University Institute of Lisbon

Department of Information Science and Technology

Discovery of sensitive data with Natural Language Processing

Mariana Rebelo Dias

A Dissertation presented in partial fulfillment of the Requirements for the Degree of

Master in Computer Engineering

Supervisor

Doctor João Carlos Amaro Ferreira, Assistant Professor

ISCTE-IUL

Co-Supervisor

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Assistant Professor

ISCTE-IUL

October, 2019

Resumo

O processo de preservação de dados sensíveis está em constante crescimento e cada vez apresenta maior importância, proveniente especialmente das diretivas e leis impostas pela União Europeia. O esforço para criar sistemas automáticos é contínuo, mas o processo é realizado na maioria dos casos de forma manual ou semiautomática. Neste trabalho desenvolvemos um componente de Extração e Classificação de dados sensíveis, que processa textos não-estruturados em Português Europeu. O objetivo consistiu em criar um sistema que permite às organizações compreender os seus dados e cumprir com fins legais de conformidade e segurança. Para resolver este problema, foi estudada uma abordagem híbrida de Reconhecimento de Entidades Mencionadas para a língua Portuguesa. Esta abordagem combina técnicas baseadas em regras e léxicos, algoritmos de aprendizagem automática e redes neuronais. As primeiras abordagens baseadas em regras e léxicos, foram utilizadas apenas para um conjunto de classes especificas. Para as restantes classes de entidades foram utilizadas as ferramentas SpaCy e Stanford NLP, testados dois modelos estatísticos — Conditional Random Fields e Random Forest – e por fim testada uma abordagem baseada em redes neuronais – Bidirectional-LSTM. Ao nível das ferramentas utilizadas os melhores resultados foram conseguidos com o modelo Stanford NER (86,41%). Através dos modelos estatísticos percebemos que o Conditional Random Fields é o que consegue obter melhores resultados, com um f1-score de 65,50%. Com a última abordagem, uma rede neuronal Bi-LSTM, conseguimos resultado de f1-score de aproximadamente 83,01%. Para o treino e teste das diferentes abordagens foram utilizados os conjuntos de dados HAREM Golden Collection, SIGARRA News Corpus e DataSense NER Corpus.

Palavras-chave: Dados Sensíveis, Processamento de Língua Natural, Reconhecimento de Entidades Mencionadas, Regulamento Geral de Proteção de Dados, Projeto DataSense

Abstract

The process of protecting sensitive data is continually growing and becoming increasingly important, especially as a result of the directives and laws imposed by the European Union. The effort to create automatic systems is continuous, but in most cases, the processes behind them are still manual or semi-automatic. In this work, we have developed a component that can extract and classify sensitive data, from unstructured text information in European Portuguese. The objective was to create a system that allows organizations to understand their data and comply with legal and security purposes. We studied a hybrid approach to the problem of Named Entities Recognition for the Portuguese language. This approach combines several techniques such as rule-based/lexical-based models, machine learning algorithms and neural networks. The rule-based and lexical-based approaches were used only for a set of specific classes. For the remaining classes of entities, SpaCy and Stanford NLP tools were tested, two statistical models – Conditional Random Fields and Random Forest – were implemented and, finally, a Bidirectional-LSTM approach as experimented. The best results were achieved with the Stanford NER model (86.41%), from the Stanford NLP tool. Regarding the statistical models, we realized that Conditional Random Fields is the one that can obtain the best results, with a f1-score of 65.50%. With the Bi-LSTM approach, we have achieved a result of 83.01%. The corpora used for training and testing were HAREM Golden Collection, SIGARRA News Corpus and DataSense NER Corpus.

Keywords: Sensitive Data, Natural Language Processing, Named Entities Recognition, General Data Protection Regulation, DataSense Project

Acknowledgements

First, I would like to acknowledge my supervisors, Professor João Ferreira and Professor Ricardo Ribeiro for their support and assistance. Thank you for providing the knowledge, suggestions and scientific contribution that made this work possible.

I want to thank all the people of INOV-INESC Inovação, especially Rui Maia and Pedro Santos, for their advice and support, for challenging me in this work and always encouraging me to new challenges. I am very grateful for the opportunity.

A big thank you to my family. I want to thank my sister for her support and all the words of strength. A very special thanks to my parents Manuela and Agostinho who supported me unconditionally, inspired me and believed in me even more than I did.

Mariana Rebelo Dias

Contents

Abbreviations

ACE	**A**utomatic **C**ontext **E**xtraction
AI	**A**rtificial **I**ntelligence
API	**A**pplication **P**rogramming **I**nterface
CNN	**C**onvolutional **N**eural **N**etwork
CoNLL	**C**onference on **N**atural **L**anguage **L**earning
CRF	**C**onditional **R**andom **F**ields
FSM	**F**inite **S**tate **M**achine
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
HAREM	Named Entity Recognition Systems Evaluation
HMM	**H**idden **M**arkov **M**odel
HTML	**H**yper**T**ext **M**arkup **L**anguage
IE	**I**nformation **E**ntity
LSTM	**L**ong **S**hort-**T**erm **M**emory
MEMM	**M**aximum **E**ntropy **M**arkov **M**odel
ML	**M**achine **L**earning
MUC	**M**essage **U**nderstanding **C**onference
NE	**N**amed **E**ntity
NER	**N**amed **E**ntity **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing
NLTK	**N**atural **L**anguage **T**ool **K**it
NN	**N**eural **N**etwork
OCR	**O**ptical **C**haracter **R**ecognition
PII	**P**ersonally **I**dentifiable **I**nformation
POS	**P**art **O**f **S**peech
RNN	**R**ecurrent **N**eural **N**etwork
SM	**S**tatistical **M**odel
XML	e**X**tensible **M**arkup **L**anguge

List of Figures

List of Tables

Chapter 1

Introduction

The amount of information available on the web as well as in companies and other sectors is getting bigger, therefore it is a need for filtering and processing so that information can be used for a purpose. The vast majority of existing documents and information is unstructured, requiring even more processing efforts to overcome these difficulties. Natural Language Processing (NLP) is an area of Artificial Intelligence (AI) that studies problems of automatic understanding and generation of human language, whether spoken or written [Dale et al.2000]. This allows obtaining from unstructured text documents, information that can be used by machines. Associated with the NLP area is Information Extraction (IE), and the main task of IE is to extract data from documents, which can be structured or unstructured texts, and one of its main subtasks it is the Named Entity Recognition (NER) [Nadeau and Sekine2007].

The NER task and concept of Named Entity (NE) appeared for the first time in the Message Understanding Conference-6 (MUC-6) [Grishman and Sundheim1996]. Over the years there have been some redefinitions of named entity recognition task and of entities classes, but the initial concept remains until today. The topic has been evolving and continues to be researched. In 2005, the HAREM [Santos and Cardoso2007] was the first joint evaluation of recognition systems of named entities in document collections written in Portuguese. The emergence of obligations for processing sensitive data has been increasing the focus on the advancement of NER. However, for languages with fewer resources, such as the Portuguese language, it is still a challenge and the results are still quite inferior when compared to English, for example. This work strives to evaluate these problems focusing on the research, implementation, and evaluation of NER systems for the Portuguese language, with the intent to build a reliable solution that can be used by organizations in a real scenario. However, the main focus is on the topic of sensitive data, and on the discovery and classification of entities corresponding to sensitive and personal information data.

1.1 Overview

This dissertation presents an effort in the study of the Natural Language Processing task, NER, and was jointly developed at ISCTE-IUL and INOV-INESC Inovação. This dissertation is part of a project financed by the Portugal2020 programme, the DataSense Project[1]. This project aims to provide a set of organizations with solutions for the treatment of sensitive data applied to the European Portuguese Language. In addition, the results achieved in this dissertation were also used in the SocialOpinion Project developed at INOV and presented at an ANI (Agência Nacional de Inovação) conference. The example of SocialOpinion Project is presented in the Appendix section, Figure 7.1.

The proposed work aims to transform many of the processes that can be carried out manually and with high cost into automatic processes that can carry out efficiently. It allows organizations to save resources and time, have confidence in the security of their data and in compliance with protocols and regulations imposed, as is the case of the General Data Protection Regulation (GDPR). A prerequisite of this work is that these tasks are done automatically through the application of Natural Language Processing and Machine Learning techniques. One of the greatest challenges of this work is the application of text processing techniques to the Portuguese Language.

The main feature of this dissertation is the development of a module based on NLP techniques to integrate into the DataSense Project. This module focuses on named entities recognition in unstructured textual documents. The aim is to improve the existing results for entity recognition and to use recent state-of-the-art techniques.

The work carried out enables many organizations to better manage their documents and customers' sensitive data, which leads to better security and compliance with the standards defined by the European Union regarding personal data. The module resulting from the work of this dissertation was already integrated into the first and second versions of DataSense and used for commercial purposes.

1.2 Motivation

In recent years, and in particular, since the year 2000, we have seen a growth in the amount of data and documents generated on a large scale, which implies an increase in textual information that is often unstructured [Chen et al.2014]. This brings new security concerns regarding information availability, particularly the way organizations handle sensitive data. The concept of Sensitive

[1]The DataSense Project is being led by Link Consulting and received co-funding from the FEDER - Lisbon 2020, PT 2020, European Union's PT 2020 research and innovation program under grant agreement cod POCI-01-0247-FEDER-038539.

Data or Sensitive Information may follow various points of view, depending on the context and purpose. In May 2018, the directive of GDPR appeared to regulate the processing of personal data in the European Union [Albrecht2016].

The process of recognizing sensitive data is still a task that is often carried out manually, respecting certain rules, which implies additional time spent and a bigger probability of errors and failures. The need to solve this type of problem in an automatic way has become ever greater, and this leads to the need to use more intelligent methodologies than those previously used. Due to this, there has been a great advance in the application of NLP tasks in the real world, and there are already some approaches and systems that work in the area of data discovery, such as Logikcull [Johnson2019][2] and Onna [Takatsuka et al.2007][3]. Despite the advances and the encouraging progress in NER, most of the real systems developed base their classification on the document's metadata instead of classifying the content [Clough2005]. Another problem is that most automatic approaches are made only for English, or other commonly used languages. Automated approaches are often dependent on existing data, regardless of language, but for many languages there is no data available, as is the case for Portuguese, so the results for Portuguese are far behind. For this context, on last realized NER events for each language, the f1-score results for English exceed 88.76% [Sang and De Meulder2003], while for Portuguese the best results do not go beyond 79% [Mota and Santos2008]. This divergence is also clear in relation to the available corpora, where for English and other languages there are dozens available, for European Portuguese, there is only one available, to our best knowledge. Therefore, for Portuguese there are many tasks at the level of recognition of sensitive data that are still performed manually, for example:

- Detection of sensitive information in particular according to the GDPR Directive;
- Recognition and Classification of sensitive data in documents with unstructured text information, without using metadata;
- Obtaining information on the type of sensitive data present in a document.

The possibility of automating this type of task would greatly facilitate compliance with security rules and imposed regulations. If we manage to overcome these limitations, the practical applications in other projects besides DataSense would be countless, and to several markets.

1.3 Objectives

This work's main objective is the implementation of a Named Entity Recognition Component for the DataSense Project, using Natural Language Processing applied to the Portuguese language.

[2]See `https://logikcull.com` (last visited 01-09-19)
[3]See `https://onna.com` (last visited 02-10-19)

As shown in Figure 1.1, this is a central component of the NLP Engine Module since all the other components in the module depend on it.

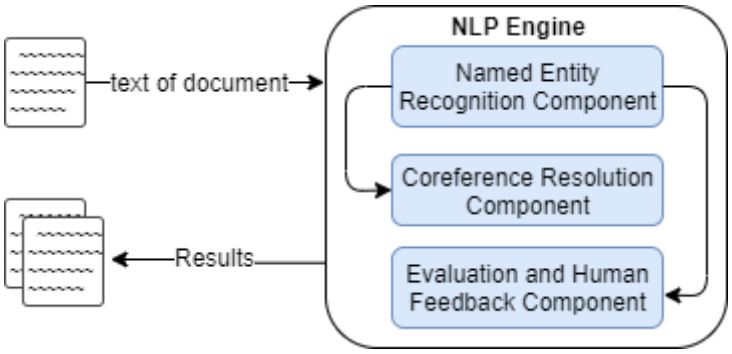

FIGURE 1.1: NLP Engine Module Architecture

The NER Component requires the development of a system for sensitive information discovery in text documents. The set of textual information to be processed by the component consists of legal documents, contracts, curricula, minutes, etc. Therefore, this dissertation should focus on the study of the NER task for sensitive data, as well as in all involving natural language processing tasks, more specifically in text preprocessing techniques, such as Part-of-Speech Tagging. With this study, we aim to achieve results that allow us the integrate the developed work into a real-world product, the DataSense Project. In addition to the core development of the NER Component and the discovery of sensitive data, it is also important that the component is developed in a modular and scalable way so that it can be used in other fields with different customizable characteristics.

In summary, the main objectives are the following:

1. Find the Natural Language Processing techniques that provide the best results for unstructured text processing in European Portuguese Language. The texts to be processed should be taken from the information repositories of the organizations, such as contracts, minutes, curricula, legal documents, etc.

2. Study the Preprocessing methods in Portuguese that may facilitate the extraction of sensitive information and to achieve better results in the task of Named Entities Recognition.

3. Identify and Classify Named Entities in the scope of sensitive data and General Data Protection Regulation personal information. The automation of processes will allow organizations to have greater confidence in their data and to comply with laws more efficiently and rigorously.

4. Develop a textual information processing component that allows for answering immediate questions about the content of the documents. In order for the user to be able to know immediately how much sensitive data is present in the text, how many types of data are present in the

text per category of sensitive data, and what are the most frequent classes of sensitive data in the text of each document.

5. Implement a scalable solution adaptable to different contexts and other types of data. This solution should be adaptable to other projects within the same scope and should be accessible simultaneously to multiple users. Data types, such as sensitive data classes, should not be restrictive, and there should always be the possibility to add the extraction of other sensitive data classes.

Thus, the motivation for this work is to understand how state-of-the-art approaches for other languages perform when applied to the Portuguese language. We also focus on trying to overcome existing problems through natural language processing, machine learning techniques and exploring deep learning concepts. By testing different approaches and performing different experiments, the best methods and techniques will be chosen that are most appropriate to this problem. And finally, be able to create a Named Entity Recognition Component to integrate into the DataSense project that achieves and satisfactory results.

1.4 Outline of the Dissertation

This document consists of seven chapters, including the Introduction (chapter 1), with the following structure:

Chapter 2 is focused on the state-of-the-art, more specifically in Natural Language Processing, and in Portuguese language problems. This chapter is a literature review of previous work and existing systems in the same areas of the present dissertation. The research areas include preprocessing techniques focused on Part-of-Speech Tagging, Named Entity Recognition and Information Extraction methods.

