



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Assessing NER Tools for Dialogue Data Anonymization

Miguel Alexandre da Silva Sarmiento Falco Pereira

Master in Computer Engineering

Supervisor:

PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2023



TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Assessing NER Tools for Dialogue Data Anonymization

Miguel Alexandre da Silva Sarmiento Falco Pereira

Master in Computer Engineering

Supervisor:

PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2023

Acknowledgements

Em primeiro lugar gostaria de agradecer todo o apoio prestado pelos meus Orientadores de tese, Professor Fernando Batista e Professor Ricardo Ribeiro. Obrigado por estarem sempre disponíveis para ajudar no que fosse preciso e obrigado por providenciarem o conhecimento necessário para a realização deste trabalho.

Gostaria também de agradecer aos meus pais, Augusto César e Maria de Fátima, por sempre terem acreditado em mim e por insistirem comigo para continuar a trabalhar.

Por último, gostaria de agradecer à minha namorada por ter sido capaz de me motivar a fazer este trabalho ao longo do tempo e por me ter ouvido sempre que precisava de fazer um brainstorm de ideias.

Lisboa, 27 de outubro de 2023
Miguel Alexandre da Silva Sarmiento Falco
Pereira

Resumo

Com o aumento do número de organizações que processam dados sensíveis, aumenta também a necessidade de as empresas assegurarem a privacidade dos seus clientes. No entanto, os métodos de segurança e proteção de dados sensíveis envolvem, frequentemente, procedimentos manuais ou semi-automáticos, os quais consomem muitos recursos e são propensos a erros.

Esta tese aborda anonimização de dados, centrando-se em modelos de Reconhecimento de Entidades Mencionadas. Em particular, investigamos e comparamos vários modelos de Reconhecimento de Entidades Mencionadas para a língua portuguesa para anonimizar automaticamente dados não estruturados.

Na abordagem de aprendizagem automática foram utilizados os modelos do SpaCy, STRING, WikiNEuRal e RoBERTta com o intuito de identificar classes como Pessoa, Localização e Organização. Contudo, a abordagem baseada em regras procura identificar classes como NIF, Email, Matrícula de carro e até mesmo Código Postal.

Consequentemente, foi construída uma ferramenta em Flask, capaz de processar dados não estruturados e anonimizá-los, mais especificamente, capaz de, dada uma string (que simule uma mensagem), anonimizar o seu conteúdo sensível automaticamente. Esta ferramenta combina diferentes técnicas para a Identificação e Extração de Entidades Mencionadas para a língua portuguesa, baseando-se em modelos de regras e de aprendizagem automática.

A junção de ambos os modelos de regras e aprendizagem automática na mesma ferramenta foi essencial para conseguirmos abranger mais classes sensíveis para anonimização, sendo que os resultados calculados para a extração de entidades da ferramenta contruída neste trabalho, engloba os resultados para as três classes calculadas com o modelo SpaCy, com a adição dos modelos de regras criados.

Palavras chave

Anonimização de Dados; Extração de Entidades; Processamento Natural da Língua; Reconhecimento de Entidades Mencionadas; Dados Sensíveis.

Abstract

As the number of organizations processing sensitive data grows, so does the need for businesses to protect and ensure the privacy of their customers. However, the prevailing methods for protecting sensitive data often involve manual or semi-automatic procedures, which can be resource-intensive and error-prone.

This dissertation addresses data anonymization by focusing on Named Entity Recognition (NER) models. Particularly, we investigate and compare various NER models for the Portuguese language to automatically and effectively anonymize unstructured data.

The models SpaCy, STRING, WikiNEuRal and RoBERTa are used in the machine learning approach with the goal of identifying classes such as Person, Location, and Organization. On the other hand, the rule-based approach seeks to identify classifications such as NIF, Email, Car Plate and even Postal Code.

Additionally, it was created a Flask API tool capable of processing unstructured data and anonymizing it, more specifically, given a string that simulates a message, automatically anonymize the message content that might be considered as sensitive. This tool combines many techniques for identifying and extracting mentioned entities for the Portuguese language, based on rule models and machine learning.

The combination of both rule-based and machine learning models in the same tool was crucial to enable the ability to encompass more sensitive classes for anonymization. The results calculated for the extraction of entities from the tool built in this work encompasses the results for the three classes calculated with the SpaCy model, with the addition of the results calculated for the rule-models created.

Keywords

Data Anonymization; Entities Extraction; Natural Language Processing; Artificial; Named Entity Recognition; Sensitive Data.

Contents

- 1 Introduction** **1**
 - 1.1 Context and Motivation 1
 - 1.2 Research Questions 2
 - 1.3 Goals 2
 - 1.4 Research Methodology 3
 - 1.5 Document Structure 4

- 2 Related Work** **7**
 - 2.1 Relevant Concepts 7
 - 2.1.1 Personal and Sensitive Data 7
 - 2.1.2 General Data Protection Regulation 8
 - 2.1.3 Natural Language Processing 9
 - 2.2 Literature Review 10
 - 2.2.1 Methodology for Literary Review 10
 - 2.2.2 Anonymization System Architecture 12
 - 2.2.3 Pre-processing 13
 - 2.2.4 Named Entity Recognition 14
 - 2.2.5 Co-Reference-Resolution 16
 - 2.2.6 Anonymizing Text Data 16
 - 2.2.7 Summary 18

- 3 Assessing Named Entity Recognition for Portuguese Conversations** **19**
 - 3.1 Rule Based Models 19
 - 3.2 Named Entity Recognition Tools 20
 - 3.2.1 SpaCy 21

3.2.2	STRING	21
3.2.3	WikiNEuRal	21
3.2.4	RoBERTta	22
3.3	Datasets	22
3.3.1	Second HAREM Golden Collection	22
3.3.2	Seguradora Dataset	23
3.4	Metrics	29
3.5	Experiments and Results	30
3.5.1	Second HAREM Golden Collection	30
3.5.2	Seguradora Dataset	32
3.6	Discussion	35
4	Anonymization Tool	37
4.1	The API	37
4.2	System Architecture	38
4.3	Pre-processing Module	38
4.4	NER Module and Anonymization Module	39
4.5	Named Entity Recognition Results	40
5	Conclusions and Future Work	41
	Bibliography	42

List of Figures

- 1.1 Design Science Research Methodology Process Model. 5

- 2.1 Flowchart of the anonymization pipeline. 13

- 3.1 Example of a full conversation between a user, a chatbot and an assistant. . . . 26
- 3.2 Example of the same conversation as Figure 3.1 without the chatbot messages. 27
- 3.3 Total number of conversation that contain a certain number of words. 28
- 3.4 Total number of conversation that contain a certain number of words between certain intervals. 28

- 4.1 Screenshot of the API’s endpoint needed to anonymize given text. 38
- 4.2 Flowchart of the API project anonymization pipeline. 38

List of Tables

- 2.1 Number of articles per year for each searching query using IEEE Xplore Digital Library. 11

- 3.1 Quantitative distribution of occurrences for each category in Second HAREM Golden Collection. 23
- 3.2 Words and messages distribution for Seguradora dataset. 24
- 3.3 Number of conversation and their respective average number of words in different scenarios using the Seguradora dataset. 24
- 3.4 SpaCy metrics results for the recognition of each entity class using Second Harem dataset. 31
- 3.5 WikiNEuRal metrics results for the recognition of each entity class using Second Harem dataset. 31
- 3.6 RoBERTa metrics results for the recognition of each entity class using Second Harem dataset. 32
- 3.7 SpaCy metrics results for the recognition of each entity class using Seguradora dataset. 33
- 3.8 STRING metrics results for the recognition of each entity class Seguradora dataset. 33
- 3.9 WikiNEuRal model metrics results for the recognition of each entity class using Seguradora dataset. 34
- 3.10 RoBERTa metrics results for the recognition of each entity class using Seguradora dataset. 34

Abbreviations

NER	Named Entity Recognition
GDPR	General Data Protection Regulation
EU	European Union
SLR	Systematic Literature Review
NLP	Natural Language Processing
POS	Part of Speech
MSF	Morpho-Syntactic Features
CVT	Cross-View Training
CRR	Co-reference Resolution
NE	Named Entity
KB	Knowledge Base
TP	True Positives
FP	False Positives
FN	False Negatives



Introduction

In this chapter we establish the groundwork for this dissertation. We will begin by explaining why we are delving into this topic, followed by the key research questions and what we hope to achieve with this work. We will also go over the methods used to conduct our research and to conclude this chapter, we will give an initial look into the structure of the entire document, so you know what to expect as you read further.

This chapter is the starting point to our exploration of smart data anonymization for the Portuguese language.

1.1 Context and Motivation

Privacy is a fundamental human right (United Nations Declaration of Human Rights, UDHR, 1948, Art. 12) and essential to any society. It helps us set boundaries to protect us from some types of unwanted interference in our lives [1], by limiting who can have access to our personal information that only concerns us.

Technology will always relate to this right, and we can confidently assert that our ability to protect our personal data has improved over time. A good example of this is the establishment of the General Data Protection Regulation (GDPR), which was implemented in 2018 and applies to all citizens of the European Union and the European Economic Area. GDPR has as its focus the improvement of data protection and the clarification of how businesses can process personal data while respecting people's right to privacy. Particularly, personal data may not be distributed or used for other purposes without explicit authorization from the individuals to whom the data pertains.

However, acquiring this consent might be quite difficult. As a result, data anonymization emerges as a viable option, capable of transforming sensitive data into data that does not jeopardize the entity's privacy.

Anonymization of data is a data processing technique that removes or modifies sensitive data while attempting to maintain its original format. The term "sensitive data" refers to any type of information that, when used, may be used to identify a person, such as a person's name, age, sexual orientation, or even bank account details.

This topic arises from the need for a company to automatically detect sensitive and private data that must be anonymized. One of the reasons for this requirement is the GDPR, but its significance extends to other contexts, such as the availability of data for scientific research.

1.2 Research Questions

In this section, we outline a set of fundamental research questions to guide the investigation into the usage of Named Entity Recognition (NER) models in the context of Portuguese text analysis and their significance in smart anonymization tasks. The questions are listed below:

1. Which are the different Named Entity Recognition models available for Portuguese? Understanding the landscape of NER models specific to the Portuguese language is a crucial starting point. This question looks to create an inventory of available tools, including their advantages and disadvantages, by evaluating existing models and their features. As such, gaining knowledge about state-of-the-art NER models for Portuguese will provide a foundation for subsequent research.
2. How do different NER models compare in their ability to accurately extract entities from Portuguese text? This question dives deeper into the performance characteristics of the different NER models. We will look at their precision, recall and F1-Score when it comes to identifying and classifying named entities within Portuguese text. We hope to determine the best models for our specific use case.
3. What are the anonymization techniques that can be used to conceal sensitive entities? In the context of natural language processing and data privacy, this question seeks to investigate the methods and strategies available for anonymizing sensitive entities. By examining established anonymization techniques, we can develop a better understanding of how to safeguard sensitive information while preserving the integrity and utility of the data.

These research questions serve as the foundation for the subsequent chapters of this dissertation. We aim to contribute to the development of the information on NER models, their applications in Portuguese, and their value in smart anonymization tasks by answering these questions.

