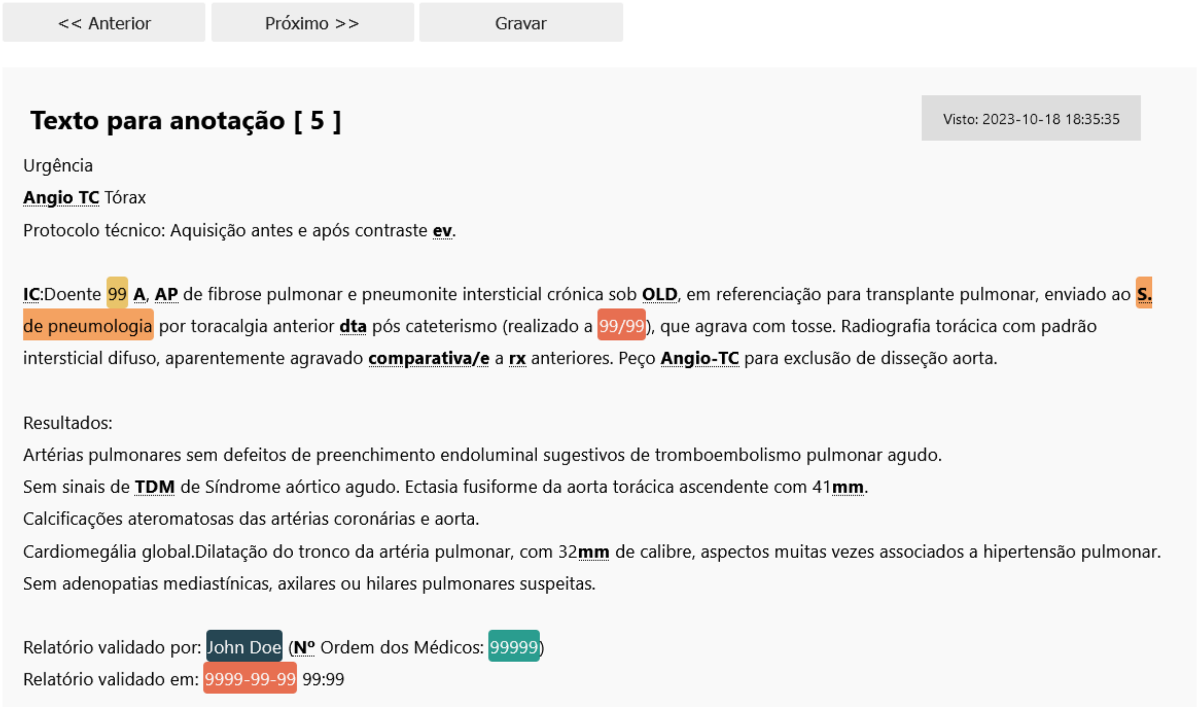
3.3.3. Annotation Tool

Searched for available annotation tools, but none provided the desired functionalities to automate the process or could treat PII categories and abbreviations, and the corresponding description, at the same time, as intended.

Moreover, keeping in mind that medical specialists should review the annotation and wanting the minimize data sharing, the goal was to have a very easy-to-use tool that could work on the already shared set-up environment (Jupyter Lab[5]). This would allow the creation of specific credentials to access the tool, without having to share data files or elaborate set-up instructions.

An annotation tool was built in Python. The annotator only has to run a Jupyter notebook[6]. Given the dataset location and an index row, an interactive widow[7] runs within the notebook, with four distinct panels: navigation, text, and add and remove panels. It also has an automatic labelling function.

The navigation panel allows loading of the previous and next texts and has a save button to save working changes in a CSV file. Figure 3.1 is an example of an annotated text as it is presented in the tool. In the figure, all sensitive information was altered.



Texto para anotação [5]

Visto: 2023-10-18 18:35:35

Urgência

Angio TC Tórax

Protocolo técnico: Aquisição antes e após contraste **ev**.

IC:Doente 99 **A**, **AP** de fibrose pulmonar e pneumonite intersticial crónica sob **OLD**, em referenciação para transplante pulmonar, enviado ao **S.** de pneumologia por toracalgia anterior **dta** pós cateterismo (realizado a 99/99), que agrava com tosse. Radiografia torácica com padrão intersticial difuso, aparentemente agravado **comparativa/e** a **rx** anteriores. Peço **Angio-TC** para exclusão de disseção aorta.

Resultados:

Artérias pulmonares sem defeitos de preenchimento endoluminal sugestivos de tromboembolismo pulmonar agudo.

Sem sinais de **TDM** de Síndrome aórtico agudo. Ectasia fusiforme da aorta torácica ascendente com 41**mm**.

Calcificações ateromatosas das artérias coronárias e aorta.

Cardiomegália global.Dilatação do tronco da artéria pulmonar, com 32**mm** de calibre, aspectos muitas vezes associados a hipertensão pulmonar.

Sem adenopatias mediastínicas, axilares ou hilares pulmonares suspeitas.

Relatório validado por: John Doe (**Nº** Ordem dos Médicos: 99999)

Relatório validado em: 9999-99-99 99:99

FIGURE 3.1. Annotated text in the annotation tool

[5] https://jupyterlab.readthedocs.io/

[6] https://jupyter.org/

[7] https://ipywidgets.readthedocs.io/

The text to annotate is shown with a specific style to facilitate the annotation process and minimize errors. For example, the text was given a bigger line height and larger padding and margins than the default values. Identified entities are highlighted with a specific colour background and contrasting text colour, depending on their identified category. Identified abbreviations' text is bolder, underlined and, when hovered, the given description is shown. The index number of the text being annotated and the last checked date are also displayed.

To annotate a new entity, the user must write the text to label and choose a category from a drop-down box. To identify an abbreviation, one has to write the abbreviation as it is in the original text and the corresponding long description.

The tool will label all strings in the original text that match exactly the input strings if there is no overlapping to already identified data. This means that a given character in the original text can only be part of a single abbreviation or entity. Abbreviations and entities can overlap.

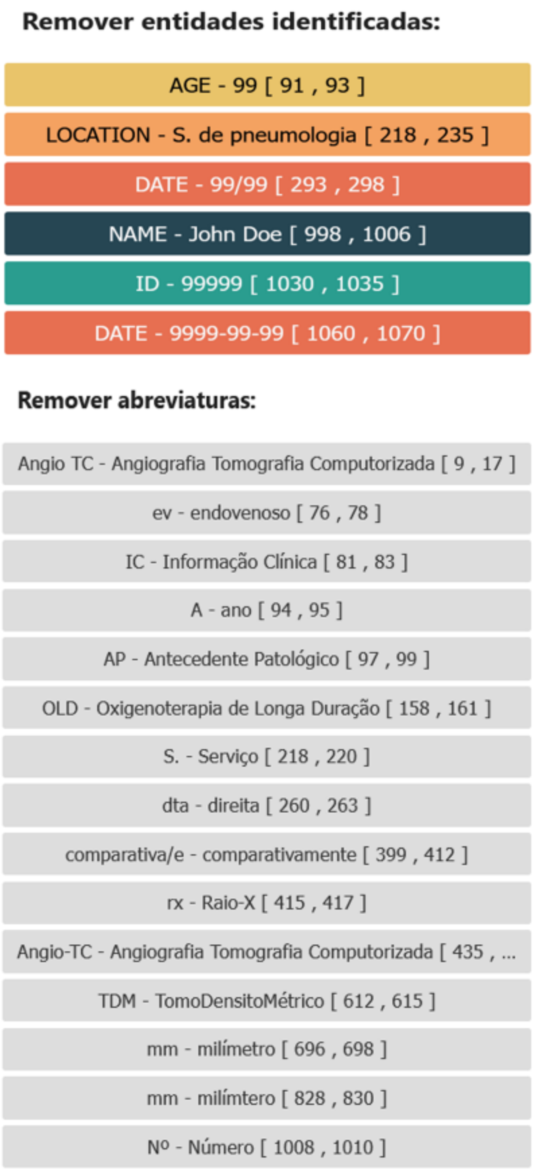
On the right side of the text, two lists of buttons, ordered by their text position, allow the removal of identified data. Each button shows an identified entity or abbreviation, with the corresponding text, start and end positions. The entities removal buttons also indicate the category and have the same category background colour used in the text. In the abbreviations, the description is also shown. Each of these listed buttons enabled the removal of information with a single click. Figure 3.2 displays the list corresponding to the annotations in the text in figure 3.1.

Upon loading a data set, the application generates auxiliary lists with previously identified entities, except for the age category. These lists are ordered by the number of characters in the identifying text, from largest to smallest, to guarantee that longer texts are identified first.

Abbreviations and their corresponding description are also listed and ordered first by abbreviation text size and then description frequency, from the most to the least common. Abbreviations with only one character and descriptions with a single question mark (unknown descriptions) are not added to this list.

When a text is loaded the application iterates these ordered lists and automatically identifies possible entities or abbreviations.

Afterwards, regular expressions automatically identify possible dates and ages: first, all cases of 3 groups of one or more digits separated by one of the characters - backslash



Remover entidades identificadas:

AGE - 99 [91 , 93]

LOCATION - S. de pneumologia [218 , 235]

DATE - 99/99 [293 , 298]

NAME - John Doe [998 , 1006]

ID - 99999 [1030 , 1035]

DATE - 9999-99-99 [1060 , 1070]

Remover abreviaturas:

Angio TC - Angiografia Tomografia Computorizada [9 , 17]

ev - endovenoso [76 , 78]

IC - Informação Clínica [81 , 83]

A - ano [94 , 95]

AP - Antecedente Patológico [97 , 99]

OLD - Oxigenoterapia de Longa Duração [158 , 161]

S. - Serviço [218 , 220]

dta - direita [260 , 263]

comparativa/e - comparativamente [399 , 412]

rx - Raio-X [415 , 417]

Angio-TC - Angiografia Tomografia Computorizada [435 , ...

TDM - TomoDensitoMétrico [612 , 615]

mm - milímetro [696 , 698]

mm - milímtero [828 , 830]

Nº - Número [1008 , 1010]

FIGURE 3.2. Buttons to remove identified entities and abbreviations

\”, period ., hyphen – or a single space - were automatically identified as dates; secondly, also as dates, were considered the same cases of 2 groups. Finally, cases of 2 digits that were not preceded or followed by another digit were marked as age entities. Order is very important in this process because overlapping is not permitted, so the objective was to identify longer occurrences first.

The application easily allows to disable the automatic identification, useful in a first annotation, but maybe not in a review process.