Chapter 3 provides the basic supporting concepts related to Sensitive Data, Personal Data and General Data Protection Regulation to better understand the concepts behind this dissertation, as well as a detailed description of the DataSense Framework specifying how this dissertation fits into it.

Chapters 4 presents the work developed in the Preprocessing, Named Entity Recognition and Post-processing areas. The decisions made and the development of the NER component are both explained in detail. Also present in this chapter are the tools, techniques, and models used for this work.

Chapter 5 consists of the description of existing metrics and resources. In other words, the evaluation metrics used to evaluate the results obtained and the existing resources at the level of the available datasets and corpus that helped in the preparation of this work.

Chapter 6 presents the entire evaluation of the work and the results obtained based on the metrics of chapter 5. The results are analyzed, discussed, and compared with other studies and results for the same data.

Finally, in the Conclusion, we discuss what was achieved with the present work so far, and identify the remaining future challenges.

Chapter 2

Related Work

The discovery of Sensitive Data or Personal Information follows different approaches, depending on the problem and the final goal. Thus, different approaches have emerged to deal with the automatic discovery of sensitive data and information extraction [Korba et al.2008]. Many of these approaches aim to anonymize sensitive data, but anonymization tasks go through the early stages of data detection and classification. This detection and classification, in the context of unstructured text information data, are performed using Natural Language Processing techniques, more specifically Named Entity Recognition [Cardie1997] [Ciravegna2001].

In this chapter, we summarize the current methods of data detection and classification and introduce other systems with similar purposes to ours. We begin by explaining what NER is and address the main existing works. NER methods can be divided into two main approaches, the first being based on the manual creation of sets of rules or the use of lexicons/dictionaries information. The second approach appears around the year 2002, with CoNLL 2002, where some new approaches emerged presenting the named entity recognition based on machine learning tasks [Klein et al.2003]. In addition, there are also works aimed at hybrid approaches, combining the two previous techniques in order to achieve better results. In this chapter, on top of the overview of the theme, we focus on the study and existing work for the Portuguese language.

2.1 Named Entity Recognition

Named Entity Recognition (NER) is a subtask of the Information Extraction (IE) task, in the context of Natural Language Processing (NLP). The purpose of NER is to enable the identification and classification of entities in unstructured text according to a set of predefined categories. Named Entity (NE) is a real-world object, such as people or places, are words ('entities') that can be classified ('named') according to a set of categories. NEs consist of terms present in a

text and that can be composed of one or more tokens [Nadeau and Sekine2007], but the concept of "entity" changes from approach to approach. The most common entity types are First Names (Personal Names, Organizations, Places), Time (dates, times), Numerical Entities (percentages, monetary values, measurements) or Personal Data (mobile phone numbers, medical and criminal history). The NER task not only identifies the entities but also classifies them according to predefined categories, however, the identification and classification are closely related problems [Sureka et al.2009]. Different NLP techniques are applied, which consist of identifying the keywords present in the text and classifying them. The NER task may follow different approaches and might have a very broad set of entity categories. The first categories appeared in MUC-6: people, places, organizations, time, and numerical expressions. This set of categories has been the most common ever since, however, other categories have been added.

MUC-6 (Message Understanding Conference 6) [Grishman and Sundheim1996] was one of the first conferences to introduce the task of Named Entities Recognition, focusing on the English language. This first event, in 1995, involved the recognition of the following types of entities: people, organizations, and places. Later, are also considered temporal entities (date and time) and numerical measures (money and percentage) [Grishman and Sundheim1995]. In the MUC events, the participating systems were scored according to two distinct axes. The first was relative to the ability to find the limits of an entity independently of a class and the other relative to the classification with the correct category. The result of each participant was calculated using the sum of both axes for the precision, recall and f1-score metrics. Twenty different systems participated in the MUC-6 where the best result was obtained with a f1-score of 96.42% [Sundheim1995]. Most participants' approaches consisted of simple tokenization tasks, using small gazetteers, word dictionaries and discovery of simple patterns. In general, the best-performing systems used semantic patterns for recognizing verbs and names [Grishman1995]. However, the best global result for identification and classification was achieved with the use of a Hidden Markov Model (HMM) [Bikel et al.1999]. After the MUC-6 other events and conferences focused on the NER task emerging.

As MUC successor, the MET [Merchant et al.1996] appeared in 1996, which directly adapted the MUC's task but for Japanese, Spanish and Chinese. Later in 2002, with a slightly different approach to the NER task, took place the Conference on Natural Language Learning (CoNLL) [Tjong Kim Sang2002a]. CoNLL is an annual conference that began in 1999, with two editions focused on named entities: in 2002 [Tjong Kim Sang2002b] and 2003 [Sang and De Meulder2003]. In this conference were evaluated systems in English, German and Spanish, and unlike the evaluation carried out at MUC, the participants were only scored with the exact identification and classification of the entities. In other words, this evaluation method was more demanding, it was necessary to correctly classify the entity limits and the category simultaneously. The recognition of entities for both the 2002 and 2003 events focused on four categories: People, Locations, Organizations and Miscellaneous. It was at CoNLL in 2002 that the IOB entity tagging format

was first used, also called the CoNLL format [Tjong Kim Sang2002b]. This format consists of writing down each token with the corresponding tag: Inside, Outside or Begin. Words tagged with O are outside of named entities, and the I-XX tag is used for words inside a named entity of type XX. Whenever two entities of type XX are immediately next to each other, the first word of the second entity will be tagged B-XX in order to mark the beginning of another entity. In addition to the IOB tags and the categories of each entity, the data used in CoNLL also had each token annotated with Part-of-Speech Tagging tags. The 2002 conference, for Spanish and German, had twelve participants, where the best results of f1-score for the Spanish test was 81.39% and for German 77.05% [Carreras et al.2002]. In 2003 there were sixteen participants, and the best results were achieved for the English test, with a f1-score of 88.76% [Florian et al.2003]. In these events, in contrast to what happened previously at MUC-6, most participants did not rely on simple approaches but based their work on supervised learning, namely, Maximum Entropy Models, and Hidden Markov Models [Sang and De Meulder2003]. However, the best results were achieved by hybrid approaches that combined machine learning methods and hand-coded methods.

After MUC, MET, and CoNLL other events linked to the NER task emerged, such as the ACE program [Doddington et al.2004]. However, all often had a great focus on English or other languages different from Portuguese, until 2005 in which HAREM focused on NER exclusively for the Portuguese language. HAREM [Santos et al.2006] is one of the largest baselines for NER in the Portuguese language. It started as a competition for the evaluation of NER, with two main events, in 2005 and 2008. In these events, were added types and subtypes of entities, in addition to the goal of identifying entities and classifying them into categories, for the classification was also considered types and subtypes. That is, in the case of the entity Lisbon, was assigned the category Local, the type Human and the subtype Region. This assessment was therefore based on the identification of the entity, the classification into categories, types and subtypes and also the semantic and morphological classification of the symbol. In the first event in 2005 [Santos and Cardoso2007], only the task of named entities was considered [Santos et al.]. However, in 2008, it also focused on the semantic relations between entities [Freitas et al.2010]. In both events, the evaluation was based on Portuguese texts of various genres, such as newspapers and web pages. The entity categories were divided into ten: Obra (Work), Acontecimento (Event), Organização (Organization), Pessoa (Person), Abstração (Abstraction), Tempo (Time), Valor (Value), Local and Coisa (Thing). Both HAREM events had ten participants, most of the participating systems used a hand-coded approach, based on rules and simple approaches. In the first HAREM, only two systems presented machine learning approaches using supervised learning knowledge-based into Spanish [Ferrández et al.2007], [Solorio2007]. The best NER score for this event was a 58% f1-score, using a rules-based approach. For the second HAREM the picture was similar, with only one machine learning-based approach, the R3M system [Mota2008]. This system was based on the training of semi-supervised learning models that used a co-training

algorithm to infer classification rules. However, the best results were obtained by the Priberam system [Amaral et al.2008], using a rules-based approach, achieving an f1-score of 57%, and the REMBRANDT system [Cardoso2008], that also using a hand-coded approach, based on Wikipedia achieving 56%, similar to those achieved for English, with the same kind of approach. From these events, also emerged some annotated corpus that enabled the continuation and evolution of the study of NER tasks for the Portuguese language. The corpus used for the evaluation on HAREM remains the only official annotated corpus freely available. There are numerous variations of the corpus used for the two evaluations, being the most complete the HAREM Golden Collection [Santos and Cardoso2006].

Since these conferences, new works and systems with the same features and corpus continued to appear, but with better results over the time [Derczynski et al.2017]. The most notable case is the existing work for the English CoNLL corpus, in which the current results already exceed by 10% of the ones achieved in 2003 [Baevski et al.2019]. Initially, the NER task was done using approaches based on manual rules [Collobert and Weston2008], which establishes a specific structure for a domain and requires intense work and human experience to create rule patterns [Appelt et al.1993]. Recent approaches to the NER task emerge, and that is already adaptable to different domains and data types, based on Machine Learning techniques [Lample et al.2016]. The most commonly used machine learning techniques in the context of NER are probabilistic methods such as Hidden Markov Models (HMM) [Ponomareva et al.2007], Maximum Entropy Markov Models (MEMM) [Borthwick et al.1998], Conditional Random Fields (CRF) [Teixeira et al.2011] and Random Forest Models [Magge et al.2018]. Even for the Portuguese language, since 2008, many authors have achieved good results through the use of Machine Learning approaches, more specifically with Conditional Random Fields models [Amaral et al.2013]. However, the best results for NER tasks, in general, are obtained through hybrid approaches, combining methods based on hand-coded techniques and Machine Learning techniques [Fresko et al.2005]. Recently, some NER experiments have achieved optimal results with the use of Neural Networks and Deep Learning [Yadav and Bethard2018]. The vast majority of entity recognition studies are based on Long Short-Term Memory (LSTM) and its variants.

Considering existing work, we can see a clear division in the techniques used for NER tasks. In this section, we present and explain the main methods for extracting entities: hand-coded techniques and machine learning techniques.

2.1.1 Hand-Coded Techniques

Named entity extraction methods based on Hand-coded techniques can follow two distinct approaches:

- Methods based on rules or grammatical patterns;

- Methods based on dictionaries or lexicons, where tokens are compared to extract entities.

These techniques can obtain good results without any training data. On the other hand, they require the development of complex rules or the existence of dictionary collections of entities [Brill and Mooney1997].

Rule-based methods using grammar rules were the first attempts at solving the NER problem [Mikheev et al.1999]. One of the first entity extraction works was published in 1991 [Rau1991], with the goal of extracting company names using a set of manually created rules and heuristics. These used word capitalization or suffix detection for information extraction. Another type of standards-based approach was the use of titles for name recognition. In this case, titles such as "Mrs" or "Miss" followed by one or more capital letters indicate these words are entities of the type Person [Mikheev et al.1999]. The first works for the Portuguese language appeared with the HAREM [Santos and Cardoso2007] but even today there are systems with the same type of approaches such as the PAMPO system [Rocha et al.2016], adopted for the extraction of entities. It first was implemented using a set of regular expressions to gather candidate expressions for entities such as capitalized words and personal titles such as "professor", and later using the results of the POS annotation. This system achieved an f1-score of 73.6% for the HAREM corpus, without classifying entities into categories. However, there are several rule-based systems that can achieve results above 85% [Ralph1995]; [McDonald1993]. The biggest disadvantage of ruled-based approaches is that they require a great deal of experience and grammatical knowledge of the language and the specific domain. They are not adaptable to different languages and domains, and their main problem is maintenance over time. Although rule-based approaches are not ideal, they achieve acceptable results, and they still achieve the best results for very specific, well-defined domains, especially when there is no training data available.

Another hand-coded technique also widely used for entity recognition is the use of dictionaries or word lexicons. These are dependent on the existence of a previous knowledge base in order to extract entities [Gattani et al.2013]. In the literature, this knowledge base is usually called gazetteer [Kazama and Torisawa2008], and its use consists of comparing the words in the text with this gazetteer to find matches. Many of the NER approaches that use a knowledge base resort to Wikipedia, which is a huge knowledge base that provides many entities [Gattani et al.2013]. Some approaches are based on searching each possible entity in the knowledge base in order to find answers in the first words of the description provided by Wikipedia [Toral and Munoz2006]. Other approaches that use specific knowledge bases for the entity category use simple stemming and lemmatization techniques to extract more than just exact match words. The plural, singular, and variants of possible entities [Bellot et al.2003] are also considered. There are also some works for the Portuguese language that follow this approach, one example being the system SIEMÊS [Sarmento2006], which participated in the HAREM event. This system used similarity rules in order to obtain correspondence between the entities, using

the REPENTINO gazetteer. After identifying possible candidates, the system uses the rules to decide on possible categories.

The use of gazetteers or lexicons is a simple approach to the NER task. However, as previously mentioned it is always dependent on the existence of a previous knowledge base for all entity categories. This approach achieves results around 70% for some classes of entities, however, it does not allow the identification of entities that are not in the knowledge base.

2.1.2 Machine Learning Techniques

The NER task has increasingly been a field of study, and many approaches have been proposed. After the initial use of hand-coded techniques, by needed to achieve better results and to have more extensive approaches to the problem, new studies have emerged based on Machine Learning. Supervised learning approaches were first developed by adopting Hidden Markov Models (HMM) [Zhou and Su2002], as well as Conditional Random Fields Models [Finkel et al.2005] to train a sequential entity recognizer, both using previously annotated data. Other methods were also applied, such as experiments based on Maximum Entropy models and the sequential application of Perceptron [Collins2002]. This type of approach could overcome the results achieved with hand-coded techniques, for example, the Stanford NER system with a supervised learning approach could already achieve a f1-score higher than 92.0% in its test set. The most used models based on Machine Learning techniques for the NER task, and those used even today, are probabilistic models. The most used are Hidden Markov Models, Maximum Entropy Models, Conditional Random Fields Models, and some more distinct approaches were based on Decision Trees models, more specifically Random Forest Models. More recently, it has emerged approaches based on Deep Learning, specifically in Recurrent Neural Network (RNN) [Lample et al.2016]. These approaches have been growing fast in recent years, but the most used model and the one that produces better results is the Long Short Term Memory (LSTM) or variants of it. Therefore, we can conclude that the study overview for entity recognition has changed, in this section we try to understand how the new approaches have been used and the results obtained with them.