1.3 Goals

The major goal of this research is to investigate and assess the performance of various Named Entity Recognition models when applied to specific data sets. This study aims to

accomplish the following general goals:

1. **Evaluation of NEE Model Effectiveness:** The first goal of this research is to systematically evaluate the effectiveness of different NER models in the context of identifying and classifying named entities within a given dataset. This work aims to establish which NER models exhibit superior performance according to certain types of metrics.
2. **Comparative Analysis:** Another important goal is to undertake a comparison of the selected NER models. This involves evaluating their strengths and weaknesses, highlighting disparities in performance across different types of entities. This research seeks to provide insights into the benefits and downsides of each NER model under scrutiny.
3. **Anonymization Potential:** The central focus of this research is to determine the suitability of NER models for anonymizing sensitive data, comparing each one of them. This research will investigate whether certain NER models are better suited for the anonymization purpose and whether their performance varies with different types of sensitive entities.
4. **Practical Applications:** This study also aims to contribute to the practical application of NER models in data anonymization processes. By identifying which models perform optimally for anonymization, this study can provide valuable guidance for organizations and practitioners seeking to implement NER-based anonymization solutions in real-world scenarios.

In summary, this research addresses the effectiveness of employing NER models for the anonymization of sensitive data. These models must be able to identify sensitive or private terms so that they can be appropriately anonymized to protect the identity of the person to whom they are being applied to. For example, for the sentence

“O meu nome é Miguel e vivo em Lisboa”

the system built in this work must be able to recognize that "Miguel" is a person's name and that "Lisboa" is a city (location) that may be considered sensible. Hence, this system must be capable of replacing sensitive data in a sentence with, for example, "*". The result of the anonymization of the previous sentence should be something like:

“O meu nome é *** e vivo na ***”.

1.4 Research Methodology

The methodology used to develop this work was Design Science Research (DSR). Its main goal is to identify the procedures and observational points that allow for the development of

various research projects, particularly in the field of engineering [2], to create the desired work. Figure 1.1, extracted from Peffers et al. [3] illustrates all the steps that must be taken into consideration when using this research methodology.

First, we must determine the purpose behind the development of this topic and identify any potential issues. This stage is crucial when it comes to defending the value of the solution that has been presented, with the main goal being to encourage the search for the solution and make the results easier to accept. Section 1.1 and Section 1.2 outline how this step developed.

The project objectives, identified in Section 1.2, that are required for the achievement of the solution must then be defined in step two, using the previously identified issues as a starting point. These goals are identified in Section 1.3 of this dissertation.

The third step in this methodology is the design and development of the proposed solution. In our work, we will be using different types of NER models and compare them later using the same datasets for all. This study will serve as the baseline to pick which model performs better and try to include it in the tool we built in this work.

Next, we have the demonstration's step. It seeks to demonstrate the effectiveness of the solution developed to fix the problems in question. In the context of this dissertation, we will describe the data collected so that we can later apply our pre-processing techniques and implement the respective NER models to extract named entities.

In the fifth step, we reach the point when the solution is evaluated. Here, we will introduce a comprehensive set of metrics designed to evaluate the performance of the NER models in question, taking into consideration the obtained results.

Finally, we reach the communication step. This entails communicating the generated solution to the public, whether through the writing of this dissertation or even through the writing of an article that might be published in a scientific journal.

1.5 Document Structure

This document is composed by six chapters, including the first Chapter Introduction. It has the following structure:

Chapter 2 focuses on introducing relevant concepts regarding Personal and Sensitive Data, as well as what is GDPR and why is it important within the scope of this dissertation. It also gives some information of very important concepts for this dissertation that are Personal and Sensitive Data and Natural Language Processing. This chapter also focuses on the state-of-the-art in Natural Language Processing and anonymization of text data. It describes the related work done in this area and how the research to find suitable and useful information was done. Therefore, a literature review is elaborated in this chapter, focusing on existing systems related to the topics covered by this dissertation. Finally,

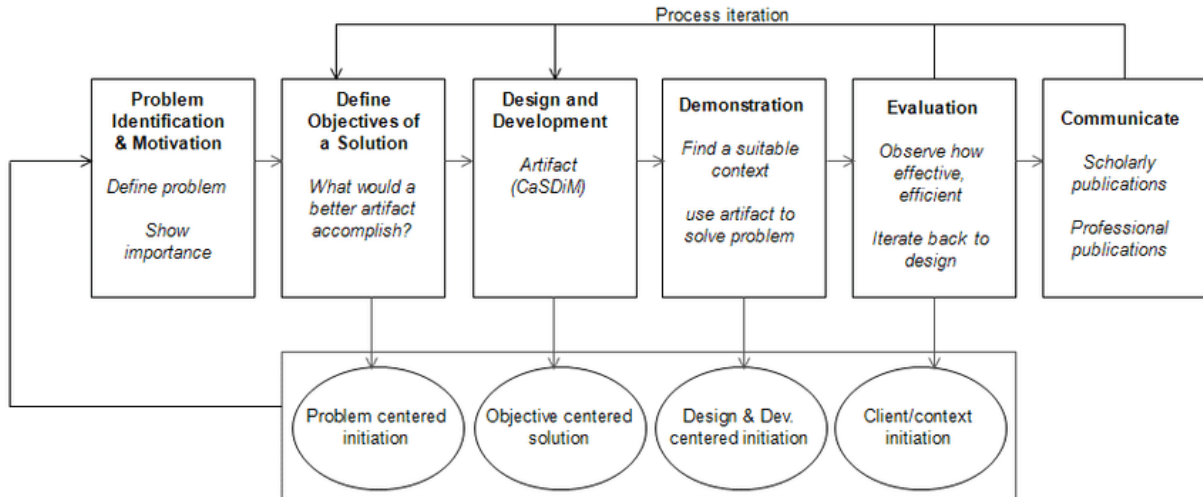


Figure 1.1: Design Science Research Methodology Process Model (Peppers et al. [3]).

this chapter will focus on explaining certain concepts that are important to the task of finding sensitive data such as Named Entity Recognition and Co-reference-Resolution. In addition, it also presents methods for processing data and an architecture of a possible Anonymization System.

Chapter 3 starts by describing the resources used to the development of this work, as well as the methods and models that were implemented. This chapter also gives an in-depth description of the datasets used for our experiments using different types of NER model. This chapter also presents the results for the Named Entity Recognition task for each module tested, using the metrics also enumerated on this chapter.

Chapter 4 explains how the solution implemented for the topic of this dissertation was built and how it works, finalizing with the results for the Named Entity Recognition task.

Lastly, Chapter 5 discusses the results presented in the Chapter 3 and Chapter 4, as well as what can be done in the future to enhance all the work done in this dissertation.

Related Work

2

The main goal of this chapter is to research and analyze potential technologies that might be related to the development of this topic. Therefore, it is quite important to follow a specific methodology while looking for articles and research papers that might offer ideas or even solutions for the evolution of the subject at hand. Considering this, we start by understanding the review methodology used in this work, as well as its relevant results, followed by describing the related work done in the area.

2.1 Relevant Concepts

This section gives a brief overview on some concepts that are important to take into consideration for the context of this dissertation. We will begin by explaining the differences between Personal and Sensitive data. After that we will give a brief description on General Data Protection Regulation and end this section by describing what Natural Language Processing is, as well as the tasks it performs.

2.1.1 Personal and Sensitive Data

Nowadays, personal data is commonly acknowledged as a very valuable asset. Personal data is a primary target for most cybercriminals and hackers due to its value in the business world.

Within the scope of this dissertation, two distinct concepts are distinguished: personal data and sensitive data, each playing an important role in the wider issue of data privacy and security.

Personal data is any amount of information that can be used to identify an individual, making it essential to understand and safeguard. This includes, but is not limited to, details such as an individual's name, phone number, address, age or even their email address. Personal data can simply be information that categorizes your presence. It is important to recognize that personal data encompasses not only explicit identifiers but also any information that, when evaluated collectively, allows the identification or behavioral profile of an individual, even if the data is anonymized or not directly linked to that individual.

The significance of personal data lies in its role as the basis of one's digital identity and the potential insights it can offer into an individual's preferences, behaviours and characteristics.

Sensitive data represents a distinct and more protected category within the area of personal data. It requires special attention and precautions, particularly in compliance with regulations like the GDPR. Sensitive data comprises information that, if mishandled or disclosed, could have significant consequences for an individual. This includes:

- Personal data: Data that might lead to the discovery of an individual racial or ethnic origin, political and religious beliefs.
- Political and religious beliefs: Information concerning an individual's political ideas or religious beliefs.
- Genetic or biometric data: Highly sensitive data, such as DNA sequences or biometric identifiers like fingerprints or retinal scans.
- Sexual orientation: Data that pertains to an individual's sexual orientation.
- Mental or sexual health: This type of data can be very sensitive and stigmatized.
- Trade Union Membership: Data indicating an individual's participation in a trade union, which may be a concern in certain contexts.

Understanding these concepts is fundamental because they work as the cornerstone for data privacy and security concerns, which are essential to the evaluation of the NER models investigated in this work. Identifying and securing personal and sensitive data effectively is crucial to both maintaining individuals' privacy and complying with data protection regulations, making both concepts central to this work general goals and objectives.

2.1.2 General Data Protection Regulation

The GDPR is the strictest privacy and security law in the world. Though it was designed and passed by the European Union (EU), it also imposes obligations on enterprises anywhere that target or collect data about EU citizens. It went into effect in 2016 after being approved by the European Parliament, and all organizations were obliged to be compliant by May 25, 2018.

According to GDPR, sensitive data, need to be specially secured, comparatively to personal data, due to the fact if such information is exploited or improperly handled, the individual may face discrimination or mistreatment.

The GDPR's goal is to offer a set of standardized data protection legislation across all member countries, strengthening individual rights and making laws more consistent and clearer. It aims to make it easier for EU people to understand how their data is used and to lodge concerns, even if they are not in the country where the data is stored.

2.1.3 Natural Language Processing

Natural Language Processing (NLP) is the branch of computer science, specifically, the branch of artificial intelligence, whose concern is giving computers the ability to comprehend text and spoken words in the same sense that humans do. By combining computation linguistics with statistical, machine learning and deep learning models, NLP enables computers to process human language in the form of text or voice data, using the writer's intent and sentiment to try to understand the meaning of the data.

Several NLP tasks have been developed to deconstruct human text and voice data in order to assist the computer in comprehending the input that is being given to it. Among these tasks are the following:

- Speech recognition - Involves converting spoken words into written text. This technology supports voice assistants like Siri and Google Assistant but also plays a critical role in transcription services and voice-controlled systems.
- Part of speech tagging - This task involves labeling individual words in a sentence with their corresponding grammatical categories, such as verbs, nouns, adjectives and even adverbs. This allows computers to understand the syntactical structure and relationships within a sentence.
- Named entity recognition - Is the process of detecting and categorizing specific entities within a text. This task is essential for extracting relevant data and detecting potentially sensitive data.
- Co-reference-resolution - This task addresses the challenge of understanding when two or more words, or phrases, within a text refer to the same entity. An accurate co-reference-resolution enhances the coherent understanding of context, especially in longer texts.
- Sentiment analysis - NLP uses sentiment analysis to determine the emotional tone and attitude expressed in a piece of text. This task assists in gauging public opinion, customer feedback and social media sentiment for example.

In this work, we aim to assess and compare existing NLP tools, focusing on their performance in the NER task to determine which models excel at identifying sensitive data. By elucidating the role of NER within the context of security in the digital age, this research seeks to enhance our understanding of effective solutions for safeguarding sensitive information, with practical implications for data privacy protection in the Portuguese language.