3.3.4. Annotation Process

Using the previous tool, nearly 1,000 *observations* and 1,000 *reports* were annotated with the automatic annotation feature enabled.

The first version of the tool had no automatic identification feature using previous annotated texts, but there were several recurring abbreviations and entities across all texts and the feature, that have been fine-tuned during annotation, proving to be extremely helpful.

The way new annotations are made, where a given string is matched in the entire text, and the existence of automatically inserted entities and abbreviations on text load, made it possible to have a large number of false entities or abbreviations. The reasoning was that it was preferable to remove false entities than to risk missing a true one. This is why it was important to have an easy way to remove annotated entities and abbreviations, like the already described buttons.

Nevertheless, the annotation of an abbreviation like *a* for *ano* (year) resulted in too many false abbreviations listed. A new function was added that, given a text and a list of numbers, would delete all abbreviations or entities matching exactly the given text and those whose starting position was not in the given list.

Even with the automatic feature, this proved to be a tedious and challenging task, specifically the abbreviations annotation, due to the lack of expert knowledge.

The annotation used the catalogues provided by CTC and UMLS search tools as primary resources, but many searches issued no results. Many times, a Google[8] search provided better results than the ones with the specific domain tool, mainly for Portuguese terms. One of the more valuable information sources was the texts themselves, many abbreviations and their descriptions were found in the texts.

After the first round, all texts were reanalysed, starting again with the first text, this time without the automatic identification feature.

3.4. Manual Annotation Validation

To ensure some data consistency, all abbreviations with more than one description were listed, with all texts considered in lowercase. From these, the ones with no given description were reviewed to evaluate if the identified descriptions for the same abbreviation would be suitable. For example, the abbreviation *PCR* could appear as:

[8] https://www.google.com/

```
'pcr': {
    'total': 49,
    'desc': {
        'proteína c-reativa': 32,
        'paragem cardiorrespiratória': 12,
        'polymerase chain reaction': 3
        'polmerase chain reaction': 1
        '?': 1
    }
}
```

For *PCR* we have one record with no description, identified with *?*, so the corresponding text was read once more to check if the other descriptions *Proteína C-Reativa* (C-Reactive Protein), *Paragem Cardiorrespiratória* (Cardiorespiratory Arrest) or *Polymerase Chain Reaction* could be used.

Also, these listings made it possible to detect description misspellings or slightly different description options, like in the *PCR* case and the misspelled description *Polymerase Chain Reaction*.

To accelerate these reviews, another feature was added to the annotation tool that allowed to navigate through a list of given indexes.

The abbreviations' lists with more than one description were reviewed for *reports* and *observations* separately, and then with both text types combined.

3.4.1. Semantic Type Identification and Grouping

To try to further characterize these abbreviations, the identified terms were categorized as given by UMLS semantic type. It is not a thorough classification, it was done using the UMLS Metathesaurus Browser[9], a UMLS Terminology Service.

For the Portuguese terms that were not found, used Google Translator[10] to retrieve an English translation. For example, the abbreviation *BO*, *Bloco Operatório*, was translated do *Operating Room*, and this term is listed with two semantic types: Manufactured Object and Health Care (HC) Related Organization, none existed in the categorized list. It was classified as HC Related Organization because it seemed to better suit *Bloco Operatório* as a HC structure.

The classification was almost entirely done with no context, using the abbreviations lists. Only when the exact term was not found, the full text was read to help in filtering

[9] https://uts.nlm.nih.gov/uts/umls/home
[10] https://translate.google.com/

similar terms in the result list. When a given term had more than one semantic type, the one more common in the already classified terms was chosen. If there was not one more common, the one that seemed to better suit the term was chosen. When a term was not found, if there was a similar term, the corresponding similar semantic type was chosen; for example *CCO*, described as *Consciente, Colaborante e Orientado* (Conscious, Collaborative and Oriented), had no match, but *conscious* appears with the semantic type *Mental Process*, the abbreviation was classified as *Mental Process.* This similar classification was not done often and a total of 210 terms were not classified (46 unique). Also, 559 abbreviations were not identified (with no description) and the 414 corresponding to symbols were not classified.

Inspired by McCray et al. [**68**] and following the semantic groups in the UMLS Metathesaurus Browser, semantic types were grouped aiming to reduce analysis complexity [**68**]. In Appendix refApendix:UMLS we present Table A.1 that shows exactly how the types were classified into eleven groups. This grouping is very similar to the given by the UMLS, with very few differences, namely: 1) *Quantitative Concept* type was considered a single group, due to its weight in the dataset; 2) A group **Other** was considered to include all groups with fewer representation; 3) The types *Patient or Disabled Group* and *Professional or Occupational Group*, that UMLS includes in the group *Living Beings*, were include in the *Other* category, to keep *Living Beings* group with only the types *Bacterium*, *Fungus* and *Virus* that, for the identified abbreviations, presented some unique characteristics.

3.5. Automatic Identification of Entities

Sensitive data annotation made it possible to evaluate automatic identification processes. This evaluation can support decisions on the use of automatic sensitive data identification in future work. This section will describe the use and evaluation of four different automatic processes on the identification of the annotated sensitive data entities, from simple regular expressions to complex specifically trained models using Flair embeddings.

3.5.1. Automatic Identification Processes

To prepare the data for automatic processing, a new pandas dataframe, with all the 2,012 annotated texts, was set up with the following data columns:

- *Filename*, the original filename;
- *Text*, the original text;

- *Entities*, list of annotated entities;
- *Abbreviations*, list of annotated abbreviations;
- *Text type*, identification if the text is an *observation* or a *report*;
- *Text with abbreviations*, the original text where identified abbreviations with a description, has been replaced with the corresponding description;
- *Entities altered*, list of annotated entities where starting and ending positions have been altered to match the new texts with the abbreviations' descriptions.

The tested automatic processes vary in complexity and goal, table 3.4 shows, for the different tested processes, the corresponding sensitive data categories.

Table 3.4. Automatic processes sensitive data categories

Automatic Process	Name	Date	Location	Identification Number
Regular Expressions	X	X	-	X
Presidio	X	X	X	-
Stanford De-Identifier	X	X	X	X
No-Harm Anony	X	-	-	-

To be used as a baseline result, the first approach for automatic identification was using the same regular expressions already used in the annotation tool to propose possible dates.

To build a baseline for the identification numbers, all five-digit numbers were labeled as clinician identification numbers, and all seven-digit numbers as patient identification numbers.

For patient and clinician names, a baseline was also set up where regular expressions were used to identify all sequences of words where the first letter was capitalized and all the following were not, ending with a not capitalized word or a character that was not a letter. For this baseline, all names that followed the expression *Report validated by:*, were also automatically identified as names.

Some experiments were done using *Microsoft´*s data protection and de-identification Software Development Kit (SDK), *Presidio*, in particular the analyzer module, with the default PII recognizers and *spaCy*'s Portuguese news and media trained pipeline, *pt_core_news_lg*[11]. These experiments involved using both the original texts and the texts where identified abbreviations were replaced by their given descriptions.

Standford de-identifier is a transformer model trained with English clinical narratives, including radiology documents, for de-identification purposes [**52**]. The model was used to identify PII in the texts using the transformers pipeline[12], with all the default settings, except the *TokenClassificationPipeline* that was set to use the *simple* aggregation strategy, labeling the results as *entity group*.

Also, *NoHarm-Anony - De-Identification of Clinical Notes Using Contextualized Language Models and a Token Classifier*[13] [**57**], was used to identify names. This is a model based on FlairBBP trained specifically with Portuguese clinical narratives.

3.5.2. Automatic Identification Evaluation

Precision, recall, and F-measure are common measures to analyze automatic de-identification systems results. Precision and recall measure the weight of the number of corrected identified records, but precision compares it with the universe of all predictions, and recall compares it with the universe of all true values. The F-measure balances both measures, taking their harmonic mean [**2**]. In this work, evaluation is performed with two perspectives: at the level of each identified PHI, and at the token level.

The first approach was to evaluate the models considering PHI as a whole. In this context, a correctly identified record, or a True Positive (TP), for a given set of similar categories, is an annotated entity that is also present in the given result set: with the same start and end values and on given lists of similar categories. For example, let's consider the labels "NAME_CLINICIAN" and "NAME_PATIENT" (present in the annotated set) similar to the entity type "PERSON" (in the result set). A TP will be a record with the same start and end positions in both and labeled as "NAME_CLINICIAN" or "NAME_PATIENT" in the annotated list and "PERSON" in the automatic result. False Negative (FN) a record in the annotated set that is not in a TP and, similarly, a False Positive (FP) is an entry in the automatically detected entities that are not a TP. This means that a result will be a string that can contain one or more tokens.

All automatic processes output results were in the form of complete PII, except for the Standford De-Identifier. For the Stanford model, although set up to aggregate results,

[11] https://spacy.io/models/pt#pt_core_news_lg

[12] https://huggingface.co/docs/transformers/main_classes/pipelines

[13] https://huggingface.co/noharm-ai/anony

tokens were not being grouped as expected, resulting in fully identified entities counted as FN, and the tokens in these counted as FP.

Another evaluation considering token count was set up. After NLTK tokenize, using the Portuguese language parameter, to tokenize each given text, a token was counted as a TP, for a given set of similar categories, if it was part of an annotated entity and part of an entity in the model output. If it was not part of an annotated entity but part of an entity in the model result, it was counted as a FP, and FN is a token part of an annotated entity but not part of an entity in the model output.

The used metrics, for complete PHI and token counts, are as follows, where # stands for the cardinally of each set:

$$\text{Precision: } P = \frac{\#TP}{\#TP + \#FP} \tag{3.1}$$

$$\text{Recall: } R = \frac{\#TP}{\#TP + \#FN} \tag{3.2}$$

$$\text{F-Measure: } F = \frac{2\,P\,R}{P + R} \tag{3.3}$$

For each automatic process, and each pair of lists of similar categories, precision, recall, and F-measure were calculated using both complete PII and token counts, for the entire dataset and each text type. To facilitate this process, TP, FP, and FN were first calculated and saved text by text and a final metrics dataframe kept all metrics registry with the data columns:

- *Date*, the date when results were calculated;
- *Model*, the automatic process used;
- *Text type*, with the types considered: *observations*, *reports* or both;
- *Count type*, PII or tokens;
- *Labels*, identification of the categories being evaluated;
- *Total annotated*, number of annotated entities in labels;
- *Total model*, number of entities identified by the automatic process in labels;
- *TP*, number of TP;
- *FP*, number of FP;
- *FN*, number of FN;
- *Precision*, calculated precision metric;
- *Recall*, calculated precision metric;

- *F1*, calculated F-measure metric;

After the results inspection, it was noted that some processes, like *Presidio* did not consider years and days as single a PII, while the Standford De-identificator did. For dates, a new evaluation was done where all PII made up of only years or days were ignored. This was done using a regular expression that filtered out, from all entities, single numbers from 1 to 31 and from 1900 to 2029.