The **Hidden Markov Models** (HMM) are one of the statistical models most commonly used for NER. In NER tasks the different states represent the name of different entities categories. Each state transition depends only on the current state and represents the probability that the next word belongs to a specific category. For the training of these models, a set of three probabilities are defined: Initial Probability, Transition Probability State and a Conditional Probability of a certain word being in a certain state or belonging to a certain category [Zhou and Su2002]. With this approach for name recognition, many existing works can achieve good results. The state-of-the-art for English with HMM has f1-score results of 93%, for other languages such as Spanish is 90% [Bikel et al.1998]. There are many approaches that apply HMM models, mostly for the English language, using data from MUC-6 and MUC-7. The best results achieved for

this dataset were a f1-score of 96%, based on data dependencies rather than just conditional probability [Todorovic et al.2008]. For the Portuguese language the results are much lower, the same approach trained with the corpus SNR-CLIC in Portuguese [de Freitas et al.2005], has on average f1-score results of 78.07% [Milidiú et al.2007].

Another common approach is the used of **Maximum Entropy Markov Models**(MEMM). This model works in the same way as the HMM, however, they assume dependencies between words, which is the correlation of entities. The MEMM are discriminatory models, which learn the distribution of conditional probabilities. Simple approaches with the MUC-7 corpus and the training of a MEMM with a hundred documents achieve f1-score results of 89.9%, the same model combined with a simple set of rules achieved results of 92.2%. [Fresko et al.2005]. Even using Machine Learning methods, the results are always improved when helped with some hand-coded rules for the specific domain. On the other hand, for the Portuguese language the results obtained with MEMM approaches and the HAREM corpus, with the same classes of entities but with a smaller number of sentences, do not pass beyond 42.8% [Carvalho2012], much lower than those obtained for the English language.

Nevertheless, one of the most used methods for tasks where is necessary to assign a category or class to a term is **Conditional Random Fields** (CRF) models [Lafferty et al.2001]. The CRF's is the most used model for NER [Lin and Wu2009], they are probabilistic mathematical models that work in a very similar way to HMMs, but are not restricted to local resources, it is not necessary to enumerate in advance all the sequences that possible results being defined a single linear distribution in relation to a sequence of categories. This implies that CRF models are able to deal with a larger set of sequences in contrast to generative models. In addition, in HMMs, while probabilities have to satisfy certain constraints in CRF models, there are no associated constraints. CRFs define the conditional probability of a state [Pinto et al.2003], given an input sequence. Equation 2.1, $o = o_1, o_2, ..., o_n$ is a sequence of words of a text of length n, and S a set of states in a finite state machine, associated with a category. Be $s = s_1, s_2, ..., s_n$ a sequence of states that corresponds to the categories assigned to words in the input sequence o. Where Z_o is a normalization factor for all sequences, *fj* is one of the m functions that describe a characteristic, and λ_j is a weight associated with each characteristic function [Teixeira et al.2011].

$$P(s|o) = \frac{1}{Z_o} exp(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(s_i - 1, s_j, o, i)) \tag{2.1}$$

There are many NER works and studies using this approach, for example, Kazama and Torisawa [Kazama and Torisawa2007] developed a NER approach with CRF, also using the Wikipedia information in english as external knowledge, and achieved a f1-score of 92.29% for the CoNLL 2003 corpus, and this result is better than the best achieved by CoNLL participants. Another example of the use of a CRF model for NER was the NERP-CRF system [Amaral and Vieira2013],

in this case, the algorithm was trained using the HAREM Portuguese corpus. This approach uses the Part-of-Speech tagging and word capitalization as features to the model. This CRF approach for the Portuguese language achieved f1-score results of 48.43%, but with a higher precision of 80.77%. One of the advantages of this approach is the possibility of assigning an infinite set of relevant features to be considered by the model. Many other approaches used as features to train de model, the closest words to the entity and also the distinction between upper and lower case letters. In general, compared to all other statistical models, the CRF is the one that produces the best results for the NER task [Agarwal et al.2011]. However, the results of this model, when combined with hand-coded technics, produce even better results than the isolated use of the model [Fernandes2018].

In an attempt to overcome the results with the statistical models, new approaches emerged. The methods based on Neural Networks and Deep Learning techniques are still recent but mostly produce better results [LeCun et al.2015]. The most used methods for the Recognition and Classification of Named Entities at this level are the Long Short Term Memory networks or variants of this model [Chiu and Nichols2016]. The **Long Short Term Memory** (LSTM) is a special type of RNN (Recurrent Neural Network), capable of learning long-term dependencies between word sequences. This type of networks was introduced in 1997 [Hochreiter and Schmidhuber1997], and was improved over time. LSTMs are explicitly networks designed to avoid the problems of long-term dependencies because they store information for long periods. RNNs such as LSTMS are composed of a chain of repetitive neural network modules [Olah2015]. In standard RNNs, this repeating module has a very simple structure, with a single layer (Figure 2.1). LSTMs also have this chain module structure, but the central module has a different structure. Instead of having a single neural network layer, there are four, interacting differently (Figura 2.2).

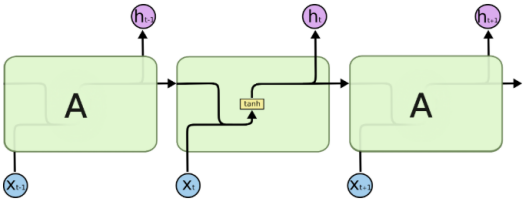

FIGURE 2.1: The repeating module in a standard RNN contains a single layer from Olah, C. (2015)

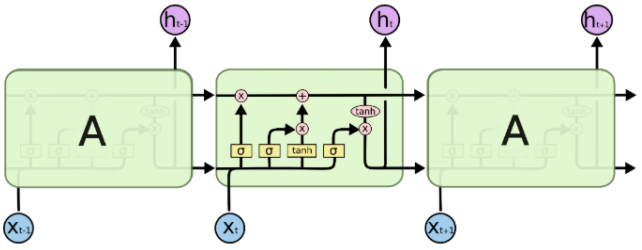

FIGURE 2.2: The repeating module in an LSTM contains four interacting layers from Olah, C. (2015)

The central part, i.e. the key of LSTMs, is the state of the central cell, the horizontal line passing through the top of the diagram. The state of the cell is like a carrier line, this line runs the entire chain, with only a few linear interactions. The LSTMs allowed taking a big step in NER, producing higher results than the existing state of the art [Graves and Schmidhuber2005]. For English, which is the most developed language in the task of NER, the results exceed 92% [Chiu and Nichols2016]. However, the best results are obtained using a second LSTM (Bi-LSTM) that reads the same sequence but two times, backward and forward [Lample et al.2016]. In these cases, the first LSTM is advanced and the second LSTM is delayed, these two distinct networks with different parameters [Peters et al.2018]. For the Portuguese language, the use of LSTM approaches also allowed better results, in the case of HAREM corpus with an LSTM approach, it is possible to see works with f1-score results close to 80% [Castro et al.2018].

2.2 Named Entity Recognition Systems

The number of systems developed in the field of NER has increased considerably in recent years. There are even some systems that implement NER for the Portuguese language, which began to emerge after HAREM [Mota et al.2007]. Each system presents different approaches based on hand-coded techniques, machine learning techniques, and some hybrid approaches. In Table 2.1 are represented some of the existing NER systems for the Portuguese language, the techniques used, the entities covered by each system and the f1-score results obtained. All presented system was evaluated with de HAREM Golden Collection Corpus.

System Name	**Hand-coded Techniques**	**ML Tecnhiques**	**Entities Categories**	**Results**
CaGE	Dictionaries, Gazetteer		Place, Person, Time, Organization	65.72%
PorTexTO	Expression patterns using co-occurrences		Time	62.31
Priberam	Lexicals, morphosyntactic and semantic classification, ontologies		Abstraction, Event, Local, Thing, Work, Organization, Person, Time and Value	64.1%
R3M_1	Rule-based to identify NER candidates	Co-training algorithm to infer classification rules	Person, Local, Organization	79.47%
REMMA	Lists of words and Wikipedia as a knowledge base		Person, Local, Organization, Event, Time, Value	46.81%
XIP	Lexicons, Lists of words, syntactic and semantic processing		Person, Local, Organization, Event, Time, Value	54.45%
Mamede et. al		CRF model	Person, Local, Organization	76.8%

Table 2.1: State-of-the-Art comparison for Portuguese NER Systems

In the table are represented seven different systems, the approaches used to solve the NER problem for the Portuguese language and the results achieved. In these systems, we can see a large variety of results depending on the approach used and the number of entities considered. The system with the lowest results having an f1-score of 46.81% and the best an f1-score of 76.8%, but with a hybrid approach and considering half entities number. All the results in the table follow the HAREM guidelines and correspond to the value of the f1-score achieved by each system. By analyzing the systems and observing the table, we easily notice the use of hand-coded techniques in most systems. R3M_1 [Mota2008] being the only with a hybrid approach and the system of Mamede et al. [Mamede et al.2016] the only one that has an exclusively applied approach to machine learning techniques.

Through the individual analysis of each system we realized that globally, the Priberam system [Amaral et al.2008] is the one that achieves the best results in the identification and classification of entities tasks. Having reached the highest result of f1-score in both, and is also the system that treats the largest number of named entities. This system was also the one that got the best score in the second HAREM event. A more realistic comparison is to consider the Priberam, XIP [Hagège et al.2008] and REMMA [Ferreira et al.2008] systems, whose common attempt to identify and classify a large number of entities are ambitious. Although in this scenario REMMA has the most higher accuracy of the three in the identification task, it has the lowest overall result. The difference in precision compared to other systems is not matched with the recall results achieved. Typically, a low recall value can be explained for two reasons: the system was built to deal with a small set of entities, or some parts of the system are underdeveloped, for example for lack of resources. On the other hand, the XIP is the best system for classifying temporal expressions, mainly because a great effort has been made to improve the results of this specific category. On the other hand, the Value category proves to be one of the weaknesses of XIP, because it has an f1-score was 42%, which is very low for a category where the other systems get all good results. However, of all the systems presented the best results were obtained we have systems R3M_1[Mota2008] and Mamede et. al [Mamede et al.2016], but for the smaller number than entities. The best results achieved for both systems were for the entities Person and Local, and the entity Organization always had much lower results in all tests.

With this evaluation of the different systems, we could see that although the work at the level of other languages is more developed it is possible to implement systems for the Portuguese language. It is also possible to see that different approaches can produce good results, but as we have seen before, the best results are achieved most of the time by hybrid approaches that combine different techniques.

Chapter 3

DataSense Framework and Concepts

In this chapter, we present a general description of the DataSense Framework, and the Named Entity Recognition Component developed with the work and results achieved from this dissertation. We also describe the necessary concepts to properly perceive the topics involves in this dissertation. The initial tasks for this work were based on the technical definition of the DataSense Framework and the associated modules.

In the context of today's society, totally focused on information and communication of data generated from countless sources, multiple documents are generated with different origins and purposes. It is normal that a large part of these documents contains confidential and/or sensitive information. It is also natural that, over time, organizations archive a vast number of documents, leading to a loss of control over their content.

In recent years, the creation, processing, and analysis of large volumes of data have become a practice in organizations in order to exploit this information for commercial purposes. The problem is that with the creation and storage of current data, new challenges have arisen. One of these being the management of sensitive or personal data, which can range from simple address or card number to biometric and medical records stored in organizations' archives. This management of sensitive data increasingly became a priority for society. Foremost due to European legal requirements, but also due to the imposition of the general population, which wants to know the way in which data is treated and processed. This is represented in the new sensitive data standard of GDPR (General Data Protection Regulation)[1] which became mandatory for European organizations in 2018.

This chapter gives a basic overview of the DataSense Framework and its architecture, as well as a basic explanation of the concepts of sensitive data, personal data and the GDPR, according to the context of this dissertation.

[1] `https://gdpr-info.eu/`

3.1 Sensitive and Personal Data

Sensitive Data and Personal Data in the context of this dissertation are two separate concepts. Personal data is any information that identifies an individual, i.e., personal data is all data that contains:

- Direct identification information, such as first name, last name, phone numbers, etc;
- Anonymous or non-directly identifiable information, which does not allow the direct identification of users, but allows the identification of individual behaviors;

Personal data can be broken into two categories: sensitive and non-sensitive. Sensitive data is all data that directly identifies an individual; Non-sensitive data is data that gives us information about the individual but through which it is not possible to reach the person. In the context of this work, we process all personal data covered by the GDPR, thus both sensitive and non-sensitive data are treated.

3.2 General Data Protection Regulation (GDPR)

Recently, the European Commission has focused efforts on the topic of data, with the aim of restricting their storage and use. Since May 2018, the General Data Protection Regulation (GDPR) has been mandatory. The data that is considered sensitive and subject to conditions of specific treatment are:

- Personal data revealing racial or ethnic origin, political opinions, and religious beliefs or philosophical
- Trade Union Affiliation
- Genetic data, biometric data processed simply to identify a human being
- Health-related data
- Data concerning a person's sexual life or sexual orientation

The GDPR aims to modernize the European Union's data manipulation legal system, strengthen individual rights and improve laws in terms of clarity and consistency.

3.3 DataSense Framework

As explained in chapter 1, the work developed in this dissertation is part of the DataSense project [Mariana Dias2019a], this is a project that aims to create a framework. This framework is being developed to help organizations deal with the new rules established by the European Union. In addition to this, it allows organizations to have a greater knowledge and management of their data and documental database.

DataSense is a highly exportable framework that enables organizations to identify and understand the sensitive data in their digital documents (unstructured textual information), in order to comply with legal and security purposes. The DataSense Framework is based on three key layers that make use of the current potential of Natural Language Processing technologies and advances in machine learning, Named Entity Recognition, Coreference Resolution, and Automatic Learning.

With the definitions presented above, it was possible to list the types of sensitive data for the DataSense Project. To this list is has also been added some types of data covered by the concept of Personally Identifiable Information (PII), which governs the use of this type of data in the United States of America[2]. Table 3.1presents all considered sensitive data, and we can notice that it was divided into three main categories: Personal Identification Numbers, Socio-economic Information and Others.

[2] `https://www.investopedia.com/terms/p/personally-identifiable-information-pii.asp`

Categories of sensitive data	Types of sensitive data
Personal Identification Number	- National identification number - Bank Identification Number - Credit Card Number - Tax Identification Number - Passport Number - Social Security Number - National Health Number - Telephone Number - Driving License Number
Socio-Economic Information	- Names - Address and Locals - Postal Code - Household Information - Date and Place of Birth - Contract/Ordered Values - Medical Data - Profession and Employer Entity - Political Guidelines - Trade Union Affiliation - Religion - Associations
Others	- E-mail address - Sexual Orientation - Mechanographic Numbers - Criminal Record

TABLE 3.1: Categories and types of sensitive data used in DataSense Framework

In addition to the recognition of personal data, the DataSense Framework should also classify entities as a way to understand and control the data. The final product must be a functional prototype that assists organizations in handling their data and complying with the imposed standards.