2.2 Literature Review

This section will contemplate the methodology used for the literature review, as well as some important tasks and approaches used by other researchers. This section will give a broader view of the work already done in the area for named entity recognition and anonymization of text data, that will serve as a starting point for the development of this work.

2.2.1 Methodology for Literary Review

The review methodology used in this study was the Systematic Literature Review (SLR). The SLR is an approach that leads to a very rigorous review of the literature. This has the ability to improve the review of literature by bringing transparency and rigor in a variety of ways so as to address a set of predetermined questions and evaluate, analyze, and synthesize previous research that has been done by researchers Pereira and Serrano [4]. Considering this, the SLR implemented is based on the recommendations made by Kitchenham [5].

Firstly, it is important to understand why a literature review is necessary in the context of this topic. The main goal of this research is to understand how smart data anonymization has changed over time and how it was accomplished. As a result, it is necessary to understand what the potential algorithms and models that can be built, or not, considering the work done by other researchers.

Two different repositories were used to search for articles in order to combine the most pertinent information possible that was related to the earlier stated goal:

1. Google Scholar¹
2. IEEE Xplore Digital Library²

It is important to note that Google Scholar allows you to search through articles that are already available in a wide range of repositories, including the IEEE.

Thus, a collection of keywords was used to provide a more effective search by allowing the filtering of all content in repositories. In this context, the terms "Data Anonymization" and "Smart Anonymization", "Machine Learning" and "Algorithms" were some of the ones used for this effect.

Using the term "Text Anonymization" in the IEEE search engine, a total of 44 articles were returned.

¹<https://scholar.google.com/>

²<https://ieeexplore.ieee.org/Xplore/home.jsp>

Table 2.1: Number of articles per year for each searching query using IEEE Xplore Digital Library.

	"Text Anonymization"	[("Named Entity Recognition" OR "NER") AND "Anonymization"]	[("ML" OR "Machine Learning") AND "Data Anonymization"]
2023	9	1	44
2022	6	4	30
2021	4	1	19
2020	2	0	16
2019	2	2	17
2018	2	0	4
2017	1	0	6
2016	2	0	6
2015	4	0	4
2014	1	0	1
2013	2	0	1
2012	2	0	2
2011	4	0	1
2010	1	1	1
2009	1	0	0
2008	1	0	2
2007	0	0	1
2005	0	0	1
Total	44	9	156

Followed by a brief analysis of several articles, we came to the conclusion that exists a key concept in the development of this work: Named Entity Recognition. As a result, when we search for articles that contain the keyword "Named Entity Recognition" in the search engine, we get a total of 1634 results, indicating that we need to create a more specific query capable of filtering out a large portion of these articles. Consequently, using the query [("Named Entity Recognition" OR "NER") AND "Anonymization"] a total of nine articles are returned. Additionally, because of the advanced search functionality of the IEEE Xplore repository [2](#), when searching using the query [("ML" OR "Machine Learning") AND "Data Anonymization"] a total of 156 articles are returned.

We can draw the following conclusions regarding the theme of this research by looking at Table 2.1:

1. A Developing Searching Field: This Table shows a clear increase in research activity, especially in recent years. The overall number of papers has gradually increased, with a significant surge in 2023, which can reflect the increasing importance and interest in the task of Data Anonymization.
2. Intersection of NER and Anonymization: The data demonstrates that research articles related to both NER and Anonymization are relatively sparse. This suggests that the

junction between these two domains has not been thoroughly investigated and may present an opportunity for further research and innovation.

3. Shift Toward Data Anonymization: There is a consistent and growing number of articles specifically dedicated to Data Anonymization, indicating an increased awareness of the importance of data privacy and the need for robust anonymization approaches. We can also conclude by looking at the lack of historical data in the early years presented in the Table that this subject is getting more attention. It is worth to note that no article published before 2005 was found.

Following this search, the titles of the articles and their respective abstracts were read in order to reduce the number of articles for analysis while keeping just those that may contribute with information that is pertinent and useful to the topic of this dissertation. It is worth noting that, while using Google Scholar, the total number of returned articles was in the tens of thousands total, which is why only the first 20 returned articles, ordered by relevance, were used in this method using Google Scholar repository 1.

After the analysis of the titles and abstracts of the selected articles, only 22 were chosen because they could provide important information for the development of this work.

In the following sections, we will look at these selected articles and focus on collecting and synthesizing valuable insights embedded within these works. This way, we can acquire a comprehensive understanding of how things have been done in the area of Data Anonymization and which tools and techniques can be used to fulfill the objectives of this work.

2.2.2 Anonymization System Architecture

A large part of the work done in this area relies on the use of NER models to identify potential terms to be anonymized.

The authors of [6] propose the architecture depicted in Figure 2.1. They begin by pre-processing the data so that it can later be used in the NER module. Both the pre-processing and the NER modules make use of the STRING chain [7], which is a Natural Language Processing (NLP) chain for Portuguese capable of performing operations such as Part of Speech (POS) tagging, chunking, and even NER.

The second module, NER, will receive a document and return a list of all possible entities in that document, each one identified by its position and respective class.

Following that, the Co-reference Resolution module determines whether two Named Entities from the same document refer to the same extralinguistic object in order to prevent the text from losing information about the original entities.

Finally, we have the anonymization phase, which aims to delete or replace sensitive entities by returning the anonymized text version.

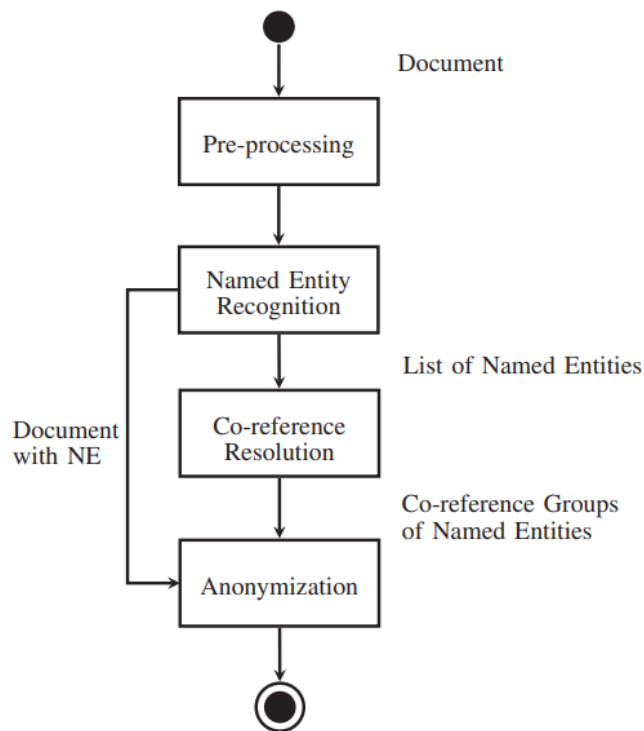


Figure 2.1: Flowchart of the anonymization pipeline implemented by Mamede et al. [6].

The authors decided to implement this modular arrangement due to the fact it allows the adaptation of the anonymization system to different languages by straightforwardly substituting the tools employed within each module.

The effectiveness of the anonymization system was evaluated using a dataset of 75 documents chosen from two distinct corpora. One corpus was made up of 50 papers from Segundo HAREM golden collection [8] with a total of 23342 tokens. The second corpus included 25 reports from the Digital Corpus of European Parliament with a total of 11497 tokens. The authors chose these corpora due to their abundance of NEs and systematic document organization. The NEs were limited to Person, Location and Organization classes, and all papers were annotated with co-references by three annotators.

The authors of [6] present a variety of anonymization methods that can be used, such as replacing the original term for a neutral character, for example, replace the term “Miguel” (which is the name of a person), for something like “XXXXXX”.

2.2.3 Pre-processing

Text pre-processing is an essential component of any NLP system because the characters, words, and sentences identified at this stage serve as the fundamental units for all subsequent processing stages Kannan et al. [9].

The work done by Dias [10] presents a module specifically designed for pre-processing, divided by three distinct tasks that are always executed in the same order. The first task is segmentation. It consists in dividing the entirety of the text, when finding punctuation marks like “.”, “?”, “!” and “...”, so that each sentence can be processed individually without the context of the previous sentence. The second task, tokenization, consists in dividing the text into n-grams, words, or groups of words. Finally, the last task, morphosyntactic analysis, will analyze and classify the words at the syntactic level using POS tagging.

Dias [11] pre-processing has some similarities to the one we just analyzed.

The first thing done in the pre-processing module is the normalization of the text and the extraction of features. The normalization step ensures that the input text has the same format and character encoding, as such, all the text encoding is converted to UTF-8. This phase guarantees that the NER models can function properly across diverse textual sources and language variances.

Nextly, they proceed to doing the tokenization of the text using a particular tokenization tool provided by Moses (Koehn et al. [12]). The choice of what tokenization tool may differ depending on the specific requirements of the NER task and the linguistic features of the text corpus.

Finally, in addition to text normalization, this pre-processing model also includes the extraction of Morpho-Syntactic Features (MSF) at word level. The extraction of MSF’s involves identifying the linguistic attributes associated with each word in the text. For instance, in languages with gendered nouns, this step detects the gender of nouns, which can be useful in disambiguating named entities. Similarly, capturing information about the number of a word (singular or plural) might help in accurately identifying entities and their contextual usage. These extracted MSFs are subsequently used as valuable inputs by downstream modules in the NER pipeline, allowing the models to make informed decisions when identifying and categorizing named entities.

It is important to note that the pre-processing methods used can differ based on the nature of the text corpus and the respective goals of the NER task.

In summary, pre-processing is a critical step because it ensures that textual data is in a standardized format, enhanced with linguistically relevant features.

2.2.4 Named Entity Recognition

Named Entity Recognition is a sub-problem within information extraction that involves processing both structured and unstructured documents to identify references to certain people, places, organizations, and businesses [13].

Using the same training dataset, Gamallo et al. [14] compare the application of a rule-based system with a neural-based tool for NER.

For the rule-based approach, they have adapted the NER module integrated in LinguaKit [15] for Portuguese. The NER that is being used is composed of two distinct rules: the first seeks to find heuristics for the purpose of identifying entities, and the second aims to categorize these entities as Locations, People, Organizations, and Others. In addition, they also use "trigger words" to cope with terms that are ambiguous, such as "Santiago", which can be considered either a name or a city. The search for trigger words in sentences containing these ambiguous terms enables a more precise classification of such terms. For instance, if the term "Santiago" appears in a sentence with the trigger word "cidade," the algorithm will classify "Santiago" as a Localization. If more than one trigger word appear in the same sentence, the first trigger word to occur will be used to classify the term in question.

The neural-based approach was based on Cross-View Training (CVT), which performs semi-supervised learning, combining supervised and unsupervised methods Clark et al. [16]. In the NER scenario, CVT makes use of unclassified data to discover the variations in the contexts in which the contested terms are used. However, in order to carry out the training task, it requires previously classified data, as well as a sizable collection of unclassified data for the unsupervised learning process. Now, using a NER model in the context of data anonymization could be quite useful because it makes it possible to identify specific entities that might be considered as sensitive.

The research done by Bridal [17] delves into the utilization of transfer learning techniques in Named Entity Recognition to anonymize a collection of medical records written in Swedish. The primary objective of this study is to assess the effectiveness of a pre-trained and fine-tuned BERT model, called KB-BERT, in the identification and classification of named entities within unstructured Swedish medical records and potential use for anonymization.

Bridal used a method recommended by Malmsten et al. [18] and implemented the NER model using Huggingface's Transformer library (Wolf et al. [19]). Using this model, a sentence of raw text could be given to it and return a dictionary comprising the tokenized terms of the sentence along with their corresponding predicted entity tags.