3.6. Chapter Conclusions

The previous sections are a detailed description of all executed tasks, from the original raw data to the calculation of the metrics used to evaluate the automatic PII identification processes. These include data exploration, the development of a customized annotation tool, the manual annotation of text to identify personal data and abbreviations, and the use of automatic classification processes to identify personal data, with various complexities, from simple custom functions using regular expressions to specialized complex ML models.

Also, the previous sections, portray the more debatable decisions made, hopefully providing reasonable justifications for the actions taken, like when describing the strategies to match exam types with the terms in the CTC catalogues.

The next chapter will describe the outputs from these tasks, keeping with their chronological order, putting emphasis on text characterization and found personal data.

CHAPTER 4

Results

A total of 2,021 texts were manually annotated, 1,014 observations, 21% of all *observations*, and 1,007 *reports*, 15% of the total. These *observations*, randomly chosen, revealed a text length average of 251.3 characters, and 42.5 token count. Annotated *reports*, also shuffled randomly, have 1,509.7 and 237.4, respectively, very similar, only slightly longer, to the values for all unique texts in the original dataset, as previously shown in table 3.2.

In the annotated texts, 3,866 possible PII were identified, and 12,202 abbreviations. Based on this annotation, this chapter will describe both text types, highlighting similarities and differences.

The identification and characterization of the found PII entities, made it possible to assess different automatic methods for the identification of personal data. The ending section will focus on the results of this assessment.

In this chapter, data will be described not only in absolute numbers, to illustrate the total of annotated text, but also with the values for 1,000 texts, to compare values per text, because there were slightly more *observations* than *reports* analysed, and values for each 500,000 characters, to enable the comparison between *reports* and *observations*, very different in text length. The number 500.000 was chosen to approximate the values per 1,000 texts, hoping it would make the information more apprehensible.

4.1. Identified Sensitive Data

Reports include 63% of all identified entities, but, if we take into consideration text length and count the number of identified entities per character, they represent only 22% of all annotated entities, which may suggest their more technical nature when compared to *observations.*

Figures 4.1 and 4.2 illustrate the number of entities identified, distinguishing text type and used categories.

The figures effectively convey the differences when considering values per text and per character, due to the large difference in length for the two narrative types. The figures also reveal the prevalence of the identities in *observations*, and, another difference

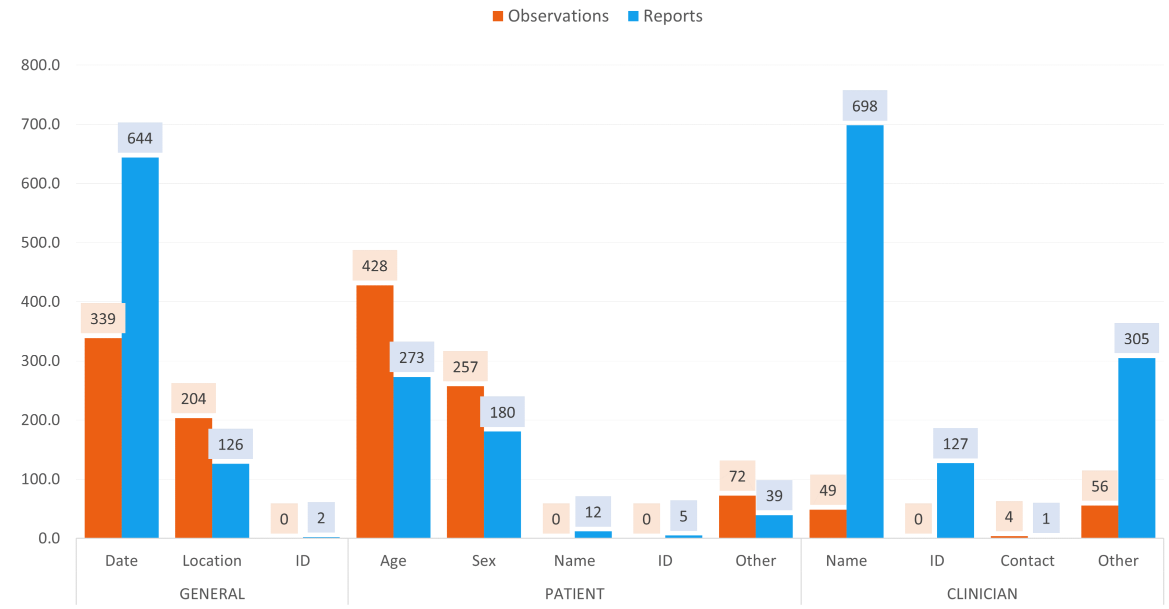


FIGURE 4.1. Identified entities in 1,000 texts

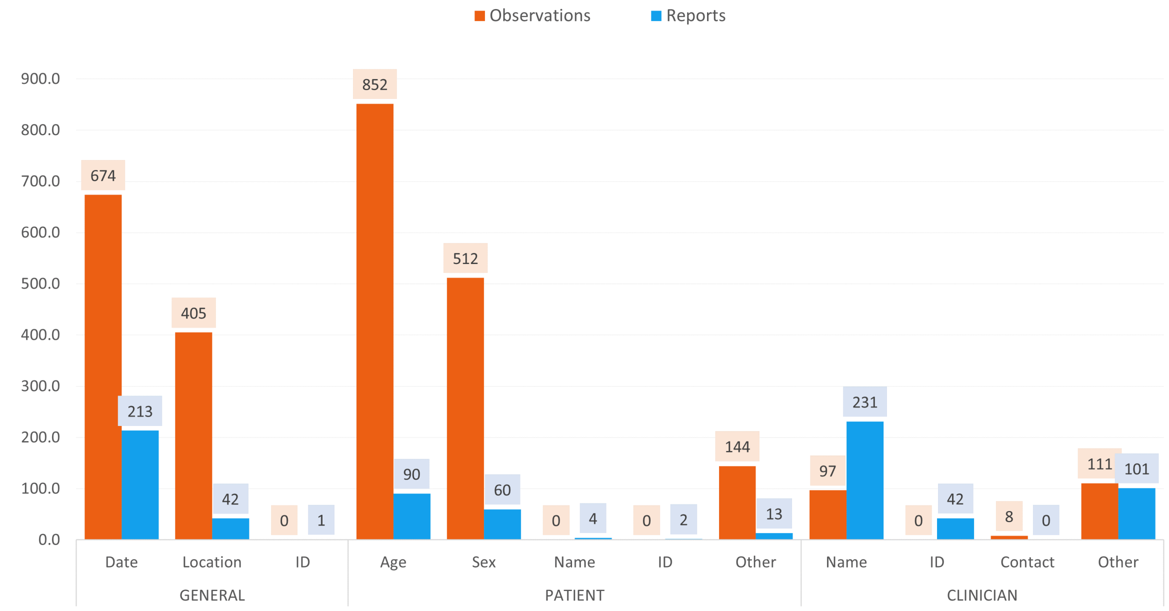


FIGURE 4.2. Identified entities in 500,000 characters

between entities in *observations* and *reports*, is the fact that *reports* have more clinician data, confirming the idea that these are very different texts.

Dates are the most common entities being 25.7% of all identified texts, followed by the names (19.6%) and patient's age (18.3%). Table 4.1 shows, per text type, not only the absolute number of identified entities in each category but also their representation per text and per text length.

Table 4.1. Number of identified entities

Type	Label	Total		In 1,000 texts		In 500,000 characters	
		N	%	Observ.	Reports	Observ.	Reports
GENERAL	Date	994	25.7	338.6	644.0	673.7	213.3
	Location	333	8.6	203.6	126.2	405.0	41.8
	ID	2	0.1	-	2.0	-	0.7
PATIENT	Age	708	18.3	428.0	273.2	851.5	90.5
	Sex	442	11.4	257.2	180.5	511.7	59.8
	Name	12	0.3	-	11.8	-	3.9
	ID	5	0.1	-	4.9	-	1.6
	Other	113	2.9	72.5	39.6	144.2	13.1
CLINICIAN	Name	758	19.6	48.7	699.2	96.8	231.6
	ID	129	3.3	-	127.2	-	42.1
	Contact	5	0.1	4.0	1.0	7.9	0.3
	Other	365	9.4	55.6	304.7	110.6	100.9

4.1.1. Date Category

From the 341 dates found in *observation* texts, 107 (31.4%) do not correspond to strings matching the regular expression used in the annotation tool. The same is true for only 13% of all dates in *reports*.

Dates not matching the regular expression have varied forms, all with non-numeric characters (except one case with added extra spaces), and most of them with the month name fully written or abbreviated. Other cases include periods like *início de dezembro* (beginning of December) or two dates as in *26 e 30 de Abril* (26 and 30 April). In the second case, we could have only considered the 30 April as a date, leaving 26 because it was a day of the month, but the option was always to identify the whole text including as much information as possible, only days with no more context were ignored.

4.1.2. Name Category

Names are very common in the texts, especially physicians' because many *reports* end with the physician's name, clinical speciality, and report date.

The 709 clinician names found correspond to 108 distinct names, a majority with more than one token, with only three findings, corresponding to three distinct strings, made of one first name, identified in the *observations*. In *reports*, the minimum number of tokens in names is two.

Patient names, only found in *reports*, are longer and have more tokens than clinician names, with an average text size of 25.6 characters, while clinician names have an average of 14.8, suggesting that the patient names appear as full names. In this case, a total of 12 patient names correspond to 9 distinct patients.

4.1.3. Age Category

After dates and clinician names, ages' references are the third most common PII category, more common in *observation* than *reports*. If there is no doubt that dates and names constitute sensitive information, HIPAA Privacy Rule only considers PII ages above 89 [**69**] and, from all 709 ages found, only 2 are 90 years old or above.

knowing that the patient's date of birth is registered in the information system, these references seem redundant. With the caveat that the circumstances that led to explicitly stating these patients' data are unknown, it is possible to recommend that patients' known information, like name and age, should not be included in admission notes or clinical reports. The same is valid for all the most common PII found like dates, names, and identification numbers.