For the DataSense Framework, a list of requirements has been agreed on initially. Tables 3.2 and 3.3 show the functional and non-functional requirements of the framework that relate to the work done for this dissertation, Named Entity Recognition Component.

Functional Requirements	Description
F1	Ability to analyze unstructured text information in multi-format humanly readable
F2	All information provided to the NLP Engine Module must be previously treated and presented in text format
F3	The NLP Engine Module shall be independent of the other components and shall provide a REST API for communication between the components.
F4	The NLP Engine Module must provide an endpoint to invoke remote execution (POST)
F6	The system is composed of a module of Recognition and Classification of Named Entities to carry out these tasks independent of the other functionalities of the system
F7	The NER Component should have the ability to automate the recognition and extraction of sensitive information in text documents
F8	The NER Component shall have the capacity to automate the categorisation of entities identified as sensitive information in text documents
F9	The assignment of categories for the classification of sensitive data will take place according to a set of predefined categories and types
F10	The system should have the ability to answer questions about the content of the data analysed and allow the request of information from the data managed
F11	The system must have a preprocessing text pipeline in order to normalize the data that serves as input to the templates
F12	In addition to the Machine Learning models, this component must also have a rules-based sub-component capable of detecting regular sensitive data
F13	Due to the similarities between numerical data, the NER Component should use disambiguation mechanisms

Table 3.2: NER Component Funtional Requirements

Non-Functional Requirements	Description
NF1	The system to be developed in a programming language advantageous to data processing, for example, Python
NF2	The project's text processing should be based on best practices and open-source tools
NF3	The system should have a modular architecture, consisting of several autonomous and independent modules
NF4	Datasense should choose for a standard representation of the data classification, the most appropriate being the CONLL column format with the IOB notation
NF5	The DataSense system shall communicate with the modules via a REST API using HTTPS communications (via SSL)

Table 3.3: NER Component Non-Funtional Requirements

In order to develop a framework with a structure capable of achieving the objectives and requirements, DataSense was proposed as a modular project. The framework must be able to process data in real-time and respond to user requests.

In Figure 3.1, the functional architecture of the framework is represented. Its logical components can be observed, which represent the blocks responsible for DataSense functionalities and

how they communicate with each other. The framework consists of a set of five independent modules, and this dissertation focuses only on the NLP Engine module and its Named Entity Recognition Component. This component and its development are detailed in the next chapters of this document.

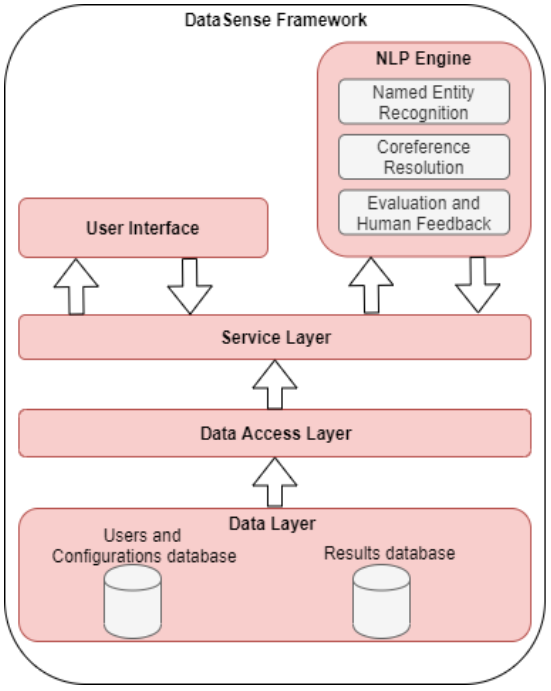

FIGURE 3.1: DataSense Framework Functional Architecture

It is possible to see that the NLP Engine Module does not communicate directly with the user, there is a Service Layer that makes an initial treatment of the data before it is given as input to the module. The Named Entity Recognition Component (NER) communicates directly with the Coreference Resolution and Evaluation and Human Feedback components. The output from the NER component is necessary for the operation of all other components, being the most relevant component in the whole operation of the framework.

All modules have different objectives and specific characteristics, and can be described as:

- **User Interface :** The interface is the visual component of DataSense that allows users to interact with the tool.

- **Service Layer:** The Service Layer is responsible for communication, providing a REST API for these communications. There are 4 main sub-components that make up this layer: **API for users**, **API for settings**, **API for executions** and **API for results**.

- **NLP Engine:** The NLP Engine is composed of three sub-components. Is responsible for do all of Natural Language Processing and Machine Learning work of the Framework. It is in this component that should be made the discovery and classification of the data, where the

relationships between the data are found, and there are models evaluated, and the user feedback is considered.

- **Users and Configuration Database:** The User and Configuration Database is responsible for storing user accounts and their settings.

- **Results Database:** The Results Database is responsible for storing the results from the application of the NLP Engine methods.

Figure 3.2 presents the technical architecture of the project. Since this dissertation focuses on the NLP Engine Module, we can see that some of the technologies are: Python, Flair, AllenNLP, SpaCy, and Flask. There are some other technologies that were used, specifically in the NER Component that are not shown in the image, but that will be described in the following chapters.

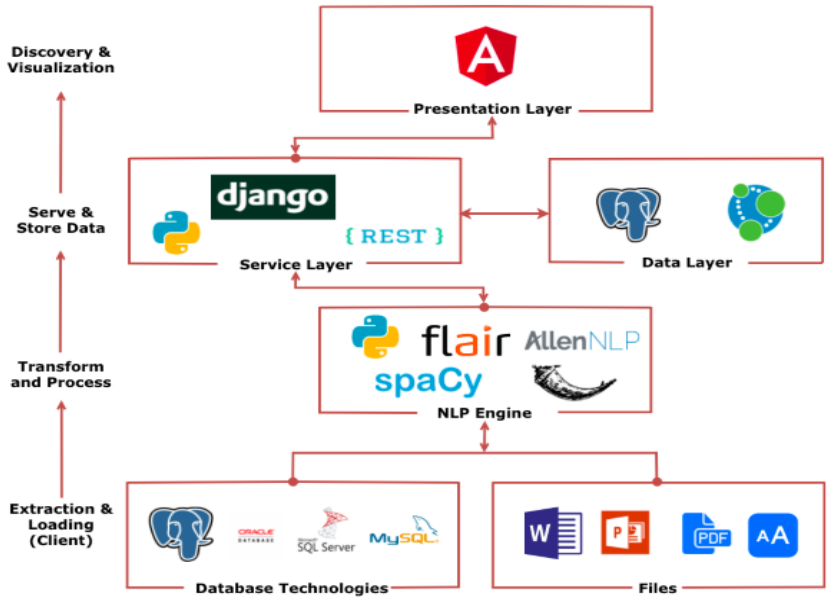

FIGURE 3.2: DataSense Framework Technical Architecture

Given the context of this dissertation in the DataSense Project, from this point on, the next chapters refer exclusively to the Named Entity Recognition Component.

Chapter 4

Named Entity Recognition

In this chapter, we present the procedures, techniques, and tools used for the development of the component of Named Entity Recognition (NER), as well as a detailed explanation of the development steps used to achieve the objective. As the main goal is the recognition of sensitive Portuguese data and processing of non-structured text documents, we divide the problem into three main sub-problems: (i) preprocessing of the text, (ii) recognition and classification of named entities, and (iii) post-processing. This chapter presents the proposed solution and architecture for the NER component of the DataSense project that is scalable and allows the identification of sensitive data.

4.1 Overview

In order to not have a closed development, a modular architecture was adopted for the development of the NER Component, following a specific processing chain (Figure 4.1). This architecture allows the different modules that belong to the chain to be configurable and instantiated several times independently. Furthermore, this architecture allows in the future that other modules can be added without additional cost.

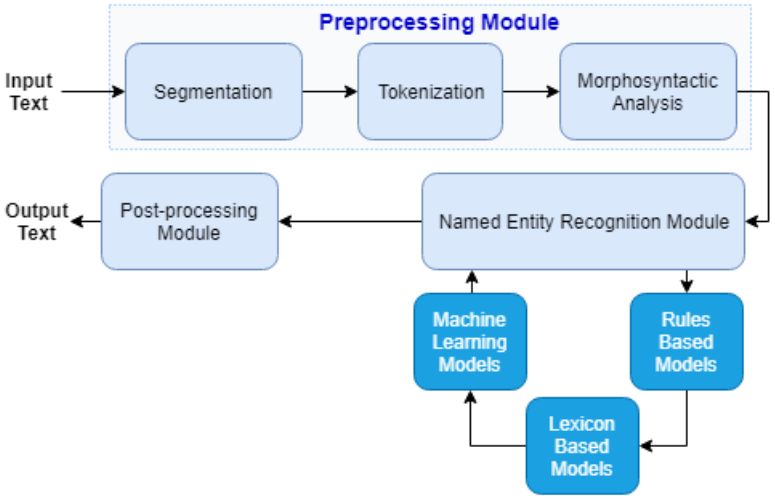

FIGURE 4.1: Named Entity Recognition Component chain

Figure 4.1 represents the architecture of the NER Component, and the processing sequence carried out, which always consists of the presentation of the figure. The input of the component consists of text only in the Portuguese language that has been already been previously treated at the level of images and tables that could be present in a document. Tools such as OCR (Optical Character Recognition) are used to extract text from PDF documents, among others. All text processing and extraction are independent of this module and the work for this dissertation. This type of tasks is performed by the other components of the system as we saw in chapter 3 and Figure 3.1.

Given the input text, the main goal is the Recognition and Classification of sensitive data, taking into account the group of classes of sensitive and personal data (Personal Identification Number, Socioeconomic Information and Others). This is done by following the processing chain and all the techniques presented above. The three main steps presented: Preprocessing, Named Entity Recognition and Post-processing consist of the tasks that are divided into modules. Each one of these modules is concerned in solving a specific problem to reach the output text:

Preprocessing: Performs the necessary input text processing in order to feed it into the next module, NER module. It is divided into three distinct tasks;

Named Entity Recognition: This is the core module of the component where Recognition and Classification of sensitive data are performed. This module is composed of a pipeline of three other independent components;

Post-processing: The post-processing module has the function of formatting the results obtained by the previous modules according to what is intended by the user;

4.2 Preprocessing Module

The Preprocessing Module is the first element in the chain. It is responsible for preprocessing and treating the input data, performing a set of preprocessing tasks so that the text can serve as input to the next module.

Preprocessing is one of the most important tasks of all Natural Language Processing (NLP) and Information Retrieval (IR) studies [Kannan and Gurusamy2014]. Preprocessing text means giving to the text another format so that it can be analyzed, something that an algorithm can digest. Text preprocessing can be divided into four fundamental parts: Cleaning, Annotation, Normalization, and Analysis. Depending on the problem and the intended solution, not all four parts have to be applied. Preprocessing tasks consist of all the tasks performed in preparing text, sentences, or words before the application of the models [Vijayarani et al.2015]. In this chapter, the tasks and methods of preprocessing used to carry out this dissertation are presented.

As we saw in Figure 4.1, the preprocessing module is divided into three parts that are always executed in the same order one after the other.

Segmentation: So that each sentence is processed individually without depending on the context of the previous sentence. We start by dividing the entire text into sentences by the end of sentence punctuation marks: period (.), question mark (?), exclamation(!) and suspension points (...). This configuration can be parameterized, and this division into sentences may not be performed, but since the objective is to process large text documents, and the noise-induced to the NER models was not advantageous, we kept the chain with the segmentation in sentences. In Listing 4.1, it is possible to see an example of the Segmentation in the Portuguese language.

```
IN: [Revisão das Directrizes de Oslo.  As Directrizes de Oslo são um documento das  Nações
    Unidas, destinado a orientar o uso de meios militares e da defesa civil nos cenários de
    emergências naturais, tecnológicas ou ambientais em tempo de paz.]
OUT: ['Revisão das Directrizes de Oslo.', 'As Directrizes de Oslo são um documento das  Nações
    Unidas, destinado a orientar o uso de meios militares e da defesa civil nos cenários de
    emergências naturais, tecnológicas ou ambientais em tempo  de paz.']
```

Listing 4.1: Text Segmentation Example

Tokenization: Tokenization is performed second on the module chain. The tokenization component does the division of the text in n-grams, words or sets of words. The number of n-grams can also be parameterized, and this tokenization consists of representing the text as a vector of individual or sets of words. In this tokenization, some decisions have been made in terms of punctuation, which consists of separating all the non-alphanumeric characters from the words. All punctuation marks except the hyphen (-), the at (@) and the slash (/) are separated by a blank space from all alphanumeric characters in the text. By default, the parameterization used in the processing chain consists of the division into unigrams as in Listing 4.2:

```
IN: ‘(Nome:Ana Cardoso Data de Nascimento:14/06/1983, Telemóvel’:916415055)
OUT: ([Nome],[:],[Ana],[Cardoso],[Data],[de],[Nascimento],[:],
     [14/06/1983],[,],[Telemóvel],[:],[916415055])
```

Listing 4.2: Text Tokenization Example

Morphosyntactic Analysis: After the tokenization, we perform the morphosyntactic analysis of all separate text in unigrams. The text is analyzed and classified with Part-of-Speech Tagging using different techniques and tools. The task of Part-of-Speech (POS) tagging consists of analyzing and tagging all the words in the text at the syntactic level, as we can see in Listing 4.3:

```
IN: ([Nome],[:],[Ana],[Cardoso],[Data],[de],[Nascimento],[:], [14/06/1983],[,],[Telemóvel
    ],[:],[916415055])
OUT: Nome NOUN
     : PUNCT
     Ana PROPN
     Cardoso PROPN
     Data NOUN
     de ADP
     Nascimento NOUN
     : PUNCT
     14/06/1983 NUM
     , PUNCT
     Telemóvel NOUN
     : PUNCT
     916415055 NUM
```

Listing 4.3: Text Morphosyntactic Analysis Example

The output result of the Preprocessing Module is the final result obtained by the morphosyntactic analysis, and this result serves as an input to the next module (NER Module) and facilitates the recognition and classification of the named entities. For the morphosyntactic analysis, some distinct approaches were tested in order to achieve the best result of Part-of-Speech Tagging for the Portuguese language. Section 4.2.1. of this document details all the models and tools used to research the best POS tagging solution for Portuguese.