The results of Bridal research underscore the significance of accurately detecting Person names in medical records for anonymization purposes. While the model demonstrated exceptional recall (1.0) in identifying Person names, its precision (0.7087) was relatively lower, having a negative impact on downstream NLP tasks. The model used also achieved a total recall score of 1.0 in identifying Location entities, meaning it was able to find all the terms that were locations, but again, at the cost of a lower precision score of 0.45.

2.2.5 Co-Reference-Resolution

The Co-Reference-Resolution (CRR) task serves a vital role in various NLP tasks, such as document similarity, text document summarization and even question-answering systems [20].

Within the context of NLP, the task of CRR can be important because it allows us to determine whether two named entities are referring to the same extra-linguistic object.

Co-reference-Resolution is the task of determining linguistic expressions that refer to the same real-world entity in natural language [21], as such, this technique involves the process of establishing connections among multiple references that pertain to the same entity.

This task is especially important for maintaining context and coherence in longer text or conversations.

Typically, this process unfolds in two distinct stages:

1. Intra-sentence resolution: This stage pertains to the identification of an antecedent within the same sentence as the anaphora.
2. Inter-sentence resolution: In this phase, the objective is to connect the anaphora to an antecedent located in a different sentence than that of the anaphora.

To illustrate this concept, consider the following sentence: “Hannah received a scholarship. The scholarship covered all her expenses.”. In this instance, the feature terms include {Hannah, scholarship, all her expenses} and the CRR task is indispensable in order to establish the connection between pronoun references, like “The scholarship” and “her”, and their respective noun terms.

When trying to perform the anonymization task, the authors of Mamede et al. [6] created a CRR module because named entities that referred to the same object could be replaced by different expressions, enabling the loss of information regarding the original entities. As such, the solution to maintain and avoid any loss of information in the original document, is to replace all occurrences of the Named Entities (NE’s) that relate to the same extralinguistic term by the same expression.

2.2.6 Anonymizing Text Data

Safeguarding sensitive information has become a top priority in an increasingly linked digital society. Anonymization techniques emerge as critical tools in this regard, providing a systematic approach to securing sensitive data while keeping its utility. This section delves into the various anonymization techniques that can be used to the development of this work.

Dias [11] presents a solution to the task of anonymizing Portuguese text. Their anonymization module contemplates four different types of anonymization methods, which can be extended with more methods. These methods include:

- **Suppression:** This is the method used by the majority of text anonymization. It consists of using a neural indicator that replaces the original text, suppressing the NE found, such as "XXXXXX".
- **Tagging:** This method requires replacing a NE with a label indicating its class or unique ID. All occurrences of entities referring to the same object in the text can be substituted with the same label from the CRR module output. In their approach, they concatenate the NER tool's class with a distinct numeric identifier. For example [**Organization123**].
- **Random substitution:** This method seeks to replace a NE by another random NE from the same class and Morpho-Syntactic Features. This technique was implemented using default lists of random entities for each class in different languages. The MFS's are determined by a NE's headword (the primary word in the NE). If the entity is not found in the dictionary, then default MSF's are assumed for the entity, such as masculine, singular and nominative case. With these MSF's, a search is conducted within the default entity list for a matching entry of the same class and attributes. If a suitable entity is not found, the tagging method becomes the alternative.
- **Generalization:** This refers to any approach that substitutes an entity with another that mentions an item of the same category but in broader terms. For instance, the terms "University of Lisbon" could be replaced by "University" or even "Institution". To execute this generalization method, the module needs to access a Knowledge Base (KB) for retrieving the higher-level classes of a given entity. Their current implementation employs Wikidata (Vrandečić [22]) as the chosen KB, which serves as an open and collaborative KB, accessible online.

Murthy et al. [23] addresses a comparative study of data anonymization techniques. This work also covers the Generalization and Suppression techniques, providing a comprehensive understanding of these methods. However, they investigate three more techniques that can be used for data anonymization:

- **Distortion:** This method refers to a process that alters the data into a different form, which can subsequently be restored to its original state using the initial data.
- **Swapping:** Murthy et al. [23] describes this method as a process of rearranging the wanted variables within each column randomly. For example, if we have a table with a column of multiple names then we would have to swap all the names within that column. If we wanted to implement swapping in this data anonymization research,

we would need a KB of names so that when we find a NE that is a Person, we can replace it by a random name retrieved by the KB. However, due to the randomness of this process, there is a possibility of replacing a particular name with the same name itself.

- **Masking:** This method involves a technique where characters within the designated attribute(s) are altered to different characters, rendering the variable inconceivable. For example, they replace any numeric number from 1-9 with 1 and any lower-case a-z with z, or Z if it is upper case. They also present that the challenge about masking lies in the increased resources needed to verify and modify the value.

In conclusion, the variety of anonymization techniques discussed in this section highlights the importance of protecting sensitive information. These methods collectively serve as critical tools for the dual purpose of safeguarding data and preserving its usefulness, thus playing an important role in addressing contemporary data privacy challenges.

2.2.7 Summary

In the preceding sections, we presented various important concepts and techniques that are essential for the development of this work, such as NER, co-reference-resolution and techniques to pre-process and anonymize text data.

The work done by Mamede et al. [6], described in Section 2.2.2 serves as the starting point for the elaboration of this research. The architecture they presented is really similar to what we want to achieve when building the anonymization tool. Also, by looking at the architecture they presented, we can discover that certain tasks like pre-processing and NER are critical points to analyze and research.

Additionally, the work done by Dias [10] and Dias [11] provided really important input that needs to be taken into account when dealing with certain types of data. In their works, they emphasized the significance of standardizing textual data and enriching it with linguistically relevant features, which are crucial in the context of Named Entity Recognition.

In Section 2.2.4, we expanded our investigation to NER because it plays an important role in our research. Bridal [17], in particular, directed our attention to the application of transfer learning techniques using models like KB-BERT, emphasizing their relevance in detecting and classifying named entities within unstructured text.

This chapter sets the foundation for the following sections of this dissertation, where we present our experiments and findings in the context of these previously explored topics. These findings will contribute to a deeper understanding of how different NER models perform on certain types of data.

Assessing Named Entity Recognition for Portuguese Conversations

3

This chapter will approach the resources and methods used to carry out the NER tasks required for subsequent anonymization. It will delve into a thorough description of various methods, techniques and models that can recognize entities, specifically for the Portuguese language.

In this Chapter, we will describe the datasets that will serve as the foundations for our assessments.

Furthermore, this chapter will introduce all the evaluation metrics that will allow us to quantify and qualify the performance of each model presented.

Lastly, this chapter will make a comprehensive analysis of the results for each model performing the NER task, ultimately providing valuable insights into the effectiveness of these models and techniques in the context of Portuguese conversational data.

3.1 Rule Based Models

Named Entity Recognition does not solely focus on using machine learning models to find entities within the text. Some of it can be found by applying certain rules to the text. As such, the entity classes detected using these models are all be directly linked with sensitive data. Within this category we made rules to find the following types of entities:

1. Número de Identificação Fiscal (NIF): This entity class is associated to the tax identification number belonging to a person. The verification rules to find this type of entity are:
 - (a) Nine digit number or PT followed by a nine digit number
 - (b) Multiplying the eighth digit by 2, the seventh digit by 3, the sixth digit by 4, the fifth by 3 and so on
 - (c) Adding up the result of each multiplication
 - (d) Calculating the rest of the number division by 11
 - (e) If the rest result is 0 or 1, the check digit (last digit of the NIF) needs to be 0.

- (f) If the rest R is different than 0 or 1, the check digit needs to be the subtraction of 11-R
- 2. Phone Number: This entity class represents Portuguese phone numbers. The verification rules to find this type of entity are:
 - (a) Nine digit number that starts with a 9 and followed by either a 1, 2, 3 or 6: 9[1236][XXXXXXXX]
 - (b) Numbers that follow the rule above but preceded by a Portuguese code: +351 or 0351
- 3. Postal Code: This entity class encases all data that can be defined as a postal code. It does not have any complex regex associated with it as we only consider certain data to be a postal code if it matches the Portuguese postal code format: XXXX-XXX. The X's can be any digit from 0 to 9, meaning the postal code is only identified if four digits are followed by a hyphen followed by three digits.
- 4. E-mail Address: This entity class corresponds to any email address that matches the following format: "[XXXX]@[XXXX].[XXX]". The X's can be any alphanumeric character or punctuation mark.
- 5. Car Plate: Lastly, a regex to detect car plates was created. This regex looks for terms that have the format XX-XX-XX and that have two pairs of digits pairs and one pair of letters, for example, "34-XY-13". Also, the letters can be either lower or upper case for it to match.

3.2 Named Entity Recognition Tools

One of the primary tasks of NLP is Named Entity Recognition. As explained in Section 2.2.4, NER has as its primary goal the identification of words that may refer to a specific entity within the text. There are many tools that can be used to perform this task, like SpaCy or even Stanford CoreNLP. However, the NER model provided by Stanford CoreNLP does not have a pre-trained model for the Portuguese language and for that reason we will not be using it in this work.

The main entities we will be trying to recognize using these Machine Learning models are Persons, Locations and Organizations. It is worth to note that tools like SpaCy can detect entities that are based on rules such as email addresses, enriching the overall NER task it performs.

As such, the tools which NER models will be explored in this work are SpaCy and STRING.

3.2.1 SpaCy

SpaCy¹ is a free, open-source library for advanced NLP in Python. SpaCy is designed for production use and assists in developing applications that process and "understand" large volumes of text. Therefore, it can be used to create systems for information extraction and natural language understanding, as well as to pre-process text for deep learning.

This tool contains pre-trained pipelines and supports tokenization and training for over 70 languages. It features state-of-the-art speed and neural network models for tagging, parsing, named entity identification, text classification, and other tasks, as well as a production-ready training system and a simple model packaging, deployment, and workflow management.

3.2.2 STRING

STRING [7] is a natural language processing chain for Portuguese that combines statistical and rule-based processing. STRING has a modular structure that handles all basic text processing tasks, such as tokenization and text segmentation, part-of-speech tagging, morphosyntactic disambiguation, chunking and deep parsing. STRING can execute Named Entity Recognition, information retrieval, anaphora resolution, and other NLP tasks. STRING is made available to the community through a web portal.

This tool NER model will be evaluated using our proposed dataset, making the same experience as the one we will be doing with the SpaCy NER model.

3.2.3 WikiNEuRal

The WikiNEuRal model is a multilingual NER model that generates high-quality annotations using a multilingual lexical knowledge base called BabelNet² and a transformer-based architecture like BERT.

BabelNet is a multilingual lexical knowledge repository that provides information about words and their meaning in many languages.

WikiNEuRal combines the utility of neural models with the knowledge-based approach of BabelNet. Additionally, WikiNEuRal presents a new approach for creating an interpretable word embedding, which introduce a domain adaption algorithm, enhancing the performance gains on test settings [24].

Additionally, the WikiNEuRal model has been trained and evaluated on multiple languages, including Portuguese, the one we need for this research.

¹<https://spacy.io/api/doc>

²<https://babelnet.org/>

3.2.4 RoBERTa

RoBERTa Large NER is a state-of-the-art model for NER tasks in NLP. This model is built upon the RoBERTa model, which is a variant of BERT (Bidirectional Encoder Representations from Transformers) model, developed by Facebook.