4.1.4. Sex Category

Only explicit references to the patient's sex were annotated, in a total of 442. No implicit references were identified, but it is often easy to infer sex from the gender of related words or anatomic descriptions in the texts, raising doubts about the pertinence of this identification. Moreover, HIPAA guidelines do not include sex as a PHI.

A total of 64 abbreviations were found (14.5%), such as *F*, meaning *Feminino* (Female) and *H*, standing for *Homem* (Man). Curiously, the most common abbreviation *M*, identified 32 times, can mean *Masculino* (Male) or *Mulher* (Woman). It was possible to determine that ten of them refer to Male, other ten to Woman, but it was not possible to ascertain the meaning of the remaining 12. There are slightly more explicit references to men (61.1%) than women, but, ignoring the patient's sex distribution, it is not possible to know if this only reflects patient distribution or if there is some trend in the narratives.

4.1.5. Location Category

Geographic or location references are also common, much more in *observations* than *reports*. Per text, 61.7% of identified locations are in *reports*, and 90.6% if we consider text length. Found locations were categorized as country, locality, institution, and service, where services include references to bed and floor numbers. Only three countries and two cities are identified. Figure 4.3 shows the number of locations per text and per category.

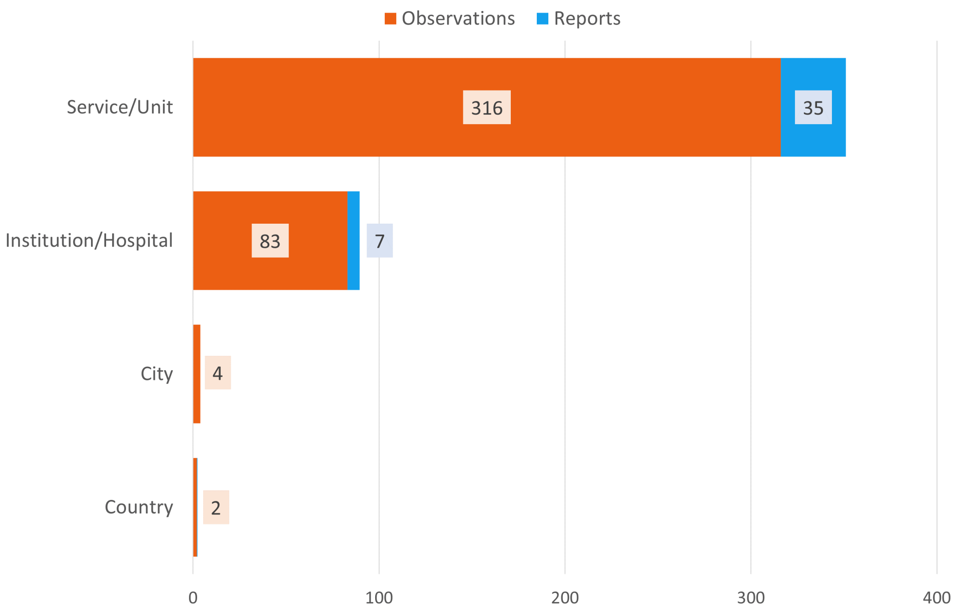


FIGURE 4.3. Identified locations per type

The majority of locations (82.9%) refer services or wards, and most services (79.3%) are referred to as abbreviations, surely well known in this Hospital, but maybe not in other settings.

All 52 identified institutions, corresponding to 23 unique texts, are Hospitals or similar, and also in this case there is a prevalence of abbreviations (84.6%), giving strength to the idea that the inclusion of an abbreviation dictionary could help in an automatic identifier task.

One other relevant characteristic of these locations is that there are a few with a high frequency, for example, the most common service appears 68 times, and the second one 57, in a total of 276 service references.

4.1.6. Identification Number Category

A total of 136 identification numbers were found in the texts, all of them in *reports*, once again showing the differences in these texts.

Of the 129 identification numbers found, 35 correspond to unique doctors' registration numbers. These are 5-digit numbers where the first digit ranges from 2 to 6 and are often explicitly identified at the end of the *report* following the clinician´s name.

There are 117 *reports* ending with the same text format, similar to the one in figure 3.1, from 34 different physicians, suggesting that the texts are inserted following a predefined rule, most probably automatic, that maybe could be removed, and can easily be automatically identified:
Report validated by: *doctor's name* (Doctor's Order Number: *DDDD*)
Report validated on: *YYYY-MM-DD HH:MM*

Patients' identification numbers are all seven digits long, not matching the nine digits of the Portuguese Health Service number, or any other broadly used Portuguese identification number, such as citizen or fiscal numbers. The numbers were labeled as patient identification because they appear in the texts always following patients' names.

The other two possible identifiable numbers, included in the general broader category, are a pharmaceutical batch number and a study reference. Ignoring what type of information can be inferred from both, opted to to include them in this review.

4.1.7. Contact Number Category

There were four contact numbers found in *observations*, all with five digits, probably internal telephone extensions, and one standard 9-digit Portuguese telephone number in the *reports*.

Of the 5-digit contacts, all but one have 9 as the first digit, making it possible to differentiate them from the physician's identification number, also 5 digits long. The one 5-digit contact starting with a 5, similar to a physician's identification number, was labeled as a contact due to text context and because it did not match any of the physicians' identification numbers.

No patient contacts were identified in the texts.

4.1.8. Other Category

As expected, the most difficult information to characterize is the one in the other category.

For patients, these include physical and behavioural features, such as smoking habits with 28 references, sometimes detailing the number of packs a year, a commonly clinical used metric. Also found were 27 references to obesity, occasionally with qualificatives like severe or an explicit degree, or alcoholism with 12 references.

Arguably, all medical history can be thought of as identifiable information. Annotation tried to identify information that could be sociably recognizable, meaning that could be identified by a neighbor or co-worker.

There are 22 references to previous incidents, in most cases that led to current Hospital admission, with different levels of detail, such as the ingestion of an infusion with a specific type of mushrooms, a surfing accident, self-harm behavior, or a fall in a public space. One can argue that some can be specific enough to identify the patient.

There are eight references to recent births and pregnancies. Three *observations* state that the patient is included in a specific study, two mention the patients' professions and two *reports* have patients' height and weight.

Two *reports* describe less common clinical family history, one of them stating a familiar's age at the time of a specific procedure. Other possible identifiable single references include general descriptions like "raça branca" (white race) or more specific ones like single-leg amputee.

In the other category, for the physicians, the vast majority of annotated texts refer to the clinician's clinic specialty, that could fit in a possible profession category. The two texts that are not medical specialties are similar in nature, as they refer to professional careers: the abbreviation *TSDT* appears 5 times and stands for *Técnico Superior de Diagnóstico e Terapêutica* (Senior Diagnostic and Therapeutic Technician), and there is one identified abbreviation for "Professor" as "Prof.".

4.2. Identified Abbreviations

4.2.1. Identified Abbreviations Description

A total of 12,202 abbreviations were identified consisting of 1,254 distinct values.

To make reading easier, in this work, the annotated abbreviations will appear in their most common form, and the corresponding descriptions, or long form, with the first letter of any word in upper case. The original texts don't necessarily follow this format. For example, *Tomografia Computorizada* (Computed Tomography) appears in diverse forms in the original texts, like *TC*, *Tc*, and *tc*.

This means that a string abbreviation can have different descriptions, depending on its context, and the same description can appear in different annotated forms. A distinct count will be a distinct pair of annotated text and given description, both strings considered in a case insensitive comparison.

Again, there are obvious differences between the two text types. *Observations* have fewer identified abbreviations in total (4,499, 36.9%) but, when considering text size, abbreviations are far more common in *observations* representing 77.9%.

There are two abbreviations much more frequent than all the others: *mm* with the description *Milímetro*(Millimeter), identified 1,225 times and *TC* as *Tomografia Computorizada* (Computed Tomography), with 1.184 occurrences.

All abbreviations described as *Millimeter* appear as *mm*, but the description *Tomografia Computorizada* (Computed Tomography) appears a total of 1.291 times, included in the descriptions of texts such as *TAC* and *T.C.*. Also, the abbreviation *CT*, described as *Computed Tomography* has 2 occurrences. Moreover, because the annotation process followed the principle to identify the longer found text, *mm* and *TC* are included in other abbreviations. There are more 117 annotations that include *mm*, such as *mmHg* (Millimetre of Mercury), or *mm2* (Square Millimeter); and more 317 abbreviations with *TC* or *CT*, such as all the 308 abbreviations standing for *Angiografia Tomografia Computorizada* (Angiography Computed Tomography), that include the identified texts *Angio-TC*, *Angio TC*, *AngioTC AngioTAC* or *(Angio)TC*, among others.

One curious fact is that, although, as described above, when considering text size, almost 78% of all abbreviations found are in *observations*, the *mm* abbreviations are 82.7% in *reports* (96.7% in absolute values). *TC*, described as *Tomografia Computorizada*, follows the other abbreviations on the dataset with 73.6% in *observations* when considering text size.

As seen, *Millimter* or *Computed Tomography* is the most common abbreviation, depending on the counting strategy. The two are far more frequent than the third more common text that is *cm*, standing for centimeter, highlighting the prevalence of length units.

The descriptions given were also marked as being in Portuguese, English, or other, keeping in mind that the description language option followed the rule that whenever the original abbreviation matched the Portuguese language then Portuguese was chosen for the corresponding description. The descriptions that are not classified as Portuguese or English are Latin expressions, such as *bid*, *bis in die*, a common medical expression that means twice a day, scientific names as *Staphylococcus* or universal unit symbols like *mm*.

The abbreviations with no description and symbols were not classified as language is concerned. From a total of 11,229 abbreviations, 7,255 (64.6%) are in Portuguese, 1,414 (12,6%) in English, and the remaining 2,560 are a majority of universal units (2,468), but also include bacteria and fungi names.