4.2.1 Part-of-Speech Tagging Implementations

POS tagging is a key feature in most NLP tasks, especially in the tasks of named entity recognition [Shishtla et al.2008], and therefore a very important task in this work. The goal is to assigns each word its category, i.e. Part-of-Speech tagging is the process of assigning a lexical marker to each word in the text. That is, to classify words into categories based on their role in the

text, the context in which they are inserted and their neighbourhood [Branco and Silva2004]. The number of classes or tags may depend on the dataset, model or algorithm applied, but there are a set of classes that are always present and that are the most important. These classes are those we can see in Table 4.1, that is a subset of the POS Tags used on NLTK library and firstly presented on the Penn Treebank Project [Santorini1990]. The table represents the set of most relevant classes to this work and those that were used in the NER model. In the table we see in the first column the POS acronyms. The acronyms are not consistent in all libraries of NLP, for example, using the SpaCy library the class 'Proper Noun' is represented as 'PROPN' and in NLTK as 'NNP'. In the second and third columns of the table the acronyms description and an example of possible words for each tag are presented.

POS Tag	Description	Example
ADJ	adjective	small (pequeno)
ADP	adposition	with (com)
ADV	adverb	there (ali)
AUX	auxiliary	is (é)
CONJ	conjunction	and (e)
DET	determiner	the (o)
INTJ	interjection	Hello! (Olá!)
NOUN	noun	boy (rapaz)
NUM	numeral	seventeen (dezassete)
PRON	pronoun	he (ele)
PROPN	proper noun	Lisbon (Lisboa)
PUNCT	punctuation	!
SCONJ	subordinating conjunction	if (se)
SYM	symbol	$
VERB	verb	study (estudar)
X	other	ahahah
SPACE	space	

Table 4.1: Part-of-Speech tags, description, and examples

The most important tags in the context of this work are **Noun** and **Proper Noun**, generally referring to people and places which is the personal data we want to identify.

The results associated with the NER task are highly dependent on POS tagging [Ritter et al.2011] and there are different ways to perform this task. In order not to compromise the following tasks one of the main goals is to achieve the best results in the POS Tagging task. After the study of the state-of-the-art, three different implementations were tested and analyzed to select the best to integrate into the NER Component. In this section, we explain the three implementations and the results obtained can be seen in chapter 6 of this document. In the first and second experiments, we use the POS tagging model of the NLTK library, the first being applied directly

to the second one retraining the model, and in the third experiment, we used the SpaCy library POS model.

1st Experiment: For this experiment, we tested the Part-of-Speech tagging model of the NLTK (Natural Language Toolkit) library [Loper and Bird2002].

In Listing 4.4, we can see the result of a contract excerpt processed by this model. In the result, we see that the words 'contrato' and 'milhões' were labelled with the tags 'NN' which mean noun, singular. The word 'dólares' has the tag 'NNS' which also corresponds to a noun but in the plural. The words 'Reino' and 'Unido' with the tag 'NNP' which corresponds to a proper noun. Another example is the tag 'VBZ' which is associated with the word 'fechado', that corresponds to a verb on 3rd person singular.

```
O NNP
contrato NN
foi NN
fechado VBZ
no DT
Reino NNP
Unido NNP
por IN
100 CD
milhões NN
de FW
dólares NNS
```

Listing 4.4: 1st Experience Part-of-Speech Tagging Example

The Part-of-Speech Tagging model provided by the NLTK library uses an averaged perceptron tagger. This POS classification model operates using a variety of algorithms, including Decision-Tree models, naive Bayesian models, the Mallet and Weka [Malecha and Smith2010] machine learning package. Although we can see that the morphological analysis performed in the example sentence is not completely wrong, unfortunately, NLTK does not really support multi-lingual tagging apart from English and Russian. We can, however, get our own POS-tagger by training on a foreign language corpus, which is what is shown in the next experiment.

2nd Experiment: This experiment consists of the model provided by NLTK, re-trained with the Floresta Sintáctica Corpus. The Floresta Sintatica Corpus is provided by Linguateca[1] and is a set of morphosyntactically annotated sentences [Freitas et al.2008]. It has four parts, which differ in terms of textual gender, mode (written vs. spoken) and degree of linguistic revision: the **Bosque**, which has been completely revised by linguists; the partially revised **Selva**, the **Floresta Virgem** and the **Amazónia**, which have not been revised. For this experiment, only the first and second parts were used, the Bosque and the Selva. These corpora are composed of

[1] https://www.linguateca.pt/Floresta/

24189 sentences and 490 thousand words. The Bosque is the smallest corpus with 9368 sentences and 190,000 words, the Selva is composed of 300 thousand sentences and 14821 words. Both corpora contain 22 different classes in which the tag 'n' that corresponds to the nouns is the one with the highest number of occurrences. In Figure 4.2 is possible to see the distribution of the classes in the corpora.

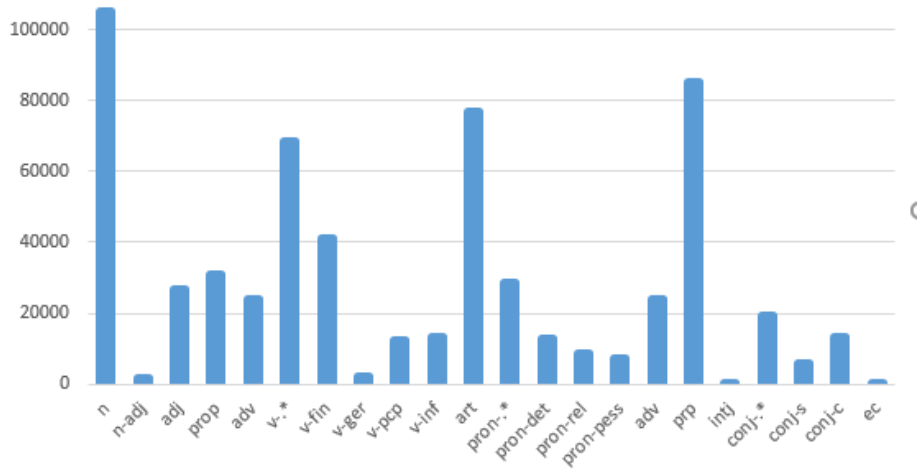

FIGURE 4.2: Part-of-Speech Tagging classes distribution on Bosque and Selva corpora

The first task to retrain the model was the transformation of the corpus so that the tags would correspond to those of the NLTK model represented in table 4.1. The conversion used for the tags can be seen in the Appendix section (Table 7.1).

After having the corpus with the generic tags, the NLTK model was retrained this time with the data in Portuguese. For the retraining of the Part-of-Speech classification model, a set of six different models provided by the NLTK library was used in the chain. The chain of models is represented in Figure 4.3.

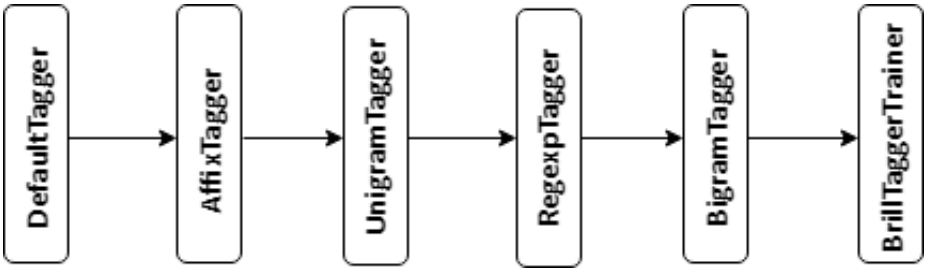

FIGURE 4.3: NLTK retraining model implementation chain

From the six models used, each one has its own features and a distinct purpose in the processing chain:

DefaultTagger: The first model is only intended to classify all words with the NOUN tag. In case none of the other models can solve the word, the default value will always be the NOUN.

AffixTagger: This model classifies each word individually according to the prefix/suffix. AffixTagger chooses a token tag based on an initial or final subset of its word string.

UnigramTagger: The third classifies the word according to a word dictionary or the most common classification. In this implementation, UnigramTagger finds the most likely tag for each word in the training corpus and then uses this information to assign tags to new tokens.

RegexpTagger: This is used to correct some special cases in which the other Taggers usually make mistakes. RegexpTagger tags the tokens by comparing their word strings to a series of regular expressions. In this implementation, a set of regex was defined mainly for the tag "ADV" (adverb), in specific for the words: "no"(não), "yes"(sim), "by"(pelo), "this"(neste), "this"(nesse), "that"(naquele) and "always"(sempre), as well as for all variations, male, female, plural and singular of these words.

BigramTagger: Classifies the word using the context of the previous word to delete ambiguities (e.g. common words that can be both a verb and a noun). The tagger considers only the preceding words in the same sentence to choose the tag for the token.

BrillTaggerTrainer: The final model is trained to recognize patterns where other taggers make mistakes and create replacement rules. It is the most time-consuming and memory-consuming step but produces the greatest gains.

These models were executed in a chain with the Floresta Sintáctica corpus to produce a model for the Portuguese language. This model was implemented for sentence processing, using NLTK tokenization. Therefore, giving as input to the model a sentence the result consists of a set of words and their associated tags. In the Listing 4.5, it is possible to see some improvements from the first experiment. The words 'de', 'por' and 'no' have already been correctly tagged with the tag 'ADP'. We also see that the first word 'O' has already been tagged with the tag 'DET' and 'foi' as a verb with the tag 'VERB'.

```
O DET
contrato NOUN
foi VERB
fechado VERB
no ADP
Reino NOUN
Unido NOUN
por ADP
100 NUM
milhões NOUN
de ADP
dólares NOUN
```

Listing 4.5: 2nd Experience Part-of-Speech Tagging Example

3rd Experiment: This third experiment is similar to the first because it is a direct application of a model, in this case provided by the SpaCy library. The decision to do this was due to the fact that SpaCy has made available a POS tagging model for the Portuguese language. SpaCy's POS tagging model is a statistical model that has been developed to provide a high-performance mix of speed and accuracy [Olney et al.2016]. This POS tagging follows the Universal Dependencies scheme[2], where the universal tags do not encode any morphological features and only cover the word type. In addition to the simple statistical models, SpaCy also uses CNN multi-task training with data available on the Universal Dependencies and WikiNER corpus [Colic and Rinaldi2019]. The provided model receives as input sentences and returns tokens with the respective POS tag. An example can be seen in Listing 4.6.

```
O DET
contrato NOUN
foi AUX
fechado VERB
no ADP
Reino PROPN
Unido PROPN
por ADP
100 NUM
milhões SYM
de ADP
dólares NOUN
```

Listing 4.6: 3rd Experience Part-of-Speech Tagging Example

We can see that there are already some changes. In this experiment, the word 'foi' was classified as 'AUX' which assists of the verb 'fechado'. The words 'Reino' and 'Unido' in this experiment were classified as 'PROPN'. One of the most visible changes is also the tag 'SYM' which corresponds to the symbol assigned to the word 'milhões'.

4.3 Named Entity Recognition Module

As we saw in Figure 4.1, the Named Entity Recognition Module is the second module in the NER component chain [Mariana Dias2019b]. The input of this module consists of the output of the Preprocessing Module, like the one we saw in Listing 4.3. The output of this module must be the input text annotated in the CoNLL format. The CoNLL format [Ramshaw and Marcus1999] is one of the most commonly used annotation methods in NER tasks, also known as IOB tagging (or BIO). The three initials BIO mean Begin, Inside and Outside. Words tagged with O are outside of named entities, and the I-XXX tag is used for words inside a named entity of type

[2] https://universaldependencies.org/u/pos/

XXX. Whenever to prevent two entities of type XXX are immediately next to each other, the first word of an entity will always be tagged with B-XXX. Listing 4.7 shows a possible result produced by the NER module.

```
A O
Maria  B-PER
vai  O
para   O
a  O
Serra  B-LOC
da  I-LOC
Estrela  I-LOC
```

Listing 4.7: IOB tagging and NER Module output example

It is in the NER module that the models and systems for recognition of sensitive data are implemented. The result produced by the module is the text classified with the respective classes of sensitive data.

Based on the sensitive data defined for the DataSense project, in Table 3.1, the classes of entities to recognize in this module were defined. In this module, we use the same division into categories: Personal Identification Number, Socio-Economic Information and Others. Table 4.2 shows the set of classes of entities for this dissertation.

Categories	Entities Classes	Sensitive Data Included
Personal Identification Number	NumIdentificacaoCivil	Personal Identification Number
	IdentificacaoBancaria	Bank Identification Number, NIB and IBAN
	NumCartaoDeCredito	Credit Card Number and American Express
	NumIdentificacaoFiscal	Tax Identification Number
	NumPassaport	Passport Number
	NumSegSocial	Social Security Number
	NumUtenteDeSaude	National Health Number
	ContactoTelefonico	Telephone Number
	NumCartaConducao	Driving License Number
Socio-Economic Information	Pessoa	Person Names
	Local	Addresses, Locals, Place of birth
	Organizacao	Organizations, Affiliations, Employer Entity
	Tempo	Dates, Dates of Birth, Contractual dates
	Valor	Values, Ordered values
	Med	Medical data
	Profissao	Jobs, Professions
Other	CodigoPostal	Postal Code
	EnderecoEletronico	E-mail address

Table 4.2: Classes of entities considered in this work, based on GDPR and DataSense Project

For this work, we identified eighteen different classes of sensitive data. The recognition of named entities is based on three different sub-modules: Ruled Based Models, Lexicon Based Models, and Machine Learning Models.

4.3.1 Ruled Based Models

As we saw in chapter 2, several Information Extraction, and Named Entity Recognition approaches are based on rules. This first component of the NER Module implements different rule-based models to discover some of the entity classes. The entity classes that are discovered at this stage of the component chain are all associated with sensitive data related to the Personal Identification Numbers category, including also postal codes, email addresses, and some date formats. The choice of rules for the discovery of this type of data is due to the fact that they are data types that have well-defined formats and follow a pattern. In addition to the rules in some cases, there is an extra validation. This validation is performed on all personal numbers for which there is a control validation, check digit or checksum. With this validation, it is possible to disambiguate and have a greater certainty of cases such as the telephone number and the tax identification number, both containing 9 digits. Another feature of this implementation with the rules-based model is the context. The context has been added to the model in order to solve errors in some of the data types, mainly those of the Personal Identification Number category. The context consists of a specific word or set of words for each class of entity that must exist in the text in order to confirm the result achieved with the rules. The use of context is parameterizable and is only used in specific document types that have more textual information, and it consists of confirming if any of the words present in the context list are in an offset of five tokens (by default) before or after the found entity.

Figure 4.4 represents the flowchart of the ruled-based model. The model processes token to token, although it receives all text divided by tokens the processing is done in uni-grams. On the flowchart is possible to see two initial confirmations: **"Otag?"** and **"IsDiggit?"**. The first consists of verifying if an entity class has already been assigned to the token and if there is already a class assigned previously this should not be changed. The second confirmation exists because all classes except the E-mail address class contain numbers, and it is unnecessary to go through all other confirmations if this is not fulfilled.