BERT has revolutionized NLP tasks by introducing pre-training and fine-tuning technique that allows models to learn contextualized representations from large amounts of unlabeled text. The RoBERTa Large NER extends this architecture specifically for NER tasks.

There are many types RoBERTa models, but the one we will be using (the large model) is one that can handle larger sequences of text, allowing it to understand broader contexts. This model uses more data for pretraining, making it more proficient in understanding a wide range of language nuances and it also possess a dynamic masking strategy, which enhances its ability to understand relationships between words.

The training methodology of RoBERTa large NER involves pre-training the model on a large corpus of data, using unsupervised learning techniques. During pre-training, the model learns to predict masked words within sentences, enabling it to capture contextual information from vast amounts of unlabeled text [25].

RoBERTa Large has the skill to classify entities such as people, places and organizations and it uses cross-lingual representation learning at scale [26], making this tool usable for this research since it can be applied to a Portuguese corpus.

3.3 Datasets

This section will start by giving a comprehensive description of both the datasets used to perform the NER task in this work. Firstly, we will start by describing the Second HAREM Golden Collection dataset, and finish this section describing a dataset provided by a company, which contains conversations between users and an insurance’s company chatbot.

3.3.1 Second HAREM Golden Collection

The CDSegundoHAREMclassico dataset is essential for evaluating the performance of NER models in the context of the Portuguese language.

This dataset, which is a subset of the larger HAREM corpus, is particularly well-suited for NER research in Portuguese, as it contains a significant volume of annotated text data for model training, validation, and evaluation.

The Second HAREM collection possesses a total of 1040 documents and was created by adding all training material provided prior to the Golden Collection documents. The

Table 3.1: Quantitative distribution of occurrences for each category in Second HAREM Golden Collection.

Category	Number of Occurrences
PESSOA	2127
LOCAL	1424
TEMPO	1194
ORGANIZACAO	1100
OBRA	495
VALOR	356
ABSTRACCAO	353
COISA	344
ACONTECIMENTO	331
OUTRO	94
SEM CATEG	29
Total	7847

remaining text was then sourced from the CHAVE collection, which includes Portuguese and Brazilian newspapers from 1994 and 1995 [27].

This dataset encompasses the following categories: PESSOA (Person), LOCAL (Location), TEMPO (Time), ORGANIZACAO (Organization), OBRA (Work), VALOR (Value), ABSTRACCAO (Abstraction), COISA (Thing), ACONTECIMENTO (Event), OUTRO (Other) and SEM CATEG (Without Category).

This dataset comprises a total of 7847 named entities, which distribution per category can be visualized in Table 3.1. As we can see, PESSOA stands out as the most common named entity, while OUTRO is the least frequently observed category.

This corpus employs an XML-based annotation format, incorporating supplementary information that will not be utilized. Every named entity identified in the corpus is annotated beginning with an 'EM' tag and closed with '/EM'. However, even though there are many different named entities that can be found in this corpus, we will only be working with the named entities PESSOA, LOCAL and ORGANIZACAO, since they can be considered sensitive information.

As the work done by Dias [10], the data was transformed from XML to a Conll format with IOB tagging in order to make the evaluation easier once we obtain results, since the XML format can not be used as an input to feed the algorithms.

3.3.2 Seguradora Dataset

The dataset used to evaluate the NER Module includes a collection of messages that represent conversations between users and an insurance company chatbot.

Table 3.2 contains information on the total number of messages in the dataset as well as the total number of messages sent by the intervening parties.

Table 3.2: Words and messages distribution for Seguradora dataset.

	User	Chatbot	Assistant	Supervisor	All
Total number of messages	26507	31317	8066	44	65934
Average words per message	4	11	11	14	8

Table 3.3: Number of conversation and their respective average number of words in different scenarios using the Seguradora dataset.

	Original	Without chatbot	User only
Conversations	3785	3425	3316
Average words per conversation	145	55	29

This dataset has a total of 65934 messages and is made up of 31317 chatbot messages, 26507 user messages, 8066 assistant messages, and 44 supervisor messages. Thus, all messages together make up a total of 3785 conversations, each of which contains an average of 17 messages.

Table 3.2 also shows the average number of words used in every message sent by each of the intervenes in the dataset. This information allows us to determine the size of the data and which intervening parties may be more important when it comes to the existence of information worth to be anonymized.

After analyzing the messages, it was concluded that the chatbot’s messages were irrelevant to the task of anonymization. Although the chatbot messages make up most of the data available in this dataset, the pre-defined chatbot messages are only used with the goal of guiding the user in explaining their issue through a series of questions to which the user must respond in order to formulate a potential solution to their issue. We can see an example of a chatbot message above:

(QUICKREPLY) - Não encontrei resposta à sua pergunta, será que quis dizer:
 Dados para pagamento - multibanco e débito direto.

As a result, all messages sent by the chatbot were removed. By looking at Table 3.3, we can see how the total number of conversations and the average number of words per conversation change when the chatbot’s messages are completely removed and in the scenario where only the users’ messages are taken into account.

As we can see, the dataset contains conversations in which the most important intervenient, the user, does not send any messages.

By looking at Figure 3.1 we can see an example of the interaction between a user and a chabot, comprising a total of 39 messages. The table shown in the figure has a total of three columns. The first column, Privilege, corresponds to what intervenient is sending the message and the second column Phase corresponds to the phase of the conversation, in this case, the phase will be Chatbot when the user is interacting with the chabot, or vice-versa, or Assistant when the interaction happens between a user and an assistant. The

third column, Text, corresponds to the text data sent in the message. In this example, we can see three different intervenients, a User, a Chatbot and an Assistant. If we take a closer look at all the messages sent by the chatbot, we can conclude that none of these messages contribute with useful data to the NER task we are trying to research in this work.

Figure 3.2 shows the same conversation as Figure 3.1 but with all the chatbot messages removed. By doing this, we were able to reduce the size of the conversation from 39 messages to 26, removing some of the useless data. Additionally, we can see that at a certain moment, the user just decides to talk to an assistant instead. By looking at the assistant messages, we can conclude that some of its messages may contain useful information to the purpose of anonymizing sensitive data. For example, the assistants always introduce themselves with their first and last names, in this case “Joana Silva” is a Person’s name, which is essential to the NER task of this work. Another thing we can conclude by looking at Figure 3.2 is that the messages are not really balanced in terms of total words used.

The dataset that we will consider in this work will only comprise messages sent by users, assistants, or supervisors, excluding all the messages sent by assistants or supervisor that were part of a conversation where the user was not a part of, resulting in a total of 34617 messages with an average of 54 words per conversation. Both Assistant’s and Supervisor’s messages were considered due to the fact they can contain sensitive information to be anonymized because they usually refer to the people they are talking to as their real name or even just confirm some information that was given to them earlier in the conversation.

Figure 3.3 illustrates the distribution of conversations with a certain total of words. By looking at the graph, we can draw two conclusions about the dataset:

1. Short conversations predominate: Most conversations in the dataset are short, with a word count between 1-199. This suggests that the dataset is heavily skewed towards shorter conversations which may impact the performance of the NER models, especially if they are sensitive to the length of the conversation.
2. Long Conversations are Rare: The number of conversations with a word count greater than 600 words is insignificant, indicating that long conversations are rare in the dataset.

Figure 3.4 has the same data as Figure 3.3 but the x-axis intervals were adjusted to provide a more balanced representation of the total number of conversation with a certain number of word in this dataset. Without even being grouped into intervals, we can confirm that conversations with fewer words are more dominate. It is worth to note that most conversations in this dataset have a total of only four words.

After extracting the wanted messages from the dataset, the resulting dataset was split into train (98%) and test (2%).

Privilege	Phase	Text
User	Chatbot	Prezada equipe do seguradoratipo, Encomendei o seguro do meu carro e aqui estão os detalhes da minha apólice: N.º Cliente: 0000000000 N.º Apólice: 0000000000 N.º Contribuinte: 0000000000 Para pagar o valor de 00.00 mencionei no formulário anexo via débito direto, quando entrei: ENTIDADE 000000 N.º AUTORIZAÇÃO 000000000000 Mostra um erro. Por favor, diga-me que me envie o número correto da Entidade para débito direto. Muito obrigado Cumprimentos Irfan Tanoli
Chatbot	Chatbot	Olá, sou o seguradoraN, o assistente virtual da seguradora. (Para voltar ao início escreva recomeçar)
Chatbot	Chatbot	(QUICKREPLY) - Posso ajudar em algum destes assuntos? (Dica: escreva o nome da opção que pretende. Ex.: Simulações)Gestão ContratualCond ParticularesReferência MBPpagamentosEnvio FaturaCarta VerdePedir AssistênciaSinistrosSimulaçõesAlteraçõesAnulação
User	Chatbot	Pagamentos
Chatbot	Chatbot	(QUICKREPLY) - Qual a opção pretendida?Referência MBPpagamento MBWAYDébito DiretoDevolução
User	Chatbot	Débito Direto
Chatbot	Chatbot	(TEXT WITH BUTTONS) - O Sistema de Débito Direto permite-lhe maior comodidade, pois o pagamento é feito através da sua conta bancária, mediante autorização prévia, com 00 de antecedência da data limite.E ainda beneficia de 0% de desconto. Caso pretenda saber o estado do pagamento do seu seguro, fale com um dos nossos assistentes.
User	Chatbot	Assistente
Chatbot	Chatbot	Para tratamento adequado da sua questão será encaminhado para um assistente - dias úteis das 0h às 00h.
Chatbot	Chatbot	(TEXT WITH BUTTONS) - Tratamento Dados PessoaisQueremos dar-lhe todas as garantias de que necessita para se sentir seguro ao utilizar o nosso serviço.Os dados recolhidos podem ser utilizados para celebração ou execução do contrato.Pode encontrar mais informação sobre a forma como tratamos os dados em: https://www.seguradora.pt/politica-privacidade-seguranca
Chatbot	Chatbot	Qual é o seu nome?
User	Chatbot	Continuar
User	Chatbot	Isac Tito
Chatbot	Chatbot	(TEXT WITH BUTTONS) - Confirma que este é o seu nome?
User	Chatbot	Sim
Chatbot	HOLD	Start Hold
Chatbot	HOLD	Aguarde um momento, um assistente irá responder-lhe assim que possível. (Para cancelar esta conversa escreva: CANCELAR)
Chatbot	HOLD	End Hold
Assistant	Assistant	Bom dia, está a falar com Joana Silva.
User	Assistant	Encomendei o seguro do meu carro e aqui estão os detalhes da minha apólice: N.º Cliente: 0000000000 N.º Apólice: 0000000000 N.º Contribuinte: 0000000000 Para pagar o valor de 00.00 mencionei no formulário anexo via débito direto, quando entrei: ENTIDADE 000000 N.º AUTORIZAÇÃO 000000000000 Mostra um erro. Por favor, diga-me que me envie o número correto da Entidade para débito direto. Muito obrigado Cumprimentos Irfan Tanoli
Assistant	Assistant	Tem de pagar no multibanco o primeiro pagamento. Sempre que o inicio do contrato ocorre nos 00 dias seguintes compra, o primeiro pagamento tem de ser feito no multibanco
Assistant	Assistant	Entidade: 00000 Referência: 0000000000
Assistant	Assistant	Valor(€): 00,00€
User	Assistant	ok
User	Assistant	Muito obrigado
Assistant	Assistant	O segundo pagamento já será feito por débito.
Assistant	Assistant	Existe mais alguma questão em que possa ajudar?
User	Assistant	quando faço o pagamento. Devo enviar e-mail ou Whatsapp?
User	Assistant	Se eu vender o carro e cancelar a apólice dentro de 0 meses, receberei o reembolso?
Assistant	Assistant	Quando fizer o pagamento, a carta verde será enviada para o email
Assistant	Assistant	sim, se o carro for vendido, poderá enviar o comprovativo da venda e será feito o reembolso pelo periodo não decorrido.
User	Assistant	muito obrigado por todas as perguntas
User	Assistant	bom dia
Assistant	Assistant	Obrigado pelo seu contacto, continuação de um bom dia.
Assistant	Chatbot	(EVENT - closeTicket) - Assistant Joana.Silva@seguradora.pt closed assistant conversation
Chatbot	Chatbot	(TEXT WITH BUTTONS) - Chat com assistente terminado.Se pretender receber um PDF com a conversação realizada escolha a opção abaixo.
User	Chatbot	PDF Conversa
Chatbot	Chatbot	Indique-nos o seu email
User	Chatbot	emailuser@email.com

Figure 3.1: Example of a full conversation between a user, a chatbot and an assistant.