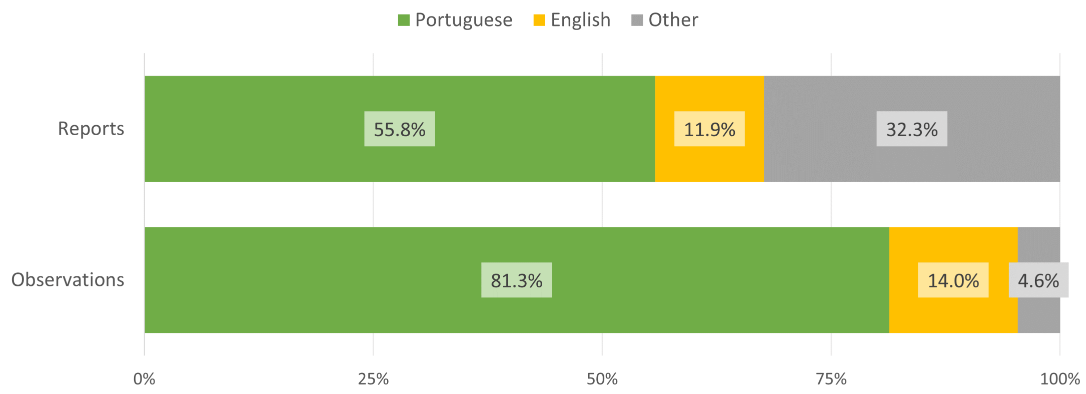


FIGURE 4.4. Identified abbreviations per language and text type

As seen in figure 4.4, the language distribution reflects the stronger use of universal units in *reports*, with 32.3% of abbreviations in these texts not classified as Portuguese or English, hinting that *reports* are more technical narratives.

4.2.2. Abbreviations Semantic Groups

A total of 58 semantic types were identified, categorizing a total of 11,019 abbreviations, and these were grouped into 11 semantic groups. Table 4.2 depicts the absolute number of identified abbreviations per semantic group and the values per 500,000 characters and per text type.

Of the 210 terms not categorized, most of them are not clinical terms. The most common text with no category is the term *vs* meaning *versus*, with 64 records, but there is also a common clinical term *Parâmetro Inflamatório* (Inflammatory Parameter) that appears as *PI* and *PInf* both in *observations* and *reports*, that remained uncategorized. Probably a domain specialist could have found a similar term in the UMLS Metathesaurus.

Symbols are the group with a stronger presence in *observations*, reinforcing the idea that *observations* are more informal and less technical than *reports*. The most common symbol found was + representing the conjunction *and*, like in the text *Doente com LLA + tonturas* (Patient with ALL + dizziness), and also *positive* as in "COVID +". Also / is used as either the conjunctions *and* and *or*, like in *exclusão de patologia/hemorragia intracraniana* (exclusion of intracranial pathology/hemorrhage). The characters -> are frequently used to illustrate relations in the text, as in *COVID + -> doente intransportável* (COVID + -> untransportable patient), or*11->17->21 (PCR)*, highlighting the parameter evolution.

Table 4.2. Identified abbreviations, per semantic group

Semantic group	All texts			Observ.	Reports
	N	%	Unique	(*)	(*)
Procedures	2,919	23.9	177	2,104.1	605.5
Quantitative Concept	2,725	22.3	81	304.3	839.7
Disorder	1,690	13.9	179	1,878.9	241.4
Concepts & Ideas	1,223	10.0	146	987.8	236.1
Anatomy	819	6.7	129	661.9	158.1
Living beings	442	3.6	29	521.6	58.1
Physiology	386	3.2	54	142.2	102.6
Chemicals & Drugs	332	2.7	85	420.8	38.9
Organizations	276	2.3	35	345.7	33.0
Devices	154	1.3	32	55.3	42.7
Other semantic type	53	0.4	20	69.1	5.9
Symbol	414	3.4	35	588.8	37.9
No semantic type	210	1.7	46	235.1	29.7
No description	559	4.6	287	572.9	87.9
Total	12,202		1,254	8,584.3	1,676.2

(*) Values in 500,000 characters

Many symbols are simply used as visual references, like the common use of # as a bullet point in lists. There are some cases where # can also be used in trauma references, again, maybe a clinical specialist could have identified these cases.

It is curious how text type distribution varies within these semantic groups, as seen in the abbreviations text type distribution per semantic group, taking into consideration text size, depicted in figure 4.5.

Procedures is the most common semantic group, 23.9% of all data, as shown in table 4.2, and its distribution is almost identical to the global distribution, which may be expected from a very representative group. But, quantitative concepts, 22.3% of the

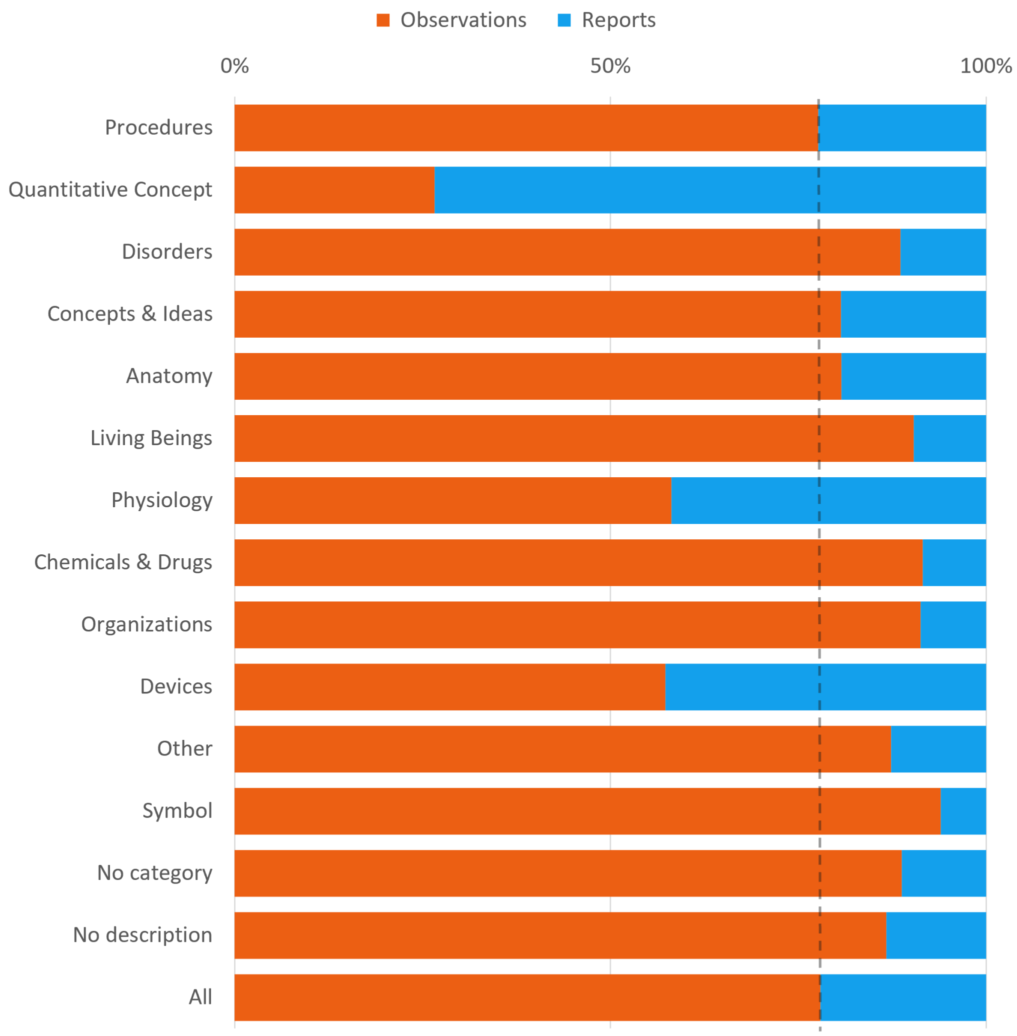


FIGURE 4.5. Identified abbreviations distribution per semantic group, considering text size

total, are much more common in *reports*. Only *Physiology* and *Devices* also exhibit a *report* representation larger than the total.

Procedures includes the types *Diagnostic, Therapeutic or Preventive and Laboratory Procedures*, that correspond to 27.2%, 7.1% and 4.9% of the total of identified abbreviations, respectively. It also includes the types *Molecular Biology Research Technique*, *Research Activity* and *Health Care Activity*, but these are rare, with a combined total of sixteen records, with only seven unique.

As already discussed, *Tomografia Computorizada* (Computed Tomography) is the most common *Diagnostic Procedure* term appearing 1,291 times, and it represents almost half of all procedures (44.5%), followed by *Angiografia Tomografia Computorizada* (Angiography Computed Tomography) with 308 records and *Ressonância Magnética* (Magnetic Resonance) with 115. All other procedures appear less than 100 times. *Extracorporeal*

Membrane Oxygenation, written as *ECMO*, is the most common abbreviated *Therapeutic or Preventive Procedure*, with 72 occurrences, and *Activated Clotting Time*, *ACT*, is the most frequent Laboratory Procedure abbreviation with fourteen.

The *Quantitative Concept* type is the most common one and, if we evaluate diversity as the ratio of distinct values to all values, it is also the least diverse of the types, considering only the types with more than 10 records. Figure 4.6, depicting diversity scores for all semantic groups, illustrates how this type's diversity stands out from the others, even when compared with the other groups with very frequent abbreviations like *Procedures*.

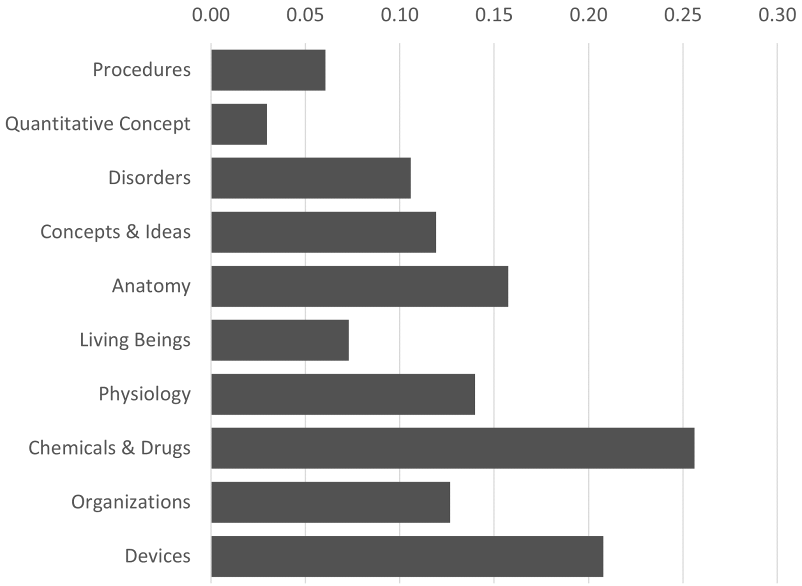


FIGURE 4.6. Identified abbreviations diversity

For almost all quantitative concepts, the identified abbreviation is a universal unit symbol (90.6%). It is questionable if these universal units should even be considered abbreviations. Nevertheless, intending to understand these texts, it seemed useful to also identify and characterize them, thinking that the information can be useful for future tasks.