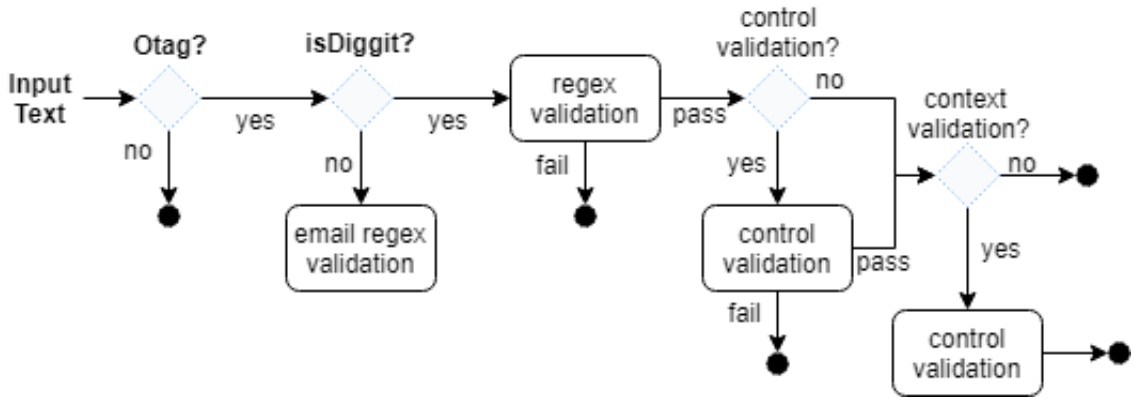

FIGURE 4.4: Ruled Based Model Flowchart

The following explanation order consists of the same order in which the set of rules was implemented.

- NumIdentificacaoFiscal: This first entity class is the sensitive class related to the tax identification number. The verification with rules for this number is quite simple, are accepted:

- Nine digit numbers
- 'PT' followed by a nine-digit numbers

At the first rule, any nine-digit number is accepted as long as it does not start with 0 (zero). In the case that the beginning 'PT' appears, the validation of the digit is the same. Since a nine-digit number can also correspond to a telephone number, in this case, it is necessary to perform a control validation. The tax numbers have a checksum, and so it is possible to confirm if they correspond to this entity. The control digit is always the last one and is calculated through the following steps:

- Multiplying the 8^{th} digit by 2, the 7^{th} digit by 3, the 6^{th} digit by 4, and so on
- Adding up the result of the multiplication
- Calculating the rest of the number division by 11
- If the rest is 0(zero) or 1(one) the check digit will be 0(zero)
- If it is another digit X, the check digit is the result of subtraction 11-X

In addition to the control validation, the context list of words was also used. The words added in the context list for the entity NumIdentificacaoFiscal were: **nif**, **fiscal**, **Contribuinte**, **Identificação Fiscal**. However, this entity may appear written in different ways, in which case the accepted formats are as follows:

- Digits all together - **226154947** or **PT226154947**
- Digits separated by space - **226 154 947** or **PT 226 154 947**

- Digits separated by dots - **226.154.947** or **PT 226.154.947**

- ContactoTelefonico: This entity represents the sensitive data telephone number. The set of rules was used and also a set of context words was created. This entity does not have any control validation, and it is only implemented for Portuguese telephone numbers or and similar. The rules used for the extraction of this entity were:

- Numbers with nine digits started by 2, 9, 7 or 8
- Numbers with fewer digits started by 1
- Numbers preceded by a Portuguese code

In the first rule are accepted all numbers beginning by 2(two) and not preceded by 0(zero) which are composed of nine digits. Also numbers beginning by 9(nine) but only if preceded by 1(one), 2(two), 3(three) or 6(six), with the following structure: 9[1236][XXXXXXX]. Finally, any number with the format: [78][XXXXXXXX], which are the Portuguese service numbers. In the second rule are accepted all numbers beginning with 1(one) that are composed of four digits, except 112(one hundred and twelve), which is the Portuguese emergency number. Codes are only accepted when combined with the other nine digits, where the accepted codes formats are: '+351', '351' and '0351'.

The list of context words for this entity is: **contacto**, **contato**, **telemovel**, **telefonico**, **telefónico**, **telefone**, **contactar**, **fax**. It consists of a set of Portuguese words that usually appear in documents related to telephone contacts. This entity may appear written in a document in different ways, and the formats that have been treated in this work are:

- Digits all together - **968702375**
- Digits separated by space - **96 8702375** or **968 702 375**
- Digits separated by hyphen - **968-702-375**
- Code together, separately or in parentheses - **+351968702375**, **+351 968702375** or **(+351) 968702375**

- NumIdentificationCivil: This entity class represents the personal identification number. This is a type of data that usually appears in documents in full or with only the first numbers. For this reason, the regex was implemented in order to accept:

- Numbers with 7 or 8 digits
- Sets of 12 character

The first possibility is a set of 7(seven) or 8(eight) digits, without any verification. The second possible way is a set of 12(twelve) characters that must comply with the following format: [XXXXXXXX][checkDigit][AA][DigitControl], where the first part X's correspond to 8 (eight) consecutive numbers, followed by a check digit that can vary between 0 (zero) and 9 (nine),

the A is two alphanumeric characters that can vary from [A...Z, 0...9], and the last digit that is used as a control digit. In this case, when the number is complete, validation is performed by confirming the check digit and the control digit. This validation is done as follows:

- By counting from right to left, double the value of each 2nd element, and the letters should be replaced according to a conversion table
- If the duplicate result is equal or greater than 10, subtract 9 from its value
- Add up all the values obtained
- Calculate the rest of the division by 10
- If the value is 0 the number is valid

Through the above steps, it is possible to validate the number found and confirm if it belongs to the entity class or not. However, there is also the context that is most useful in the first case in which the reduced version of the number appears. The set of words are: **Civil**, **Cidadão**, **cc**, **bi**, **Cidadao**, **numero do cartao de cidadão**.

As in all other cases we have already seen, the personal identification numbers can appear in different ways. In the case of this entity, the reduced version of the number is only accepted if the eight digits are together. For the other case, the following formats are accepted:

- Complete number with all digits together - **150697410ZX8**
- Number separated by spaces - **15069741 0 ZX 8**

- NumPassaport: This class corresponds to the Passport Number. This type of sensitive data has many possibilities for the different countries, so in this work, we only implemented the extraction of Portuguese passport numbers. This rule consists only of a non-case sensitive letter followed by 6(six) digits. The initial letter can correspond to any letter of the Portuguese alphabet of A-Z and also does not exist in a rule for the next six digits. For this sensitive data, there is no type of checksum validation. It is in these cases that the context associated with the numbers is very important and allows a greater certainty of the entities found. The list of context words used for disambiguation is: **passaport**, **passaporte**. In terms of representation, this tag is also very simple. Only the following two representation formats are accepted:

- All characters together - **B785421**
- Only first character separated by space - **B 785421**

- NumCartaoDeCredito: This entity is assigned to the sensitive data related to the credit card number, but as referred in Table 4.2, it also corresponds to American Express numbers. For this type of data, no validation or control was found. This way, the extraction of this type of entity is based only on the format and on simple rules. The numbers that respect the following rules are accepted:

- Numbers of 16 digits started by 4, 5 or 6 that do not have more than four consecutive digits repeated
- Numbers of 15 or 16 digits started by 34 or 37

The first rule concerns numbers that represent a credit card. Whenever sixteen digits are found, we only confirm if the initial digit corresponds to 4(four), 5(five) or 6(six). Also, if in the 16 digits the same number is not repeated more than consecutive four times. The second rule is for American Express numbers, where we check if the number is composed of 15(fifteen) or 16(sixteen) digits, and if that the initial numbers are 34(three-four) or 37(three-seven).

This is another case in which there is a context list with the most common words and most commonly used acronyms: **crédito**, **credito**, **credit**, **Visa**, **CVV**, **CVC**, **american express**, **express**. This entity only appears in documents in two possible ways, which are:

- 16 consecutive digits- **456478961023**
- groups of 4 digits separated by space or hyphen- **4564 7896 1023** or **4564-7896-1023**

- IdentificacaoBancaria: This entity class is related to the Bank Identification number, this type of sensitive data is usually called 'NIB' or 'IBAN'. In this work, we have implemented rules only for the recognition of Portuguese NIBs. In the case of IBAN, we used the schwifty library[3] to validate the numbers regardless of the country. The rules used for the discovery of this entity were:

- Numbers of 21 digits
- Numbers with 4 character followed by 21 digits

The first rule corresponds to the NIB, where the last two digits consist of control digits. This verification was implemented by making the rest of the entire division of the first 19(nineteen) numbers by 97(ninety-seven). Subtract 98(ninety-eight) from the number obtained, if that value is equal to the last two digits we consider that the number is valid. In the case of IBAN, the library itself carries out this control check. A list of context words was also composed, with the words: **nib**, **iban**, **Bancária**, **Bancaria**. Regarding the format, the only formats accepted for this class of entity are:

- 21 digits or 25 characters all together - **PT5000000000000000000000**
- Separated by spaces - **PT50 0000 0000 0000 0000 0000 0**

- NumSegSocial: This entity corresponds to the social security number. The rules implemented for the extraction of this entity are simple since the number only allows one format. The format of this entity consists of 11(eleven) consecutive digits and must always appear together. In addition, there is a control validation that has also been implemented, that consists of:

[3] `https://pypi.org/project/schwifty/`

- Calculate the sum of the first 10 digit products by (29, 23, 19, 17, 13, 11, 7, 5, 3, 2)
- Calculate the rest of the division by 10
- Subtract the result by 9
- If the result is the same as the last digit, the number is valid

In addition to the validation, a context list is used with the following words: **segurança**, **seguranca**, **social**, **NISS**.

- NumCartaConducao: This entity is the driving license number. This number has undergone several changes over the years, so there are several possibilities, and the older numbers are different from the more recent ones. In this case, the implemented ruleset has a goal the extraction of Portuguese numbers only. The validation of this type of entity is very similar to the validation of the NumPassaport entity, the set of used rules is as follows:

- 1 letter followed by 7 digits
- 2 letters followed by 6 digits

The initial letters are not case sensitive and may correspond to any letter of the Portuguese alphabet. This number is another one for which there is no control check, so the only disambiguation mode is the context in which it is inserted. The list of context words is: **licença**, **licenca**, **carta**, **condução**, **conducao**, **cédula**, **cedula**. Regarding the format, in both cases, there may be a hyphen or a space separating the initial letters from the digits. So, the only accepted formats are:

- All character together - **L2358629**
- Separated by space - **L 2358629**
- Separated by a hyphen - **L-2358629**

- NumUtenteDeSaude: This entity should extract sensitive data concerning the national health number. For this type of data, there is no validation, and the only rule that must be fulfilled is:

- Numbers with 9 digits started by 3 or 4

This entity is quite simple, and for this reason, the context should be used to disambiguate when this type of data is found in the text. The list of context words is: **utente**, **saude**, **sns**. The only accepted format is all consecutive numbers without spaces.

- CodigoPostal: This entity class corresponds to the data related to the postal code. This rule has been implemented respecting only the Portuguese format used for postal codes. There is no control verification, or any list of context is made to this entity. In this case, no list of

context words has been added because in the text the postal code is usually associated with an address and not with the set of fixed words. This entity must comply with a simple regex with the following format: XXXX-XXX. Where the X's correspond to any digit between 0-9. This is the only format accepted for this entity, i.e. four digits followed by a hyphen followed by three digits.

- EnderecoEletronico: This entity corresponds to the E-mail address and is different from the others because it is not numeric sensitive data, and as such, it has no validation or context list. Even so, it follows a very restricted pattern, being the best way to discover it through a simple set of rules. This entity must respect the following format: [xxxxxx]@[xxxx].[xxx], where the x value represents any alphanumeric character or punctuation mark.

- Tempo: Finally, the entity 'Tempo' which means time in Portuguese aims to extract from the text expressions that can be sensitive data. This entity is not only discovered in this rules-based model, due to the vast diversity of possible formats, we only consider the following:

- Numeric date with format Day, Month, Year
- Numeric date with format Year, Month, Day

On both rules implemented above, only numerical characters are accepted, except for the month. All numbers between 1(one) and 31(thirty-one) are accepted as 'Day', and 0(zero) is also accepted as the first character as long as the number is composed of two digits. The 'Month' field accepts all numeric values between 1(one) and 12(twelve), as well as every month written out in full in Portuguese or English. Finally, the 'Year' field is valid for all the numbers of two or four digits. For this entity, a validation was made to confirm if the dates corresponded to real dates. In this mode, the '31^{st}' day of the month '02' is not accepted, nor the '29^{th}' day of month '2' if the year is '2019'. The date formats accepted are:

- Separated by space - **05 05 2019**
- Separated by point, hyphen or bars - **05.05.2019** or **05-05-2019** or **05/05/2019**

4.3.2 Lexicon Based Models

As we saw in Figure 4.1, the NER Module is composed of three sub-components. In this section, we focus on the development of Lexicon-Based models, which is the second component in the processing chain. The choice of this approach is due to the lack of corpus in Portuguese classified for the task of Named Entity Recognition and the good results achieved with this type of approach [Ratinov and Roth2009]. These lexicon-based models combine the results of morphological analysis, a set of lexicons and techniques of stemming and lemmatization. The goal is the recognition of the entity classes: PESSOA, LOCAL, PROFISSAO, MED, VALOR, and TEMPO. For each entity, we used different lexicons with their own characteristics. This

type of implementation consists in comparing the tokens present in the text with the lexicon and understanding if they correspond to the same entity.

The first entity class is PESSOA, this entity corresponds to the names of people. For this implementation two different lexicons were used, one of the female names and the other of male names, available on the Public Administration Data Portal[4] and containing all names registered in Portugal since 2007. The total of both lexicons has more than 4569 names. The names in the lexicons consist only of forenames, i.e. first name, and no surname are included in this lists, but in the discovery of entity PESSOA, the purpose is to identify the full name. For that, the part-of-speech tagging classification was used, more precisely the 'NOUN' and 'PROPN' tags. Besides these, the tag 'ADP' is also used for the compound names such as: 'Maria de Sousa', where the token 'de' also belongs to the entity. In the case of class PESSOA and class LOCAL the word shape is taken into consideration, i.e. the first character must be capital.

The LOCAL entity follows the same implementation used for the names above. The words must belong to the lexicon and be classified as 'NOUN' or 'PROPN' at the morphological level. The acceptance of 'ADP' as a sensitive entity is maintained as long as this word appears in the text among two other entities that belong to the same sensitive entity. The lexicons used in this case are also two, and each entry may correspond to more than one word, as is the case of 'United Kingdom'. The first lexicon used corresponds to all the countries and capitals of the world[5], and has 13 155 entries. The second lexicon with more than 18 034 entries corresponds to the set of all Portuguese cities, municipalities, and parishes, available in Gov Data[6].