Privilege	Phase	Text
User	Chatbot	Prezada equipe do seguradora tipo, Encomendei o seguro do meu carro e aqui estão os detalhes da minha apólice: N.º Cliente: 0000000000 N.º Apólice: 0000000000 N.º Contribuinte: 0000000000 Para pagar o valor de 00.00 mencionei no formulário anexo via débito direto, quando entrei: ENTIDADE 000000 N.º AUTORIZAÇÃO 000000000000 Mostra um erro. Por favor, diga-me que me envie o número correto da Entidade para débito direto. Muito obrigado Cumprimentos Irfan Tanol
User	Chatbot	Pagamentos
User	Chatbot	Débito Direto
User	Chatbot	Assistente
User	Chatbot	Continuar
User	Chatbot	Isac Tito
User	Chatbot	Sim
Assistant	Assistant	Bom dia, está a falar com Joana Silva.
User	Assistant	Encomendei o seguro do meu carro e aqui estão os detalhes da minha apólice: N.º Cliente: 0000000000 N.º Apólice: 0000000000 N.º Contribuinte: 0000000000 Para pagar o valor de 00.00 mencionei no formulário anexo via débito direto, quando entrei: ENTIDADE 000000 N.º AUTORIZAÇÃO 000000000000 Mostra um erro. Por favor, diga-me que me envie o número correto da Entidade para débito direto. Muito obrigado Cumprimentos Irfan Tanoli
Assistant	Assistant	Tem de pagar no multibanco o primeiro pagamento. Sempre que o inicio do contrato ocorre nos 00 dias seguintes compra, o primeiro pagamento tem de ser feito no multibanco
Assistant	Assistant	Entidade: 00000 Referência: 0000000000
Assistant	Assistant	Valor(€): 00,00€
User	Assistant	ok
User	Assistant	Muito obrigado
Assistant	Assistant	O segundo pagamento já será feito por débito.
Assistant	Assistant	Existe mais alguma questão em que possa ajudar?
User	Assistant	quando faço o pagamento. Devo enviar e-mail ou Whatsapp?
User	Assistant	Se eu vender o carro e cancelar a apólice dentro de 0 meses, receberei o reembolso?
Assistant	Assistant	Quando fizer o pagamento, a carta verde será enviada para o email
Assistant	Assistant	sim, se o carro for vendido, poderá enviar o comprovativo da venda e será feito o reembolso pelo periodo não decorrido.
User	Assistant	muito obrigado por todas as perguntas
User	Assistant	bom dia
Assistant	Assistant	Obrigado pelo seu contacto, continuação de um bom dia.
Assistant	Chatbot	(EVENT - closeTicket) - Assistant Joana.Silva@seguradora.pt closed assistant conversation
User	Chatbot	PDF Conversa
User	Chatbot	emailuser@email.com

Figure 3.2: Example of the same conversation as Figure 3.1 without the chatbot messages.

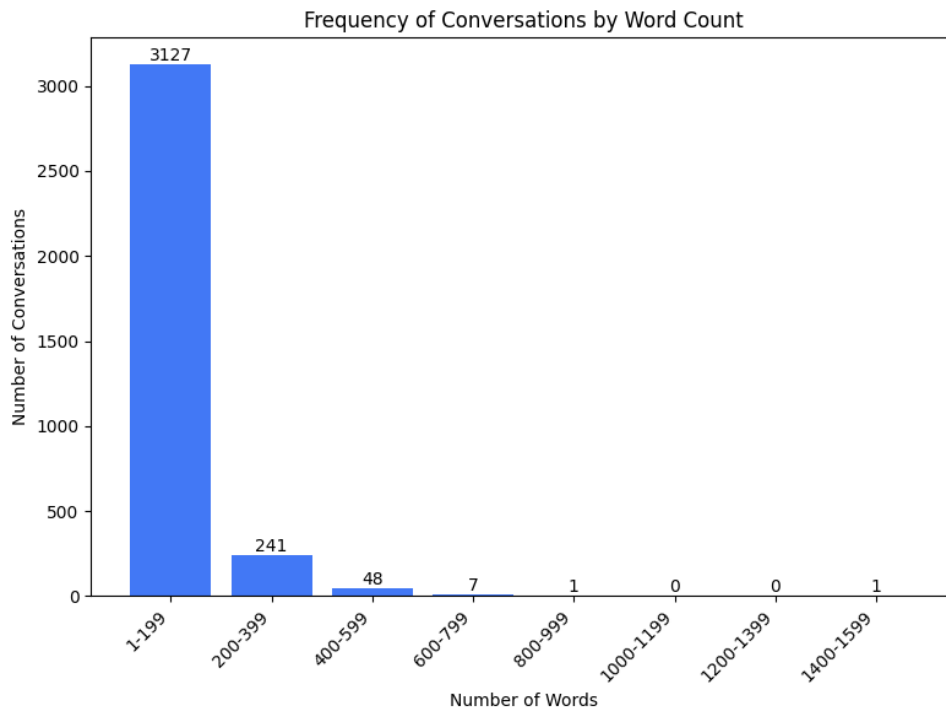


Figure 3.3: Total number of conversation that contain a certain number of words.

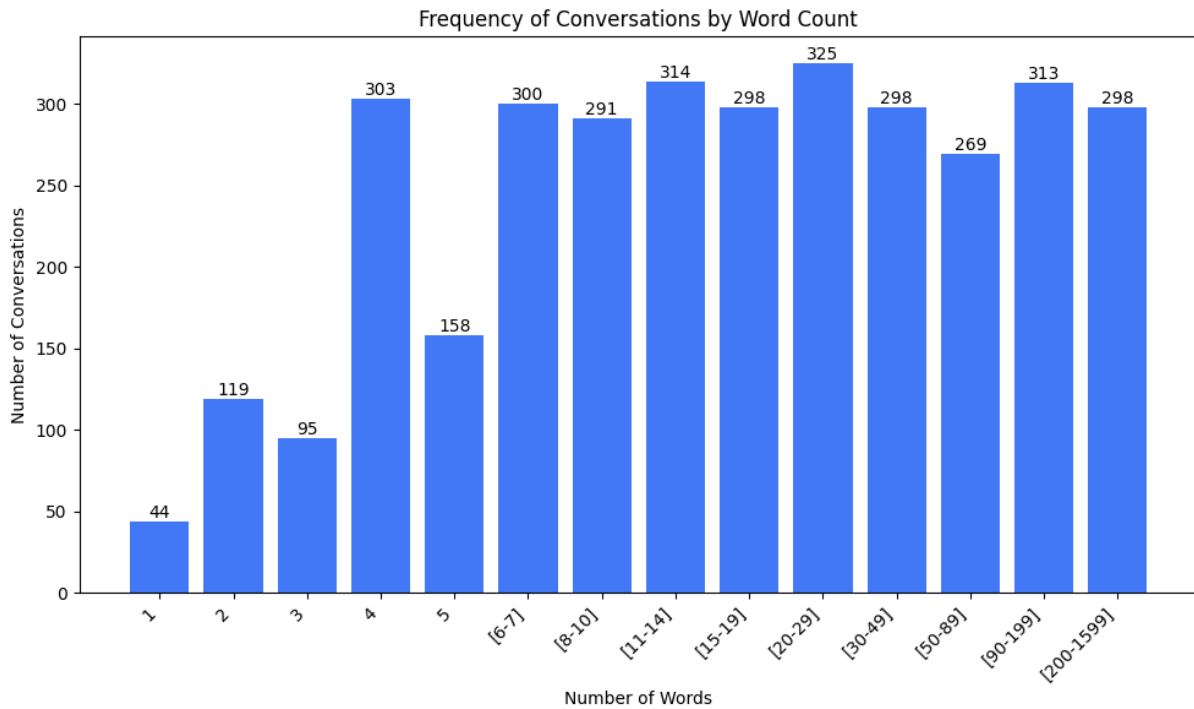


Figure 3.4: Total number of conversation that contain a certain number of words between certain intervals.

Even though we will not be training machine learning models with the training dataset (since it is not classified), we use it to find what kind of messages we are dealing with so that we can perform a better pre-processing of the data better and have a better knowledge of the vocabulary used.

The test dataset, composed with a total of 693 messages, was classified manually by two different persons (referred to as A and B) so that we can evaluate the tools proposed in Section 3.2.

It is important to have multiple people classify the same dataset for NER, especially for testing NER models. Some reasons for this fact are:

- **Quality Control:** Having two or more annotators classify the data helps in identifying discrepancies and potential errors. If annotators disagree, it highlights unclear or difficult cases that may require further examination or clearer guidelines for future annotations.
- **Insights into Ambiguity:** Entities can be expressed in various ways. Multiple annotators provide a wider perspective on how different people perceive and annotate entities.
- **Generalization and Bias Mitigation:** Multiple annotations help ensure that the model generalizes across different styles and contexts. Also having people from different backgrounds can help mitigate biased classifications leading to a more balanced evaluation.

3.4 Metrics

The calculation of metrics is a significant task in evaluating the work done in the field of NLP. However, different tasks typically require different methods of evaluation.

The main problem in NER systems evaluation is the definition of the entity, the importance of each class of entity and the divergence of existing entities, which makes the comparison with other works difficult [10]. However, the purpose of this dissertation is to consider the datasets given for the intended entity classes and, using metrics, comparing the resulting output of this work to the output of the dataset used. As such, all the entries that were classified were labeled as:

1. True Positives (TP) - Representing the terms that were classified correctly.
2. False Positives (FP) - Representing the terms that were classified as a NE when they should not.
3. False Negatives (FN) - Representing the terms that were not classified as a certain NE when they should have been.

In order to evaluate each model created to recognize each NE, the metrics of precision, recall and F1-Score were the ones chosen for that effect, as proposed by Sang and De Meulder [28]

The precision metric quantifies the number of correctly identified positive predictions made and its formula is:

$$Precision = \frac{TP}{TP + FP} \quad (3.1)$$

Recall is a measure that quantifies the amount of correct positive predictions made out of all possible positive predictions. Its formula is:

$$Recall = \frac{TP}{TP + FN} \quad (3.2)$$

The F1-Score metric is defined as the mean of precision and recall:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3.3)$$

3.5 Experiments and Results

The experiences conducted in this work have as their primary goal the identification of NE's, but only a specific few NE's. The NE's we will be looking to recognize in order to anonymize them later can be found using the different types of models described in Section 3.1 and Section 3.2. In this section, we will evaluate the tools, SpaCy, STRING, WikiNEuRal and RoBERTa, regarding the identification of the these named entities enumerated.