The use of the universal units also contributes to the lack of diversity in this group, making it harder to have different annotated texts for the same description, *Centimeter* is always noted as *cm*, which does not happen, for example with the procedure *Tomografia Computorizada* (Computed Tomography), as already discussed.

Disorders include 10 different semantic types, being *Disease or Syndrome* the most represented with 67,6% of all values. Also, the four most common disorders are of the type *Disease or Syndrome*, and these four make 67.7% of all *Disease or Syndrome* type:

Severe Acute Respiratory Syndrome, with 336 records, *Coronavirus Disease*, with 264, *Tromboembolismo Pulmonar* (Pulmonary Thromboembolism), with 208, and *Acidente Vascular Cerebral* (Stroke), with 67. Only the fifth, with 50 records, *Adenocarcinoma*, is of the *Neoplastic Process* type.

Disorders and *Devices* are the groups with more English terms, mainly due to the abbreviations *COVID* and *SARS*, and several pacemakers' references, like *CRT-D* as *Cardiac Resynchronization Therapy Defibrillator.*

UMLS *Concepts & Ideas* group includes the types *Quantitative Concept*, *Spatial Concept*, *Intellectual Product*, *Conceptual Entity*, *Qualitative Concept*, *Temporal Concept*, *Regulation or Law*, *Functional Concept*, *Idea or Concept.* For this analysis, it was decided to keep *Quantitative Concept* as a separate group due to its size and particular features.

In *Concepts & Ideas* group, the most common abbreviated term, with 303 records, is a Portuguese form to refer to medical professionals, used as a name prefix: *Dr.*, or the feminine *Dra.*; in these texts also appears as *Dr*, *Dr.$^{\underline{a}}$* or *Dra*, among others. It stands for *Doutor*, for males, or *Doutora*, for females, meaning doctor. These prefixes can help identify clinician´s names and distinguish them from patient names.

Also common are the temporal concepts *hour*, written as *h*, with 117 records, and *day* as *d*, with 89 records. Also frequent, with 106 annotated abbreviations, is the term *Antecedente Patológico* (Pathological History) from the semantic type *Intellectual Product.*

The most common *Anatomy* terms are the ones that often follow a diagnostic procedure, like a computed tomography or a MRI: the *Body Part, Organ, or Organ Component Cranioencefálico* (Cranioencephalic), with 200 abbreviations found, and the *Body Location or Region Tórax, Abdómen e Pélvis* (Chest, Abdomen and Pelvis). The three terms included in the latter description also appear abbreviated in different combinations, like *Abdómen e Pélvis* or alone, in a total of 105 times. These represent 25.0% of the total of the *Anatomy* terms. There are also various abbreviations standing for arteries, a total of 101, with more or less detail, like plain *Artéria* (Artery) to *Artéria Femoral Comum Esquerda* (Left Common Femoral Artery).

4.3. Sensitive Data Automatic Identification

Automatic identification of sensitive data can be a very helpful feature, for example, to produce suggestions for a human annotator. The previously developed toll only used regular expressions to identify dates. But, with the goal of identifying other possible suggestions, a variety of automatic identification processes were used and evaluated.

4.3.1. Regular expressions

The first automatic identification processes tried, using regular expressions, identified, in the original texts, a total of 1,591 dates, 1,597 names, and 148 identification numbers. For names, using regular expressions, there was no distinction between patient names and names of non-patients, but for the identification numbers it was assumed that the clinician's identification numbers are 5 digits long, and the patient´s identification numbers have 7 digits.

4.3.2. Presidio

Presidio models identified possible PII in the original texts in the following categories: *PERSON*, *LOCATION*, *DATE_ TIME*, *URL* and *PHONE_NUMBER*. For evaluation purposes, these labels had to be mapped to the annotated categories.

The 5,301 *LOCATION* entities were mapped to the corresponding location annotated label, although there were only 333 Locations in the annotated set. The model identified as location many abbreviations, such as *ECMO* or *COVID-19*. Even considering the texts where the abbreviations were replaced with their corresponding long description, the model output had 4,716 Locations, with many unexpected FP results like *Escherichia* or *Celcius*. *Presidio*'s documentation[1] describe these default locations as politically or geographically defined, such as cities, countries but also mountains or bodies of water, among others, but the default PII categories are set up for the English language.

The *PERSON* category, with 3,003 records in the result set, which is supposed to include full names (first, middle, and last names, or initials) was matched with both name labels, patient and clinician, that had only a total of 721 entities. The majority of FP results, like *Doença de Crohn* (Crohn's disease) can be explained due to language and domain specificities, and there are also misspelt Portuguese words as FP results, like *justiifuqem*, probably standing for *justifiquem* (justify).

The 617 *DATE_ TIME* entities were compared to the 994 annotated dates. *Presidio*'s results, as in the annotated set, do not include single years or single days and exhibit a precision of 1 in *observations* narratives and 0.9967 for all texts. The annotated dates not identified by *Presidio* are the ones with months in Portuguese, like *agosto de 2020* (August 2020) or dates with no year, like *04.Mar* or *21/07*. A curious fact is the recall score gap when comparing *reports*, with 0.6139, and *observations* with 0.2857, an extremely low score. When compared to the regular expressions baseline, despite the low sensitivity,

[1] https://microsoft.github.io/presidio/supported_entities/

Presidio's results still perform better with an F-measure of 0.7584 versus the 0.6369 baseline score.

Almost all the 15 entities identified as *PHONE_NUMBER* are dates, and none corresponds to the annotated clinicians´ contacts. Also, no URLs were present in the annotated text. The 15 entities identified as *PHONE_NUMBER* and the 73 labeled as *URL* were ignored.

Table 4.3. Relevant automatic identification F-Measures for all texts

Annotation label	Date	Name		Identification number		
		Patient	Clinician	Patient	Clinician	Other
Regular expressions	0.6264	0.6264		**1.0000**	**0.9520**	*n.a.*
Presidio	0.6630	0.5007		*n.a.*	*n.a.*	*n.a.*
Stanford De-Identifier	**0.8616***	*n.a.*	0.6141	0.8584		
NoHarm Anony	*n.a.*	**0.9903**		*n.a.*	*n.a.*	*n.a.*

* single years and days removed

4.3.3. Stanford De-Identifier

The output from the *Stanford De-Identifier* model had 6 different PII categories. The most common, and with the best performance, was *DATE*, with 969 records and an F-score of 0.8616. The Stanford result set included single days and years as dates, but these were not in the annotated set by choice. To compare results, all PII consisting of single years or days were removed, leaving a total of 858 dates in the filtered output set. Similar to *Presidio*'s results, the 200 FN dates are almost entirely dates with Portuguese months´ names and numeric dates with no year, but there are also some FN dates with the usual European format *yyyy-mm-dd* or *dd/mm/yyyy*. From a total of 64 FP result tokens, 6 are the string *janela* (window) and the 2 *marcha*(march or walk), perhaps mistaken for misspelt months *January* and *March*. Once again, the *reports*' results, with an F-measure of 0.8747 are better than the *observations*' score of 0.8241.

The *Stanford De-Identifier* model classified 112 tokens as identification numbers. These were compared to all identification numbers annotated (patients, clinicians, and others), in a total of 136 tokens. It failed to identify several 5-digit numbers, many of them written with the *M* letter as a prefix. The model did not perform better than the regular expressions, but results have to be analyzed taking into consideration that the used regular expressions were custom-made to fit the found identification numbers after text annotation.

Healthcare Workers (HCW) is a *Stanford De-Identifier* category that, with a more suitable name, identifies the same type of information of the 758 strings annotated with the label *clinician name*. A total of 524 *HCW* entities were found. When considering text types, although with a similar F-measure for *observations* (0.6544) and *reports* (0.6101), the results are very different for precision and recall: *observations* have a sensitivity of 0.9271, and *reports* texts a precision of 0.8040.

The identified entities in the categories *HOSPITAL*, with 82 records, *PATIENT* with 15, and *PHONE* with 5 did not show any resemblance to the annotated data and were thus ignored.

4.3.4. NoHarm-Anony

The *NoHarm-Anony* only identifies names, but with very good results. This is the only language and domain specif model tested and it identified 758 names that were compared to the 770 names in the annotated dataset. Considering token count, only 23 token results were FN, and, in those, twelve records correspond to a first name with only an initial and a period and four names are not traditional Portuguese names. Thirteen FP token results were detected, ten of them being the strings *Homem* (Man) or *Mulher* (Woman). For all texts, using token counting, *NoHarm-Anony* scored a final F-measure of 0.9903, with 0.9447 for *observations* and 0.9928 for *reports*, with sensitivity higher than precision.

4.4. Chapter Conclusions

From the initial 12,000 files, only 4,834 distinct *observations* and 6,771 distinct *reports* were considered for the following analyses.

Observation and *report* narratives are very different texts in many aspects: size, structure, and content, specifically PII content.

For the most commonly identified PII categories, the majority of the found information seems redundant, because it is registered in the information system, like patients' names, birthdays, and identification numbers, as well as clinicians' names, identification numbers, and specialities. Maybe some of this information was written down to be highlighted in a given clinical context, but many times it seems a question of convenience or habit.

Identified dates also seem often redundant. Dates are used as references to previous exams that are registered in the system, or to state the date of the occurring exam. Maybe the exam date could be omitted and references to previous exams could be made using an internal surrogated number with no meaning outside the system scope.

Abbreviations analysis reinforces the difference between the two text types and, in combination with semantic type classification, presents suggestions for information to be extracted from the narratives, like findings measures or specific pathologies.

Reports have more relevant metrics calculated because *observations* don't have identification numbers or patient names. The fifteen relevant metrics evaluated for *reports*, considering original texts and token evaluation, have an average F-measure score of 0.7904, and the ten relevant metrics for *observations* only have 0.5937. This can be explained by the more structured and formal text in *reports*.

Although there are significant differences in performances for *observations* and *reports*, that is not the case for all the best-performing models, as shown in figure 4.7. It can be concluded that the high performance is consistent in both text types.