For the entities, PROFISSAO and MED we have also used the comparison with lexicons using a different approach. The two lexicons are from Wikipedia[7] information. The entity PROFISSAO corresponds to the sensitive data of people's jobs or professions, and the lexicon has 649 different entries. The second entity MED corresponds to medical data more specific names of diseases, allergies, intolerances, and medicines. To use the lexicons for the recognition of these two types of entities we used the resources available in the SpaCy library. SpaCy provides an API with entity matcher and phrase matcher functions. PhraseMatcher allows you to efficiently combine large lists, and match sequences based on those lists and the comparison is not done in a linear way but with stemming and lemmatization techniques and ignoring case sensitive and plural or singular words. In addition to a more efficient comparison, PhraseMatcher also accepts a set of rules that can be defined. For example in the case of the entity PROFISSAO accepting patterns that consider morphological analysis, if there is a set of tokens in the text classified with the tags 'NOUN' followed by 'ADJ'. This type of rule serves to extract cases like 'Director Científico'(Scientific Director) which appears many times in documents and is not part of the lexicons.

[4] `https://dados.gov.pt/pt/datasets/nomesfeminino/`
[5] `https://www.geodatasource.com/`
[6] `https://dados.gov.pt/pt/datasets/`
[7] `https://pt.wikipedia.org`

Different from the previous entities, the entity VALOR was not implemented with the help of a library. This class should extract from the text all the existing values, which may correspond to the value of a contract, the value to be paid for a fine, etc. In this type of entity, the value can be written both numerically and in full, and to cover both cases, we used the part-of-speech tagging classification. For the words or symbols that come associated with the values, besides the use of the tag 'SYM' of POS tagging, a lexicon has been created with the most relevant words that should be considered. This lexicon consists of words such as 'dollar', 'euro', 'millions', etc. In this case for comparison with the lexicon, we used the stemming of the words of the text.

The last entity implemented was TEMPO, in the ruled-based model section, we already explained one implementation made for this entity, but with this model, the goal was the extraction of full dates. Full dates correspond to the date in the text as follows: '12 de Maio de 2019'. For this purpose, a lexicon was created with all the months in Portuguese and English, as well as their abbreviated forms. In addition to the use of the lexicon for the recognition of this entity, we also used the morphological classification analysis, specifically the tags 'NUM' and 'ADP'.

Some of these entities were also implemented with Machine Learning models. The goal of this work was to understand how to achieve the best results for each class of sensitive data.

4.3.3 Machine Learning Models

The Machine Learning Models is the last sub-component on the chain of the NER module, as we saw in Figure 4.1. In chapter 2, we conclude from the current state-of-the-art analysis that for the most ambiguous entities and for which there are no well-defined rules, the best results are achieved through machine learning methods and the most recent approaches are based on the study of neural networks. In this section, we present the Machine Learning approaches for NER used in this dissertation. These approaches were carried out for a smaller set of entities: PESSOA, LOCAL, TEMPO, VALOR, and ORGANIZACAO. For these entities and in this experiment, we had three different approaches:

- The first was the use of Named Entity Recognition tools, SpaCy and Stanford CoreNLP
- The second approach, we implement the two statistical models most commonly used in the tasks of Named Entity Recognition, Conditional Random Field and Random Forest
- The last approach is a study of a neural network in which a Bidirectional LSTM was chosen

For the different approaches implemented, were used the corpus described in chapter 5, i.e.: **HAREM golden Collection** and **SIGARRA News Corpus**.

4.3.3.1 Named Entity Recognition Tools

The first experiments were carried out through the use of Natural Language Processing Tools. In order to choose the best tools for the NER task, some tools were compared [Heuss et al.2014], and some criteria were defined. The main criterion was to choose tools capable of making multilingual recognition, focused on the Portuguese language. The remaining criteria were that the tool should be completely free of charge and that it should allow the training with customized classes of entities. After the study of the Natural Language Processing tools available and taking into account our criteria, two tools were chosen: SpaCy[8] and Stanford CoreNLP[9]. From the two tools, only SpaCy has a NER model for Portuguese, however, both tools allow training models for specific domains and language, so for this work both tools were trained with corpus in Portuguese.

We used for training the models the HAREM golden Collection and SIGARRA News Corpus, but taking into account the scope of this dissertation, some entities were discarded in both corpus. The used entities are: PESSOA, LOCAL, TEMPO, VALOR, and ORGANIZACAO, except the entity VALOR, which is not present in the corpus SIGARRA. Since each tool has its own input format the corpora had to be transformed to serve as input for training the models. The transformations, characteristics, and steps used to train the models for each tool were:

- Stanford CoreNLP: This provides a set of human language technology tools, one of which is the Stanford NER, used in this dissertation. Stanford NER is a Named Entity Recognizer Java implementation, provided under the GNU General Public License[10]. In this work we used the latest version available in 2018, version 3.9.2, which contains improvements in the NER pipeline and supports Java 11. The NER models available by default supports languages such as: Arabic, Chinese, English, French, German and Spanish. However, the source code is included in the available license, and the package includes components for command-line invocation and a Java API that allows model training for other languages. Stanford NER's named entities classifier model is also known as CRFClassifier, i.e. the library uses a Conditional Random Fields model[Ritter et al.2011] for the Named Entity Recognition task.

As each tool has specific requirements for training the NER model, in this case, it was necessary to make changes to the two corpora. Both HAREM and SIGARRA have been transformed to be used as inputs for the training model. The Stanford NER model requires as an input file, a file like the one we can see in Listing 4.12, that is the tokenized text, where each line contains a token and its entity class separated by three spaces. To tokenize the text, we used the preprocessing module (chapter 4.2). In this format the entities are represented only by their

[8] `https://spacy.io/.`
[9] `https://stanfordnlp.github.io/CoreNLP/`
[10] `https://nlp.stanford.edu/software/CRF-NER.html`

tag, e.g. 'TEMPO', in the case of tokens that have no entity associated is assigned the tag 'O'.

```
é     O
o     O
sétimo     O
filme     O
de     O
António-Pedro     PESSOA
Vasconcelos     PESSOA
,     O
68     VALOR
anos     VALOR
```

Listing 4.8: Stanford NER input format

After having both corpora in the format accepted by Stanford NER, it was possible to train the model. For the training of the CRF model, in addition to the training file, the features to be used in the training process are also defined. In this model, the only feature that could be parameterized would be the tolerance value. The tolerance value is used when the optimization function tries to find an unlimited minimum of the target function from the initial value, having to keep precision within the tolerance value. In this case, the parameterization of this value in the model did not imply significant gains, so it was not used. The evaluation and analysis of the results for both corpora are in chapter 6 of this document.

- SpaCy: This is an open-source library for advanced natural language processing in Python. It is a library designed specifically for use in production, and this was a positive point that led to the choice of this tool. SpaCy's Entity Recognition model is a library's own model, called the Thinc linear model, which consists of a multi-tasks CNN trained for the different languages [Jiang et al.2016]. The standard NER model provided by SpaCy includes a large number of languages such as: German, Greek, English, Spanish, French, Italian, Dutch and also Portuguese. The existing SpaCy model for Portuguese is a Convolutional Neural Networks trained in the Universal Dependencies[11] [Rademaker et al.2017] and WikiNER corpus[12] [Ghaddar and Langlais2017]. The model assigns context-specific token vectors, POS tags, dependency analysis, and named entities. It supports the identification of PER(Person), LOC(Local), ORG(Organization) and MISC(Miscellaneous entities) entities. Because the model is trained on Wikipedia, it may perform worse in many other text genres, but SpaCy also provides a CCBY-SA 4.0 license for model training. This training gives us the freedom to add new arbitrary classes to the NER model and train with different text genres.

[11] https://universaldependencies.org/

[12] https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

In this dissertation, the SpaCy model was trained with the two available corpora: HAREM and SIGARRA, this training adds two new entities to the model: VALOR and TEMPO. In other words, the entities already made available by the library were used, and the SpaCy NER model was trained only for the VALOR (Value) and TEMPO(Time) entities. The SpaCy version used for this dissertation was version 2.1.0, the most recent version at the beginning of this work. As in the previous implementation of the Stanford NER model, the data also had to be transformed in order to serve as training input to the model. The SpaCy requires input data in standoff format, which consists of a list of tuples, containing the original text and the set of offsets and classes of entities. In this dissertation, we chose to have two files for each corpus. The first with the text, as we see in Listing 4.9. The second file, with the offsets of the entities, the initial and final position of the entity in the text and the respective entities separated by three spaces, like the example represented in Listing 4.10 that referring to the text given in Listing 4.9.

```
George Steiner esteve recentemente em Portugal e de acordo com notícia no Ciência Hoje
```

Listing 4.9: SpaCy text input format

```
1  15   PESSOA
39 47   LOCAL
75 87   ORGANIZACAO
```

Listing 4.10: SpaCy label input format

After transforming the corpora into the standoff format, it was possible to train SpaCy's CNN model and make predictions with the new model. In addition to the input data, the SpaCy NER model does not allow any extra parameterization. The only parameter that can be customized is the number of iterations of the model. In chapter 6 we show the results achieved with the training of the two corpora and describe details about the training of the model.

4.3.3.2 Statistical Models

After the previous experiments, the next goal was to train our own models in order to have greater manipulation at the level of features and hyperparameters. After the study of the existing state-of-the-art, we decided to choose two Statistical Models. A statistical model (SM) is the mathematical data model that incorporates probabilities for the data generating mechanism and has identified unknown parameters that are usually interpretable and of special interest [Neter et al.1996]. Statistical models are usually specified as a mathematical relationship between one or more random variables and other non-random variables [Adèr2008].

NER approaches with statistical models typically require a large amount of training data, and this type of approach has not been used as much to avoid part of the annotation effort [Lin and Wu2009] [Nothman et al.2013]. Since we have two corpora, although not very extensive, this difficulty can be overcome. With this, the goal with the choice of statistical models is to

implement supervised learning models with the annotated corpus HAREM Golden Collection and SIGARRA News Corpus. The models chosen for this approach were a **Conditional Random Fields Model** (CRF) and a **Random Forest Model**. The implementation details of each model are explained in this chapter.

For the training of the two models, both corpora had to be previously transformed. In this case, the aim was to build two files, one file for each corpus. Each file consists of three columns, as we can see in Listing 4.11. The first column corresponds to the words tokens, the second to the POS tag and the last to the entity class tag. The annotations of the entity classes follow the CoNNL format, that is, a named entity with IOB tagging.

```
Até     ADP     B-TEMPO
1947    NUM     I-TEMPO
,       PUNCT   O
o       DET     O
jovem   ADJ     O
William PROPN   B-PESSOA
M.Gaines PROPN  I-PESSOA
não     ADV     O
passava VERB    O
de      ADP     O
um      DET     O
```

Listing 4.11: Corpus format to be used in statistical models

The classes of entities discovered with these models are: PESSOA, LOCAL, TEMPO, VALOR, and ORGANIZACAO. The methodologies used for the implementation of Conditional Random Fields and Random Forest models were:

- Conditional Random Fields: The CRF model is one of the statistical models implemented for this work. Based on the implementation of Korobov M. and Lopuhin K., for the corpus CoNLL2002, available at GitHub[13], as well as the implementation of the NERP-CRF system [Amaral and Vieira2014] [Amaral and Vieira2013]. The set of tasks performed to define and extract features was based on the work of McCallum A. and Li W [McCallum and Li2003]. The implemented model was trained with two Portuguese corpora, HAREM and SIGARRA, after both went through the entire transformation process.

The implemented model is a non-directional graphical model used to calculate the conditional probability of the output nodes values based on the values assigned to the corresponding input nodes [Lafferty et al.2001]. This model makes a first-order Markov independence assumption, so it can be understood as a conditionally trained finite state machine. The model has $x = (x_1, ..., x_m)$ as input sequence, where $\mathbf{x}$ are the words of a sentence. We have $y = (y_1, ..., y_m)$ as the output sequence states that corresponds to the classes of named entities, where $\mathbf{Y}$ is a

[13] https://github.com/TeamHG-Memex/eli5/blob/master/notebooks/sklearn-crfsuite.ipynb

set of finite state machines(FSM) with each state associated to an entity class. We model the conditional probability through $p(y_1, ..., y_m|x_1, ..., x_m)$. CRFs define a conditional probability of an output state given an input sequence, by the Hammersley-Clifford theorem [Cheung2008]:

$$P_\Lambda(y|x) = \frac{1}{Z_x} exp(\sum_{m=1}^{M} \sum_{k} \lambda_k f_k(y_{m-1}, y_m, x, m)) \tag{4.1}$$

Where Z_x is a normalization factor for all state sequences, λ_k is a learned weight for each feature function, and $f_k(y_{m-1}, y_m, x, m)$ is an arbitrary feature function over its arguments. The feature function has been set to have a value of 0 in most cases and to have a value of 1 if y_{m-1} is the state 1, where 0 corresponds to the tag 'O' and 1 correspond to the tag 'PERSON'. The feature function can access the entire input sequence, including queries on previous and next words, so $f_k(\cdot)$ can range between -$\propto$....+$\propto$. Higher λ weights make their corresponding FSM transitions more likely, so the weight λ_k should be positive. CRF models define the conditional probability of a class sequence based on the total probability of the state sequences, $P_\Lambda(l|x) = \sum_{y:l(y)=1} P_\Lambda(y|x)$, where l(y) is the sequence of class entities corresponding to the states in sequence y. Defining that the normalization factor, Z_x, is the sum of the "values" of all possible state sequences (Equation 4.2), the number of state sequences is exponential in the length of the input sequence, **M**.

$$Z_x = \sum_{y \in Y^M} exp(\sum_{m=1}^{M} \sum_{K} \lambda_k f_k(y_{m-1}, y, x, m)) \tag{4.2}$$

The weights of a CRF model, $\Lambda = \left\{\lambda, ...\right\}$, are defined to maximize the probability of conditional log-likelihood of an entity class in a training set. The second summation is a Gaussian prior to the parameters (with variance σ) which provide smoothing to help cope with the scarcity of training data.

$$L_\Lambda = \sum_{j=1}^{N} log(P_\Lambda(l^j|x^j)) - \sum_{k} \frac{\lambda_k^2}{2\sigma^2} \tag{4.3}$$

When training classes make the sequence of states unequivocal, the likelihood function in exponential models such as CRFs is convex, and so we can find a global optimal value [Limketkai et al.2007].