Using the metrics described in Section 3.4 we calculate the total number TP, FP and FN for each requested entity class and use them in the calculation of the enumerated metrics.

3.5.1 Second HAREM Golden Collection

In this section, we show the results of all the NER experiments conducted on the Second HaremGolden Collection. We focus on evaluating the performance of the NER models in recognizing specific entities within this domain, revealing the effectiveness of the chosen tools in recognizing sensitive entities.

As such, Table 3.5 and Table 3.6, and show the results when the respective models are applied to the Second Harem dataset.

Table 3.4: SpaCy metrics results for the recognition of each entity class using Second Harem dataset.

	Person	Location	Organization
Precision	0.794	0.523	0.484
Recall	0.627	0.734	0.416
F1-Score	0.700	0.611	0.448

Table 3.5: WikiNEuRal metrics results for the recognition of each entity class using Second Harem dataset.

	Person	Location	Organization
Precision	0.807	0.632	0.491
Recall	0.671	0.774	0.431
F1-Score	0.733	0.696	0.459

3.5.1.1 SpaCy

The application of the SpaCy model yielded good results for the NER task using the Second Harem dataset. By looking at Table 3.4 we can see that this model was better at identifying Person entities, achieving a precision score of 0.794, than Location and Organization entities. Also, we can see that this model demonstrated its ability in identifying a substantial fraction of the Location entities present in the text by looking at its recall value of 0.734. On the other hand, it is crucial to highlight that the performance of the SpaCy model in identifying Organization entities was low (compared to the other two), achieving scores of 0.484 and 0.416 for precision and recall respectively.

3.5.1.2 WikiNEuRal

The Babelscape WikiNEuRal model for NER revealed notable strengths and some areas for improvement. The model identified Person entities with high precision, and it also performed well in recognizing Location entities ending up with a recall score of 0.774.

However, the model’s performance in identifying Organization entities showed room for improvement, as we can see by looking at its scores in Table 3.5. This suggests the need for fine-tuning to improve the recognition of Organization entities.

3.5.1.3 RoBERTa

By looking at Table 3.6, we can see that the RoBERTa model achieved the best performance for the Person category, with a precision of 0.833 and a recall of 0.746. This suggests that the model properly identified most of the Person entities while also avoiding many false positives. Also, this model presents the best F1-Score for identifying Person entities, showing that the model recognized this type of entity more consistently and accurately.

Table 3.6: RoBERTa metrics results for the recognition of each entity class using Second Harem dataset.

	Person	Location	Organization
Precision	0.833	0.746	0.515
Recall	0.746	0.773	0.649
F1-Score	0.787	0.759	0.574

Nonetheless, the results obtained for recognizing Location entities were similar to the results obtained for recognizing Person entities, showcasing the model’s ability to demonstrate consistent performance across different entity types.

For the Organization category, the precision and recall scores illustrate that the model was able to capture most of the Organization entities in the dataset, but also made many false positive errors.

3.5.2 Seguradora Dataset

As we could see in Section 3.3.2, the Seguradora dataset is made up of exported messages from conversations between users, assistants and a chatbot from an insurance’s company. As such, certain language and vocabulary are used due to the theme of its conversations which might imply totally different results from the ones we saw for the Second Harem dataset.

Therefore, Table 3.7, Table 3.8, Table 3.9 and Table 3.10 provide an overview of the results obtained by applying each model to both classified datasets (A and B).

3.5.2.1 SpaCy

It is worth noting that the results shown in Table 3.7 were the results of using a pre-trained model available by SpaCy, more specifically the “pt_core_news_md” model, which was trained on written web text like blogs, news and comments.

As we can see by looking at Table 3.7, SpaCy was not able to recognize the NE’s that are rule-based, achieving a maximum of 0.6 score of precision when identifying the NE “Person” for both experiences.

3.5.2.2 STRING

Table 3.8 displays the resulting metrics for each NE available as a return of the NER model of the STRING tool.

These results show that the STRING model exhibits various degrees of performance across different entity categories. Notably, the model performs well at detecting email addresses and car plates, with high precision and recall values.

Table 3.7: SpaCy metrics results for the recognition of each entity class using Seguradora dataset.

	Precision	Recall	F1-Score
A			
Person	0.6	0.84	0.7
Location	0.065	0.6	0.119
Organization	0.136	0.25	0.176
B			
Person	0.586	0.82	0.683
Location	0.121	0.733	0.208
Organization	0.159	0.318	0.212

Table 3.8: STRING metrics results for the recognition of each entity class Seguradora dataset.

	Precision	Recall	F1-Score
A			
Person	0.272	0.885	0.417
Location	0.5	0.5	0.5
Organization	1.0	0.033	0.065
Car Plate	0.944	0.708	0.809
NIF	0.167	0.8	0.276
E-mail Address	1.0	0.76	0.864
B			
Person	0.266	0.865	0.407
Location	0.5	0.333	0.4
Organization	1.0	0.04	0.076
Car Plate	0.944	0.773	0.85
NIF	0.167	0.8	0.276
E-mail Address	1.0	0.76	0.864

However, certain entity classes such as Organization and NIF present challenges for this model. While the model achieves 100% precision for Organizations, the recall and F1-Score are significantly lower. This suggests that the model is having difficulties to generalize its recognition of organizational entities.

3.5.2.3 WikiNEuRal

Table 3.9 exhibits high recall rates for identifying Persons and Organizations for the WikiNEuRal model, meaning that it can recognize a significant portion of entities belonging to these categories. However, the precision values vary, indicating potential problems in achieving precise recognition of entities, especially for Locations.

Both Person and Organization classes have similar precision values and although the recall values for both are high, it should be balanced with the precision metric to fully assess the model's performance.

Table 3.9: WikiNEuRal model metrics results for the recognition of each entity class using Seguradora dataset.

	Precision	Recall	F1-Score
A			
Person	0.476	0.816	0.602
Location	0.081	0.5	0.139
Organization	0.408	0.952	0.571
B			
Person	0.464	0.796	0.586
Location	0.161	0.769	0.267
Organization	0.429	0.875	0.575

Table 3.10: RoBERTa metrics results for the recognition of each entity class using Seguradora dataset.

	Precision	Recall	F1-Score
A			
Person	0.506	0.863	0.638
Location	0.286	0.6	0.387
Organization	0.101	0.435	0.164
B			
Person	0.506	0.863	0.638
Location	0.619	0.867	0.722
Organization	0.131	0.52	0.209

In addition, we can see that this model struggles in identifying Location entities correctly by looking at its precision value. This reflects the model’s tendency to include false positives within this category.

3.5.2.4 RoBERTa

RoBERTa model’s performance was also evaluated across the entity categories of Person, Location and Organization, which were the only relevant ones to be analyzed.

From the results presented in Table 3.10 we can see that this model achieves a high Recall score for identifying the entities Person and Location, implying its capability to capture a considerable fraction of instances belonging to both categories.

On the other hand, this model’s performance in recognizing Organization entities appears to be more troublesome. We can check this fact by looking at the low values of both Precision and F1-Score. The low Precision is due to the fact the model is making many incorrect positive predictions to what should be an organization. The low F1-Score suggests that there is an imbalance between precision and recall, meaning it is not achieving a good trade-off between correctly identifying Organizations entities (Precision) and capturing all positive instances of Organization (Recall).

3.6 Discussion

In this section, we will conduct a comparison of the NER models based on the data presented in Section 3.5 using the results obtained from two distinct datasets: the Second Harem dataset and the Seguradora dataset. This analysis uncovers each model's particular strengths and weaknesses, offering information on their adaptability and performance in NER tasks.

When assessing the performance of the NER models on the Second Harem dataset, it was clear that RoBERTa consistently outperformed the others, achieving the highest values of precision, recall and F1-Score. RoBERTa's precision in detecting Person entities was 0.833, proving its ability in identifying person names. Additionally, RoBERTa obtained a recall value of 0.773 for Location entities, ensuring that most entities related with locations were captured. Also, we can see that all the models had some challenges in identifying Organization entities. Notably, RoBERTa emerged as the best performer in this category as well. Nonetheless, WikiNEuRal also produced respectable results with high precision and recall for Person and Organization entities.

On the other hand, the Seguradora dataset, characterized by domain-specific vocabulary, presented unique challenges for the NER models. In this case, RoBERTa continued to exhibit remarkable consistency across all different categories. This model repeatedly demonstrated decent precision and recall values for Location entities, showing its resilience in identifying potential location terms, by looking at its F1-Score of 0.722 for this category.

It is worth to note that the results using each model were different amongst them and that both classifiers, A and B, obtained similar results for each model used for evaluation. These poor results can be linked to the dataset nature and limited number of True Positives for each category. The STRING model had the advantage of being able to identify rule-based entities, while having decent scores for the recognition of Location entities, and poor results for the other classes. Even though it had a 100% precision score in the Organization category, by looking at its recall value of 0.033 we can conclude that this model only identified a small portion of Location entities included in the text.

Examining both datasets and the associated results, it becomes evident that the performance of each model is not inherently poor. Instead, it emphasizes the significance that the dataset used plays in achieving satisfactory results when performing the NER task.

In conclusion, the choice of the NER model should be a thoughtful decision, influenced by the dataset's nature and linguistic characteristics. As such, RoBERTa consistently outperformed the other models, demonstrating its adaptability to a wide range of entity recognition tasks.

Anonymization Tool

4

In this chapter, we will explore a solution for a system that can provide the anonymization service by giving a specific prompt.

The need to create this tool comes with the fact that most machine learning with NER modules do not have features to extract entities that are usually ruled based.

Because the primary goal of this work is to anonymize sensitive data, we must prepare a system capable of receiving unstructured data and returning the sensitive data anonymized while not altering the rest of the information presented. As a result, we divided this system into three distinct modules to accomplish this goal: pre-processing, NER, and anonymization.

We will start this chapter by describing how the API was built, followed by what the architecture of this system is composed by. Finally, we will be explaining what each module does within the anonymization flow.

4.1 The API

A REST-API was design using the extension Flask-RESTPlus of Flask¹.

Flask is a Web Framework written in Python that represents a collection of libraries and modules that enable developers to write web applications without worrying about low-level details such as protocol and thread management. Additionally the extension Flask-RESTPlus provides a coherent collection of decorators and tools to describe our API and display its documentation properly using Swagger².

Using this API we try to recreate the scenario of a user interacting with a chatbot. Although it does not exist a chatbot behind this API to answer the messages sent, it is returned the same message but anonymized, accomplishing the goal of this work.

We can see an example of how the API looks like in Figure 4.1.

¹<https://flask.palletsprojects.com/en/2.2.x/>

²<https://support.smartbear.com/swaggerhub/docs/>

Anonymization of Text Data ^{1.0}

[Base URL : /]
http://127.0.0.1:5000/swagger.json

This API tries to simulate what would be a conversation between a User and a Chatbot. Although there isn't a Chatbot behind this API to answer the messages sent, it is returned the same message but anonymized.

Anonymization

POST /Anonymization/anonymize_text

Parameters

Cancel

Name

Description

txt * required
string
(query)

txt

Execute

Figure 4.1: Screenshot of the API's endpoint needed to anonymize given text.