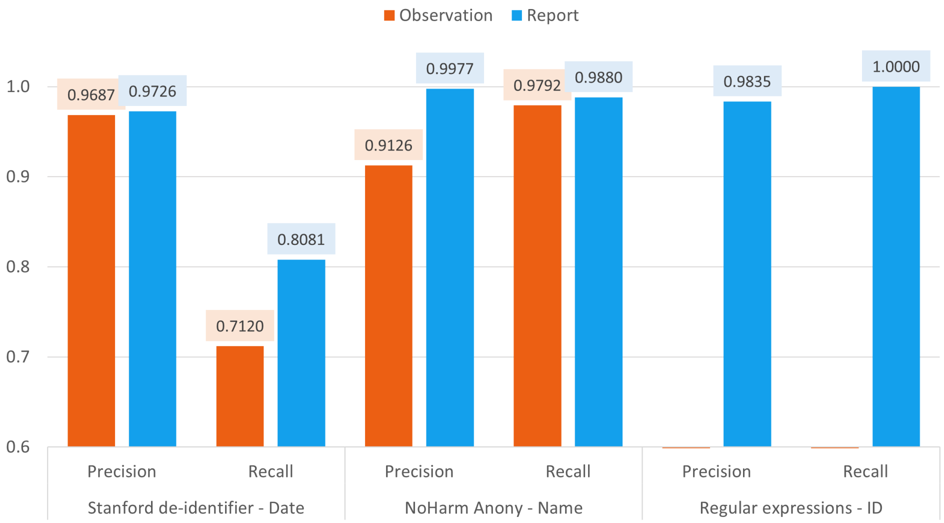


FIGURE 4.7. Automatic identification scores

The use of the text with the abbreviations replaced with their corresponding descriptions did not significantly alter any of the identification processes' performance. The exception is, as expected, the regular expression in name identification that performed much poorly in the replaced text. This regular expression searched for sequences of one or more words with the first letter capitalized, and that is exactly the abbreviations' description format.

CHAPTER 5

Conclusions

5.1. Main Conclusions

Of the 12,000 original files obtained from a Portuguese Hospital Picture Archiving and Communication System (PACS), two different unstructured text types were extracted for further evaluation: 4,834 distinct *observations* and 6,771 distinct *reports*.

Admission notes, or *observations*, are short and condensed narratives focusing on patient clinic history. *Reports* are much longer and more formal, with detailed information on the corresponding image exam or procedure, including findings descriptions, and measures. Both *observations* and *reports* are clinical texts that require specialized knowledge for a corrected interpretation, with many domain-specific terms and abbreviations.

Despite being different narratives, both contain sensitive personal information that should be removed. The identified personal information, although different in nature, seems similar enough in both text types to be treated with similar approaches.

It is also possible to conclude that automatic identification of PII can be used with success in the texts, at least for the great majority of the categorized data. But different PII categories require very different automatic identification approaches. It is evident that automatic personal data identification is not one classification task, but several distinct tasks that vary with the different PII categories.

Regular expressions exhibit very good performance results in recognizing identification numbers, although in evaluating this success one has to keep present that the expressions used were tailored to fit the found numbers. They succeed in these texts and maybe can be used in similar narratives, but there is a possibility that would perform poorly in other text types, even from the same Hospital.

The No-Harm Anony results for name identification are remarkable, and it would be interesting to explore similar options, i.e. models specifically trained for this task with in-domain Portuguese texts, for the identification of other PII categories, like locations, one of the categories with the lowest performance scores.

The main issue is probably the fact that the manual PII identification was not reviewed. Nevertheless, it is arguable that the automatic processes results can serve as a form of review, considering the following FP values examination, and generate some confidence in the subsequent work.

5.2. Main Contributions

Decision making can only be responsible if well informed. This report can help support decisions on what purposes can the texts serve, how to handle the existing personal information, and how to proceed in future works.

The work includes a description of the two different narratives. The two text types, even though both can be considered in the clinical domain, are revealed to be very different texts. This conclusion, and the given characterization, can be beneficial for following NLP tasks when deciding methods and text approaches.

The resulting abbreviations dictionary, which proved useful during manual annotation, can be a valid resource for following NLP tasks. Abbreviations' semantic characterization and analysis shone more light on the nature of both text types and suggested the possible extraction of useful features. For example, the prevalence of universal units as *mm* and *mm* in *reports* invites for extracting findings dimensions.

The manual annotation experience, and the results on automatic PII identification, support the idea that the best approach to personal data identification will probably be a hybrid approach, where human annotators and reviewers, preferably with domain expertise, clinical and legal, are supported by automatically generated suggestions. The suggested entities can be identified based on simple regular expressions and more complex models, that, similar to the No-Harm Anony, make use of contextualized embeddings, Bi-LSTM-CRF neural architectures, diverse language models, and should be trained specifically on clinical Portuguese text to identify personal data.

5.3. Limitations and Future Work

This work presented several obstacles and not all of them were completely overcome. Occasionally, the option had to be to get around those obstacles, and these options should be recognized and examined.

The first challenge was to work with data that had not been previously prepared or analyzed. It was raw data exported from a hospital information system with no additional documentation. It was known that it was comprised of reports extracted from a PACS, an information system dedicated to clinical image acquisition devices, with text data and no images. These could eventually support other ML tasks, but there was no insight into exactly what information these files could hold.

Documentation on the original data source and the data exportation process would have been useful. For example, assuming that the original source was a database, table content and information on the tables relations could have easily explained the repeated data throughout the files, and made clear how *exam types*, *observations*, and *reports* relate.

Also, a greater understanding of the entire original information system could make it possible to make contextualized recommendations. For instance, recommend the use of the CTC catalog preferred terms as exam types names, or, because some of the personal information seems to be automatically added to the end of the *reports*, the disabling of some unnecessary features.

Knowing that some texts were truncated, it would be advisable to make a new exportation to warrant the complete data, and, if possible, make accessible the exportation procedure queries.

The use of CTC catalogs as references did not deliver significant results, leading only to the conclusion that it would be beneficial to implement standardized terminology in the original information system. The UMLS tools, although much more helpful, were very limited for the Portuguese Language.

There was also, naturally, a time limitation, and the annotation process was revealed to be extremely time-consuming, even with some automatisms integrated and fine-tuned throughout the process. The initial proposal of having, at least, 20% of all *observations* and *reports* annotated, to guarantee a solid base for possible conclusion extrapolation, was not accomplished. Although 21% of *observations* were annotated, only 15% of *reports* were also reviewed.

The biggest setback was the lack of domain knowledge, in particular clinical expertise. It was not possible to have medically specialized advice during the different phases. Legal expertise on personal data issues would also have been valuable. Review, especially done by experts, clinical and legal, would validate the initial work and add value and weight to the following analysis.

On the legal questions, many questions remain purposely open throughout this report, like the possible non-categorized sensitive data, inviting a more in-depth and knowledgeable discussion.

From the most relevant PII categories defined, location was the more difficult to automatically identify. Knowing that hospital names and their services are the most commonly found locations, a list of Portuguese hospitals and corresponding internal units could improve automatic identification. Due to the prevalence of abbreviations in locations, a list with both long descriptions and abbreviations would be the better solution.

The No-Harm Anony results for name identification are remarkable, it would be interesting to explore similar options, models specifically trained in in-domain Portuguese texts, for other PII categories.

The anonymization process does not end with PII identification. It is also necessary to remove, mask, or replace with pseudonyms the identified data. Unfortunately, it was not possible to explore how these different possibilities would impact the anonymization goal.

References

[1] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization," 2022.

[2] Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[3] P. Lison, I. Pilán, D. Sánchez, M. Batet, and L. Øvrelid, "Anonymisation models for text data: State of the art, challenges and future directions," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203, 2021.

[4] B. Manzanares-Salor, D. Sánchez, and P. Lison, "Automatic evaluation of disclosure risks of text anonymization methods," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13463 LNCS, p. 157 – 171, 2022.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, fourth edition ed., 2017.

[6] G. Ignatow and R. Mihalcea, *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. SAGE Publications, Inc, 2018.

[7] S. Collovini, J. Santos, C. B., T. J., R. Vieira, P. Quaresma, M. Souza, D. Claro, and R. Glauber, "Iberlef 2019 portuguese named entity recognition and relation extraction tasks," pp. 390–410, 2019.

[8] H. Dalianis, *Clinical Text Mining*. Springer International Publishing, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," pp. 6000–6010, 6 2017.

[10] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, and H. Liu, "Clinical concept extraction: A methodology review," *Journal of Biomedical Informatics*, vol. 109, p. 103526, 9 2020.

[11] K. Ethayarajh, “How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings,” pp. 55–65, Association for Computational Linguistics, 2019.

[12] E. T. R. Schneider, J. V. A. de Souza, Y. B. Gumiel, C. Moro, and E. C. Paraiso, “A gpt-2 language model for biomedical texts in portuguese,” pp. 474–479, IEEE, 6 2021.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” 10 2013.

[14] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *arXiv preprint arXiv:1607.04606*, 2016.

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” vol. 1, pp. 2227–2237, 2 2018.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 10 2019.

[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *OpenAI*, 2018.

[19] Y. Si, J. Wang, H. Xu, and K. Roberts, “Enhancing clinical concept extraction with contextual embeddings,” *Journal of the American Medical Informatics Association*, vol. 26, pp. 1297–1304, 11 2019.

[20] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in *Proceedings of the 27th International Conference on Computational Linguistics*, (Santa Fe, New Mexico, USA), pp. 1638–1649, Association for Computational Linguistics, 8 2018.

[21] OpenAI, “Gpt-4 technical report,” tech. rep., 3 2023.

[22] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu, “Deep learning in clinical natural language processing: a methodical review,” *Journal of the American Medical Informatics Association*, vol. 27, pp. 457–470, 3 2020.

[23] D. Garets and M. Davis, “Electronic medical records vs. electronic health records: Yes, there is a difference,” *HIMSS Analytics*, pp. 1–14, 2006.

[24] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural language processing of clinical notes on chronic diseases: Systematic review," *JMIR Medical Informatics*, vol. 7, p. e12239, 4 2019.

[25] L. F. A. de Oliveira, L. E. S. e Oliveira, Y. B. Gumiel, D. R. Carvalho, and C. M. C. Moro, "Defining a state-of-the-art pos-tagging environment for brazilian portuguese clinical texts," *Research on Biomedical Engineering*, vol. 36, pp. 267–276, 9 2020.

[26] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e Oliveira, J. Copara, Y. B. Gumiel, L. F. A. de Oliveira, E. C. Paraiso, D. Teodoro, and C. M. C. M. Barra, "Biobertpt - a portuguese neural language model for clinical named entity recognition," pp. 65–72, Association for Computational Linguistics, 2020.

[27] C. Pearson, N. Seliya, and R. Dave, "Named entity recognition in unstructured medical text documents," in *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, IEEE, 2021.

[28] J. E. Harrison, S. Weber, R. Jakob, and C. G. Chute, "Icd-11: an international classification of diseases for the twenty-first century," *BMC Medical Informatics and Decision Making*, vol. 21, p. 206, 11 2021.

[29] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. Krallinger, "Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources," 2022.

[30] M. Bay, D. Bruness, M. Herold, C. Schulze, M. Guckert, and M. Minor, "Term extraction from medical documents using word embeddings," pp. 328–333, IEEE, 6 2020.

[31] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 1 2019.

[32] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *ArXiv*, vol. abs/1904.05342., 4 2019.

[33] H. Tissot and R. Dobson, "Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in portuguese," *Journal of Biomedical Semantics*, vol. 10, p. 17, 11 2019.

[34] M. Lamy, R. Pereira, J. C. Ferreira, J. B. Vasconcelos, F. Melo, and I. Velez, "Extracting clinical information from electronic medical records," in *Ambient Intelligence–Software and Applications–, 9th International Symposium on Ambient Intelligence*, pp. 113–120, Springer, 2019.

[35] A. C. Coutinho, S. F. Shiguemi, and E. A. Mendonça, "Multilingual information retrieval in thoracic radiology: Feasibility study," *MEDINFO*, vol. 12, pp. 387–391, 2007.

[36] M. Lamy, R. Pereira, J. Ferreira, F. Melo, and I. Velez, "Extracting clinical knowledge from electronic medical records," *IAENG International Journal of Computer Science*, pp. 488–493, 2018.

[37] I. Coutinho and B. Martins, "Transformer-based models for icd-10 coding of death certificates with portuguese text," *Journal of Biomedical Informatics*, vol. 136, p. 104232, 12 2022.

[38] F. Lopes, C. Teixeira, and H. Gonçalo Oliveira, "Named entity recognition in portuguese neurology text using crf," in *Progress in Artificial Intelligence* (P. Moura Oliveira, P. Novais, and L. P. Reis, eds.), (Cham), pp. 336–348, Springer International Publishing, 2019.

[39] F. Lopes, C. Teixeira, and H. G. Oliveira, "Comparing different methods for named entity recognition in portuguese neurology text," *Journal of Medical Systems*, vol. 44, p. 77, 4 2020.

[40] F. Duarte, B. Martins, C. S. Pinto, and M. J. Silva, "Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text," *Journal of Biomedical Informatics*, vol. 80, pp. 64–77, 4 2018.

[41] A. D. Reys, D. Silva, D. Severo, S. Pedro, M. M. de Sousa e Sá, and G. A. C. Salgado, "Predicting multiple ICD-10 codes from brazilian-portuguese clinical notes," in *Intelligent Systems*, pp. 566–580, Springer International Publishing, 2020.

[42] L. E. S. e Oliveira, Y. B. Gumiel, A. B. V. dos Santos, L. M. M. Cintho, D. R. Carvalho, S. A. Hasan, and C. M. C. Moro, "Learning portuguese clinical word embeddings: A multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task," vol. 264, pp. 123–127, IOS Press, 8 2019.

[43] B. Chiu, S. Pyysalo, I. Vulić, and A. Korhonen, "Bio-simverb and bio-simlex: wide-coverage evaluation sets of word similarity in biomedicine," *BMC Bioinformatics*, vol. 19, p. 33, 12 2018.

[44] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pp. 54–59, 2019.

[45] Z. Liu, B. Tang, X. Wang, and Q. Chen, "De-identification of clinical notes via recurrent neural network and conditional random field," *Journal of Biomedical Informatics*, vol. 75, pp. S34–S42, 11 2017.

[46] E. Commission, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing directive 95/46/ec," 5 2016.

[47] M. Kayaalp, "Modes of de-identification," in *AMIA annual symposium proceedings*, vol. 2017, p. 1044, American Medical Informatics Association, 2017.

[48] H. Berg, A. Henriksson, and H. Dalianis, "The impact of de-identification on downstream named entity recognition in clinical text," p. 1 – 11, 2020.

[49] D. Mahendran, C. Luo, and B. T. Mcinnes, "Review: Privacy-preservation in the context of natural language processing," *IEEE Access*, vol. 9, pp. 147600–147612, 2021.

[50] S. Nikoletos, S. Vlachos, E. Zaragkas, C. Vassilakis, C. Tryfonopoulos, and P. Raftopoulou, "Rog§: A pipeline for automated sensitive data identification and anonymisation," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 484–489, 2023.

[51] R. Jain, D. S. Anand, and V. Janakiraman, "Scrubbing sensitive phi data from medical records made easy by spacy–a scalable model implementation comparisons," *arXiv preprint arXiv:1906.06968*, 2019.

[52] P. J. Chambon, C. Wu, J. M. Steinkamp, J. Adleberg, T. S. Cook, and C. P. Langlotz, "Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods," *Journal of the American Medical Informatics Association*, 11 2022.

[53] C. Meneses-Oliveira, "Use of electronic medical records for research: New ethical challenges and possible solutions; [uso dos registos clínicos eletrónicos para investigação: Novos desafios éticos e soluções possíveis]," *Acta Medica Portuguesa*, vol. 32, no. 5, p. 332 – 334, 2019.

[54] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio, "Portuguese word embeddings: Evaluating on word analogies and natural language tasks," *arXiv preprint arXiv:1708.06025*, 2017.

[55] H. D. dos Santos, A. P. Silva, M. C. O. Maciel, H. M. V. Burin, J. S. Urbanetto, and R. Vieira, "Fall detection in ehr using word embeddings and deep learning," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 265–268, 2019.

[56] J. Santos, B. Consoli, C. dos Santos, J. Terra, S. Collonini, and R. Vieira, "Assessing the impact of contextual embeddings for portuguese named entity recognition," in *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pp. 437–442, 2019.

[57] J. Santos, H. D. dos Santos, F. Tabalipa, and R. Vieira, "De-identification of clinical notes using contextualized language models and a token classifier," in *Brazilian Conference on Intelligent Systems*, pp. 33–41, Springer, 2021.

[58] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," 9 2020.

[59] Y. Xiao, S. Lim, T. J. Pollard, and M. Ghassemi, "In the name of fairness: Assessing the bias in clinical record de-identification," p. 123 – 137, 2023.

[60] "Catálogo português de radiologia," tech. rep., 12 2021.

[61] "Catálogo português de gastrenterologia," tech. rep., 12 2021.

[62] "Catálogo português de cardiologia," tech. rep., 3 2022.

[63] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, pp. 596–606, 5 2017.

[64] X. Yang, T. Lyu, Q. Li, C.-Y. Lee, J. Bian, W. R. Hogan, and Y. Wu, "A study of deep learning methods for de-identification of clinical notes in cross-institute settings," *BMC Medical Informatics and Decision Making*, vol. 19, p. 232, 12 2019.

[65] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set," *Applied Soft Computing*, vol. 97, p. 106779, 2020.

[66] A. E. W. Johnson, L. Bulgarelli, and T. J. Pollard, "Deidentification of free-text medical records using pre-trained bidirectional transformers," pp. 214–221, ACM, 4 2020.

[67] A. Ahmed, A. Abbasi, and C. Eickhoff, “Benchmarking modern named entity recognition techniques for free-text health record deidentification,” *AMIA Summits on Translational Science Proceedings*, vol. 2021, p. 102, 3 2021.

[68] A. T. McCray, A. Burgun, and O. Bodenreider, “Aggregating umls semantic types for reducing conceptual complexity,” *Studies in health technology and informatics*, vol. 84, no. 0 1, p. 216, 2001.

[69] L. Radhakrishnan, G. Schenk, K. Muenzen, B. Oskotsky, H. Ashouri Choshali, T. Plunkett, S. Israni, and A. J. Butte, “A certified de-identification system for all clinical text documents for information extraction at scale,” *JAMIA Open*, vol. 6, p. ooad045, 07 2023.

APPENDIX A

UMLS semantic types and proposed groups

Table A.1: UMLS semantic types and proposed groups

Semantic group	Semantic type	UMLS semantic group
Anatomy	Anatomical Structure	Anatomy
	Body Location or Region	Anatomy
	Body Part, Organ, or Organ Component	Anatomy
	Body Space or Junction	Anatomy
	Body Substance	Anatomy
	Body System	Anatomy
	Cell	Anatomy
	Tissue	Anatomy
Chemicals & Drugs	Amino Acid, Peptide, or Protein	Chemicals & Drugs
	Antibiotic	Chemicals & Drugs
	Biologically Active Substance	Chemicals & Drugs
	Immunologic Factor	Chemicals & Drugs
	Inorganic Chemical	Chemicals & Drugs
	Organic Chemical	Chemicals & Drugs
	Pharmacologic Substance	Chemicals & Drugs
Concepts & Ideas	Conceptual Entity	Concepts & Ideas
	Functional Concept	Concepts & Ideas
	Idea or Concept	Concepts & Ideas
	Intellectual Product	Concepts & Ideas
	Qualitative Concept	Concepts & Ideas
	Regulation or Law	Concepts & Ideas

Semantic group	Semantic type	UMLS semantic group
	Spatial Concept	Concepts & Ideas
	Temporal Concept	Concepts & Ideas
Devices	Medical Device	Devices
Disorders	Acquired Abnormality	Disorders
	Anatomical Abnormality	Disorders
	Cell or Molecular Dysfunction	Disorders
	Congenital Abnormality	Disorders
	Disease or Syndrome	Disorders
	Finding	Disorders
	Injury or Poisoning	Disorders
	Neoplastic Process	Disorders
	Pathologic Function	Disorders
	Sign or Symptom	Disorders
Living beings	Bacterium	Living beings
	Fungus	Living beings
	Virus	Living beings
Organizations	HC Related Organization	Organizations
Physiology	Clinical Attribute	Physiology
	Mental Process	Physiology
	Organ or Tissue Function	Physiology
	Organism Attribute	Physiology
Procedures	Diagnostic Procedure	Procedures
	Health Care Activity	Procedures
	Laboratory Procedure	Procedures
	Molecular Biology Research Technique	Procedures
	Research Activity	Procedures
	Therapeutic or Preventive Procedure	Procedures

Semantic group	Semantic type	UMLS semantic group
Quantitative Concept	Quantitative Concept	Concepts & Ideas
Other	Activity	Activities & Behaviors
	Biomedical Occupation or Discipline	Occupations
	Gene or Genome	Genes & Molecular Sequences
	Laboratory or Test Result	Phenomena
	Manufactured Object	Objects
	Occupation or Discipline	Occupations
	Patient or Disabled Group	Living Beings
	Phenomenon or Process	Phenomena
	Professional or Occupational Group	Living Beings