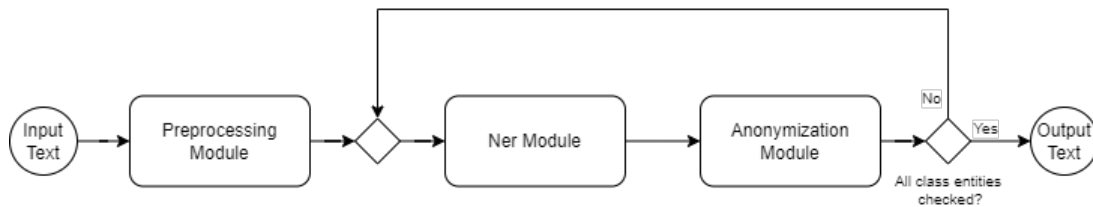


Figure 4.2: Flowchart of the API project anonymization pipeline.

4.2 System Architecture

In order to make this project work, a modular structure was used to make the project code as easy to understand as possible. When we use this type of architecture, we not only make it easier to add new functionalities (modules) without any additional cost, but also make the code more scalable and easier to read when compared to a monolithic architecture.

Firstly, our system will be waiting for the user to run the required endpoint in the API for us to get the input text needed. Secondly, the input will be processed in the processing module for it to be returned with the desired output. Lastly, the NER Module and the Anonymization Module will be responsible to take care of finding the sensitive terms within the input text and return the same sentence given as the input, but now anonymized. An example of how this architecture is shown in Figure 4.2.

4.3 Pre-processing Module

Pre-processing is crucial when dealing with great amounts of unstructured data. It is the first step in the text mining process [29].

By pre-processing the text, we are able to transform it to be formatted as our desired input for the subsequent tasks.

Due to the nature of the dataset, certain segments of automated messages produced by the chatbot were labeled as User messages, as we can see by the example bellow:

(EVENT - restart) - User restarted conversation

As such, we start the pre-processing by removing all messages generated by chatbot because they do not bring any useful information to the desired task.

Nextly, we noted that some words might contain pieces of html elements, so they were removed as well.

Finally, some words were connected by dots (“.”). Therefore, it was performed a regex to find all words that were connected with a dot, having in consideration terms that could be an email, and separate them to be in the desired format. For example, the term “Já recebi.Muito” would be rewritten to “Já recebi. Muito” and “user.user@gmail.com” would not be split.

We decided to not perform case normalization to the text due to the fact that some entities, like person names, should be written with an upper-case letter to start the word. Lowering case of all letters would make the task of NER more difficult to succeed.

Since this tool would be integrated in a chatbot to execute anonymization automatically, it would be bad to show the user a different message than what they wrote. For that reason, lemmatization, stemming and the removal of stopwords are also not performed.

4.4 NER Module and Anonymization Module

The input of the NER Module consists of the output of the Preprocessing Module.

SpaCy was chosen as the NER model implemented for this tool due its ability to identify non-rule-based entities and because its implementation is straightforward. This choice aligned perfectly with the goal of improving this tool’s ability to detect and anonymize a wide spectrum of sensitive data. Furthermore, to enhance this tool’s capabilities, a set of rule-based models were created, as discussed in Section 3.1, allowing the identification of additional sensitive entity types.

Both the NER Module and the Anonymization Module will be working together to deliver a more complete anonymized sentence. First, the NER module will find the first given class entity so that the anonymization module can anonymize them right away if they exist. The Anonymization Module will then return the anonymized string to the NER module so it can proceed with matching the next type of entity and the cycle goes on. All matching entities identified as sensitive data, will be anonymized with a specific term chosen for that entity. The steps of this pipeline are:

1. Matching and anonymization of emails to “_email_”.
2. Matching and anonymization of car plates to “_plate_”.
3. Matching and anonymization of phone numbers to “_phonenumber_”.
4. Matching and anonymization of NIF’s to “_nif_”
5. Matching and anonymization of person names to “_per_”, locations to “_loc_” and organizations to “_org_” using SpaCy.

After performing all these steps, a sentence given as input such as:

“O meu nome é Miguel e o meu email é user@gmail.com.”

would be anonymized to:

“O meu nome é _per_ e o meu email é _email_.”

As we can see by the example above, the anonymization we are using in this context is replacing the NE’s found with a label that can identify which class the term belongs.

4.5 Named Entity Recognition Results

It is also important to provide a comprehensive overview of the results produced by the NER component of this anonymization tool.

When processing each message withing the Seguradora dataset (Section 3.3.2) to identify the NE’s within, this tool performed exactly the same way as SpaCy when identifying entity types like Person, Location and Organization.

However, the distinctive aspect of the results obtained by this tool lie in the additional recognition of rule-based entity types, like E-mail Address, Car Plate and NIF, which were previously undetected. This tool utility and effectiveness in the context of data anonymization are enhanced by this increase of its entity recognition repertoire. Notably, all the rule-based entities listed in Section 3.1 were identified correctly achieving 100% in all the metrics calculated. This indicates that the tool has successfully bridged the gap, allowing the detection and anonymization of sensitive data items that SpaCy alone could not address.

5

Conclusions and Future Work

The work presented in this dissertation demonstrated that it is possible to combine Natural Language Processing techniques into real-world scenarios to solve challenges of sensitive data recognition and anonymization.

In this work we analyzed four different models capable of performing the NER task for the Portuguese language: SpaCy, STRING, WikiNEuRal and RoBERTa.

As we can see by looking at Figure 3.7 and Figure 3.8, the results of Precision and F1-Score were low for all entity types within the Seguradora dataset. This happened due to the fact that this dataset did not possess many terms that could be identified as sensitive, the number of True Positives for each class was really low compared to the number of False Positives. Even though the machine learning models used were able to identify most TP, they also identified way more entities than what they should have had. Even though the dataset used was classified manually, meaning there could be human error in the classification of the entities, the NER models used could be trained with data more suitable to the task we are performing, in this case, could be trained with messages more directly connected to the Business world and not just from blogs or even news.

Another reason for the lack of TP in some entity classes comes from the fact that some data were already “anonymized”, for example, the message:

“Qual o email que pretende que fique associado ao nif 280284061?”

should contain a valid NIF at the end, however it seems that some digits of the NIF were changed in order to protect the privacy of the person to whom this message was referring to.

One notable aspect of this research is the disparity in results obtained when applying the same NER models to two distinct datasets. The Seguradora dataset produced comparatively poor results, which strongly contrasted with the results produced by the NER models when applied the Second Harem dataset. This discrepancy can be attributed to lack of information within the Seguradora dataset, as opposed to the broader and more diverse content of the Second Harem dataset.

Regarding the second research question defined in 1.2, the complete study of multiple NER models across different datasets revealed RoBERTa as the best performer, consistently

delivering superior scores for precision, recall and F1-Score. This outcome establishes RoBERTa as the best NER model for our particular use case, demonstrating its robustness and reliability in the context of identifying sensitive entities in Portuguese texts.

Therefore, for future work, we propose to train the machine learning models with more suitable data using SpaCy, for example, which allows the training of the NER model using our own annotated data. Also, the dataset used for training and testing needs to be bigger and with more useful data related to the task we are performing and preferably with data not already anonymized in it.

Additionally, there are different types of anonymization techniques, as we saw in Section 2.2.6, that can be implemented as well. The anonymization technique used in this work was a simple masking strategy where we swap the term by a specific mask depending on the entity category. For future work, the different types of anonymization techniques could be implemented in order to provide the user a broader array of anonymization options so that the user could choose which one he wants to use depending on his requirements and preferences.

Finally, the API can also be extended as future work to include more entities that can be identified as sensitive and that were not included in this work due to the fact they did not exist once within the dataset used, like religious beliefs and sexual orientation. In the same sense, the API could implement different types of NER models and make it optional to choose which one we want to use to perform the NER task to identify sensitive entities.

Bibliography

- [1] 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)*, 2021, pp. 4188–4203.
- [2] D. P. Lacerda, A. Dresch, A. Proença, and J. A. V. Antunes Júnior, "Design science research: A research method to production engineering," *Gestão & produção*, vol. 20, pp. 741–761, 2013, publisher: SciELO Brasil.
- [3] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.
- [4] R. Pereira and J. Serrano, "A review of methods used on IT maturity models development: A systematic literature review and a critical analysis," *Journal of information technology*, vol. 35, no. 2, pp. 161–178, 2020, publisher: SAGE Publications Sage UK: London, England.
- [5] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.
- [6] N. Mamede, J. Baptista, and F. Dias, "Automated anonymization of text documents," in *2016 IEEE congress on evolutionary computation (CEC)*. IEEE, 2016, pp. 1287–1294.
- [7] N. Mamede, J. Baptista, C. Diniz, and V. Cabarrão, "String: an hybrid statistical and rule-based natural language processing chain for Portuguese," 2012, publisher: Cite-seer.
- [8] C. Mota and D. Santos, "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem," 2008.
- [9] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, S. Kannan, and V. Gurusamy, "Preprocessing techniques for text mining," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2014.
- [10] M. R. Dias, "Discovery of sensitive data with natural language processing," PhD Thesis, 2019.

- [11] F. M. C. Dias, “Multilingual automated text anonymization,” *Instituto Superior Técnico of Lisboa*, 2016.
- [12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, “Moses: Open source toolkit for statistical machine translation,” in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*. Association for Computational Linguistics, 2007, pp. 177–180.
- [13] A. Mansouri, L. S. Affendey, and A. Mamat, “Named entity recognition approaches,” *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339–344, 2008, publisher: Citeseer.
- [14] P. Gamallo, M. García, and P. Martín-Rodilla, “NER and Open Information Extraction for Portuguese: Notebook for IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks.” in *IberLEF@ SEPLN*, 2019, pp. 457–467.
- [15] P. Gamallo, M. Garcia, C. Pineiro, R. Martinez-Castano, and J. C. Pichel, “LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction,” in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2018, pp. 239–244.
- [16] K. Clark, M.-T. Luong, C. D. Manning, and Q. V. Le, “Semi-supervised sequence modeling with cross-view training,” *arXiv preprint arXiv:1809.08370*, 2018.
- [17] O. Bridal, “Named-entity recognition with bert for anonymization of medical records,” 2021.
- [18] M. Malmsten, L. Börjeson, and C. Haffenden, “Playing with words at the national library of sweden—making a swedish bert,” *arXiv preprint arXiv:2007.01658*, 2020.
- [19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.
- [20] S. S. Sonawane and P. A. Kulkarni, “Entity based co-reference resolution with name entity recognition using hierarchical classification,” in *2015 Annual IEEE India Conference (INDICON)*. IEEE, 2015, pp. 1–6.
- [21] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, “Coreference resolution: A review of general methodologies and applications in the clinical domain,” *Journal of biomedical informatics*, vol. 44, no. 6, pp. 1113–1122, 2011, publisher: Elsevier.
- [22] D. Vrandečić, “Wikidata: A new platform for collaborative data collection,” in *Proceedings of the 21st international conference on world wide web*, 2012, pp. 1063–1064.

- [23] S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "A comparative study of data anonymization techniques," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2019, pp. 306–309.
- [24] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, and R. Navigli, "Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2521–2533.
- [25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: a robustly optimized bert pretraining approach (2019)," *arXiv preprint arXiv:1907.11692*, vol. 364, 1907.
- [26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.
- [27] C. Freitas, P. Carvalho, H. Gonçalo Oliveira, C. Mota, and D. Santos, "Second harem: advancing the state of the art of named entity recognition in portuguese," in *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. European Language Resources Association, 2010.*
- [28] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.
- [29] S. Vijayarani, M. J. Ilamathi, M. Nithya, and others, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.