The features, f_k, are based on the set of features used. In this implementation, the POS tags can be seen as pre-extracted features, but more features were extracted, such as:

- Parts of words through stemming
- Simplified POS tags
- Confirmation of capital letters, lower case letters, titles and digits

- Resources from nearby words

To consider the effect of adding a new feature, a new sequence template is defined with an additional feature, g, with weight μ.

$$P_{\Lambda+g,\mu}(y|x) = \frac{P_{\Lambda}(y|x)exp(\sum_{m-1}^{M} \mu g(y_{m-1}, y_m, x, m))}{Z_x(\Lambda, g, \mu)} \tag{4.4}$$

By converting the corpus to a dictionary list format with the tokens and all associated features, we were able to train and test the Conditional Random Fields model.

- Random Forest: This is the second statistical model implemented in this work. Random Forest model is a machine learning algorithm that works through decision trees, the model is trained to create a group of decision trees with a random subset of the data. This model was chosen for the many comparisons works between this model and the previous CRF [Tran et al.2015] [Baldwin et al.2015], and because it is considered an excellent approach to classification and regression problems [Liaw et al.2002]. The training of this model is performed with the same two datasets as the previous one. The implementation carried out follows the implementation of Shoumik available at Kaggle[14] and the approach of feature extraction of Jin N. [Jin2015]. In terms of features, we tried to bring the model as close as possible to the previous one, in order to compare them.

The implemented model is a simple tree-based classification model. A Random Forest consists of a large number of deep trees, where each tree is trained using a random selection of features [Jiang et al.2007], so as to gain a complete understanding of the decision-making process. Each tree takes a path (or paths) from the tree root to the leaf, consisting of a series of decisions, held by a particular feature, each of which contributes to the final predictions. In this case the model with **M** leaves divides the feature space into **M** regions, **Rm**, $1 \leq m \leq M$. And the tree prediction function is then defined by:

$$f(x) = \sum_{m=1}^{M} c_m I(x, R_m) \tag{4.5}$$

where **M** is the number of leaves in the tree, **Rm** is a region in the space of the features corresponding to leaf **m**, **cm** is a constant corresponding to region **m**, and finally **I** is the indicator function. The indicator function returns 1 if **x**$\in$**Rm** and 0 if not. The value of **cm** is determined in the training phase of the tree and **Rm** is the extracted features, which correspond to the same features in the previous model.

[14] `https://www.kaggle.com/shoumikgoswami/ner-using-random-forest-and-crf`

Before training the Random Forest model, we converted the data into a simple feature vector for each word. That is, each vector consists of the word and the set of features used in this model.

4.3.3.3 Neural Network Model

This approach, based on neural networks, was used in order to try to follow the recent trends of the current state-of-the-art and the use of Deep Learning in NLP tasks[Sang and De Meulder2003]. The goal was to test a NER approach based on neural networks with the corpus we have available in Portuguese. After the literature review (chapter 2), we noticed that the most used approaches and also the ones that produce better results [Yadav and Bethard2018], have been using LSTM (Long Short-Term Memory).

For this experiment, the SOTA algorithm [Nie and Fan2006] was implemented following the approach of Chiu J. and Nichols E.[Chiu and Nichols2016]. This implementation was performed in Python with Keras and Tensorflow, but differs from the original paper, because we did not consider lexicon, and the Nadam optimizer was used instead of the SGD (Stochastic Gradient Descent).

In the original article, the algorithm was trained with the dataset CoNLL-2003, and 2012 and the number of tokens per corpus were much higher than the corpus available in Portuguese. The implementation is based on a Bi-directional LSTM (Bi-LSTM) [Schuster and Paliwal1997], and it also uses a convolutional neural network (CNN) to identify character-level patterns. The LSTM cells are the building block of recurrent neural networks (RNNs). While plain LSTM cells in a feedforward neural network process text from left to right, Bi-LSTMs also consider the opposite direction, and this allows the model to discover more patterns. In this case, the model not only considers the sequence of tokens after a token of interest but also before the token of interest. In Figure 4.5 of Cui et al. [Cui et al.2018], it is possible to see the main idea of a Bidirectional LSTM. We can note that x represents the input sequence, h the output sequence from forwarding or backward runs (defined by the arrow), and y is the concatenated output sequence where σ represents the forward and backward output elements.

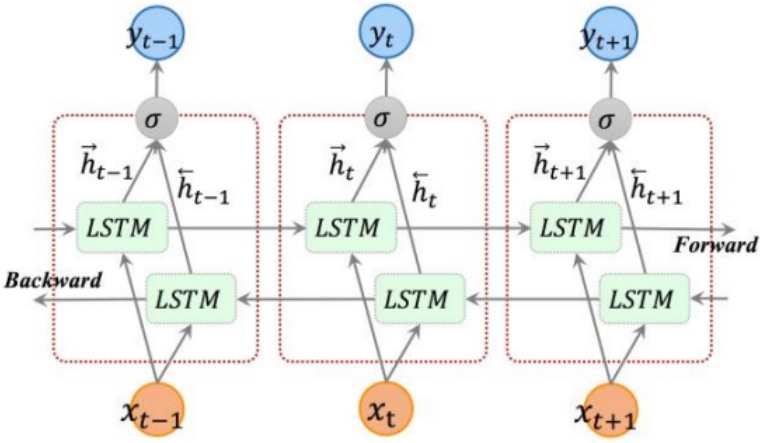

FIGURE 4.5: Unfolded architecture of bidirectional LSTM with three consecutive steps, from Cui et al. (2017)

The first step of the implementation consists of loading the training, development and test datasets. For this implementation, an embeddings representation was used for each word. All words were mapped to vectors through the embeddings representation provided by fastText[15] [Athiwaratkun et al.2018]. This means that all words and characters were mapped to real numbers that the neural network can work with. All words, except the stopwords that have been removed, are mapped using the pre-trained Portuguese dictionary of fastText. At the model architecture level, the Bi-LSTM layer forms the core of the network and is composed of three entries:

- Character-level patterns are identified by a convolutional neural network
- Word-level input from FastText embeddings
- Casing input (whether words are lower case, upper case, etc.)

Figure 4.6 presents the architecture and layers of the model we followed for this implementation [Chiu and Nichols2016].

[15] `https://fasttext.cc/docs/en/crawl-vectors.html`

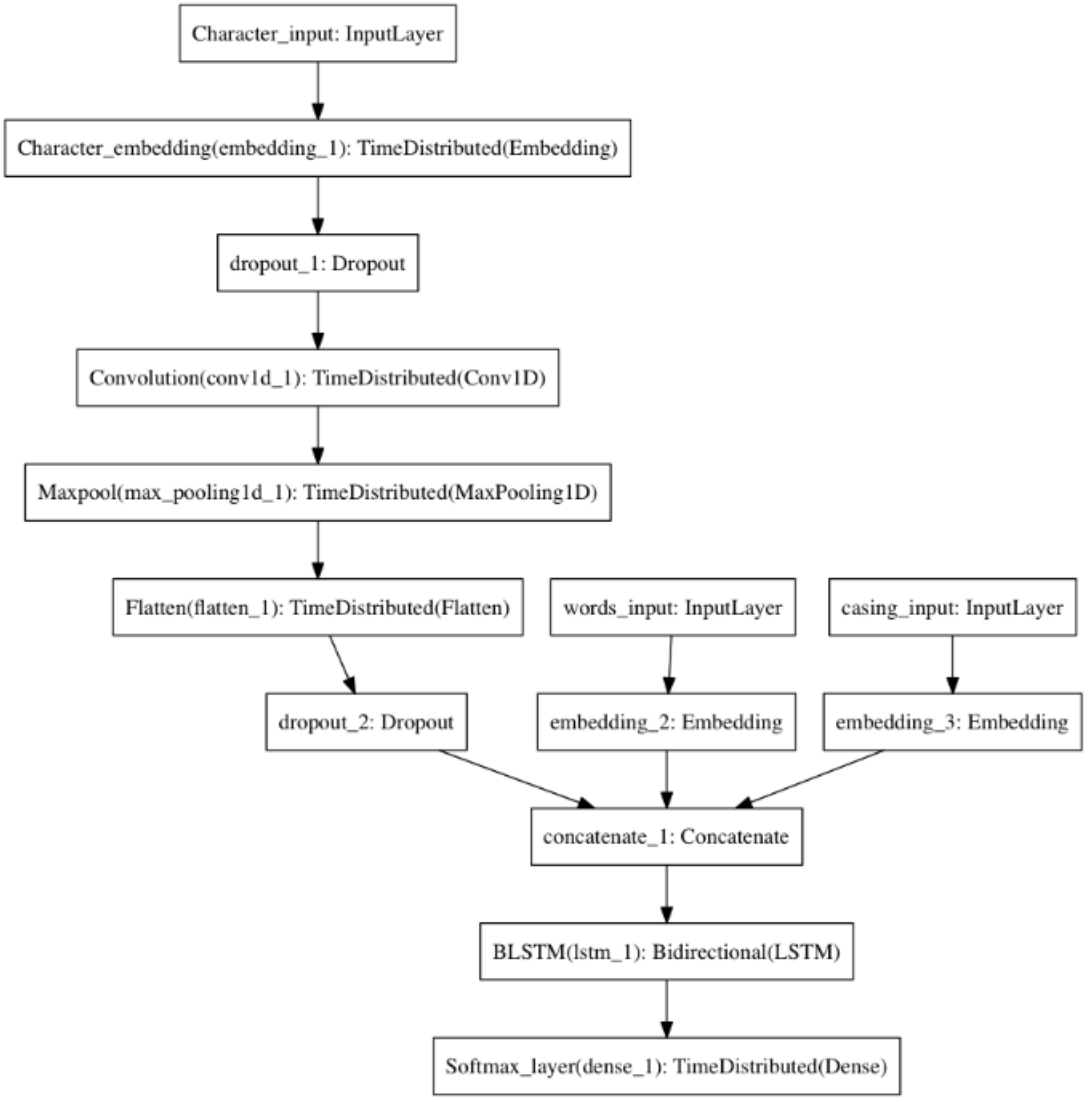

FIGURE 4.6: Model architecture and layers representation, from Chiu and Nichols (2016).

In Figure 4.6, we can see the three entity layers and the output softmax layer. After training the model the softmax activation layer generates the final outputs. The results achieved by this model with SIGARRA corpus in Portuguese can be seen in this document in chapter 6.

4.4 Post-processing Module

Post-processing is the last module of the Named Entity Recognition component chain (Figure 4.1). This module is meant to treat the results achieved from the previous NER module and return the text (output), and the entities found with the desired format to the user. This module allows the user to choose to view the result in five different ways, since there are different types of outputs that can be shown, depending on their preference. The previous NER module returns to

this module the input text with all preprocessing transformations annotated in CoNLL format, as represented in Figure 4.7.

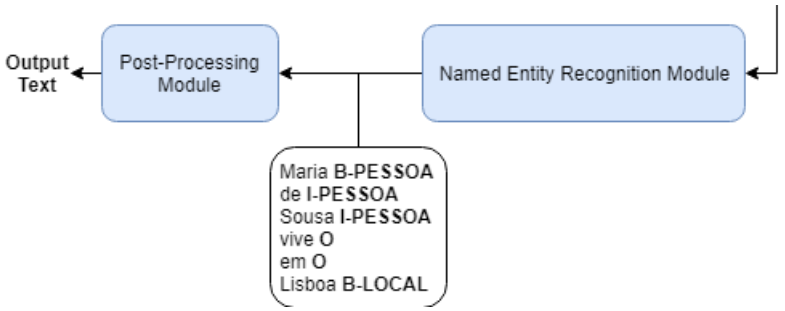

FIGURE 4.7: Post-Processing Module input

This module has as a first task the text correction. Text correction consists of rewinding the Segmentation and Tokenization tasks so that the output corresponds to the same text given as input. This way, in this first task, the NER module's output is compared with the text provided as input so that the named entity tags correspond to the initial text provided. After this the output can be presented in five different ways, depending on the user option. The output can consist of text annotated with IOB tags, text annotated in XML format with "START" and "END" tags, it can consist only of the frequency of entities found, offsets of the entities or finally in an HTML format with the representation of the entities.

The first possible output is the result annotated according to the CoNLL format. In Listing 4.12, we can see the text given as an input to the NER component, which will be used in the following examples. Listing 4.13 is the result presented to the user when he has chosen this first format. This is the simplest output format and allows the user to handle the data in a way that suits him best depending on the purpose.

```
Nascido em Março de 1960 o João Cardoso reside em Lisboa
```

LISTING 4.12: Input text to NER component

```
Nascido O
em O
Março B-TEMPO
de I-TEMPO
1960 I-TEMPO
o O
João B-PESSOA
Cardoso B-PESSOA
reside O
em  O
Lisboa B-LOCAL
```

LISTING 4.13: Text output with CoNLL format

The second output format is very similar to the first, and it was implemented due to a requirement of the DataSense Project. This second output format returns the texts as a string with "START" and "END" tags. That is, before each entity in the text there is a "<START: tag-name>" tag in which the tag-name corresponds to the entity class. At the end of the entity there is a "<END>" tag, which marks the end of the found entity. This representation format consists of the NE format used by the OpenNLP library[16]. An example of this format can be seen in Listing 4.14.

```
Nascido em <START:TEMPO> Março de 1960 <END> o <START:PESSOA> João Cardoso <END> reside em
 <START:LOCAL> Lisboa <END>
```

Listing 4.14: Text output with START : END tagging

The third possible format consists of counting the number of entities found. It is a more numerical and summarized output, not going into the details of the previous ones. It is available because one of the features of the DataSense Project was the enumeration of the entities found. Therefore, in this case, the result is only the presentation of the number of entities for each class of entity. In Listing 4.15 it is possible to see the output corresponding to the input presented in Listing 4.12. This format is a dictionary with key: value representation, in which the key consists of the class and value to the number of entities found.

```
{"TEMPO": 1, "PESSOA": 1, "LOCAL": 1}
```

Listing 4.15: Text Entity distribution tagging output

Another output format of the component is the representation of the text with offsets. This representation was a fundamental requirement of the project in order to comply with sensitive data regulations. The purpose of this format was not returning to the user sensitive data. This format returns the initial and final position of the word(s) in the text and the corresponding entity class. In Listing 4.16 we can see the output result with this format, where a key: value representation is returned, the key representing the entity class and value the offset of each entity.

```
{"TEMPO": [(11, 25)], "PESSOA": [(27, 40)], "LOCAL": [(50, 57)]}
```

Listing 4.16: Text output format with offset tagging

The last possible representation is the most visual and the one used for visualization in the DataSense platform of INOV and also in the SocialOpinion Project. This representation consists of an HTML image of the input text with labels that represent the entities found. In Figure 4.8, it is possible to see an example of the result. The HTML is generated with the help of the SpaCy library that provides different ways to represent NE.

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

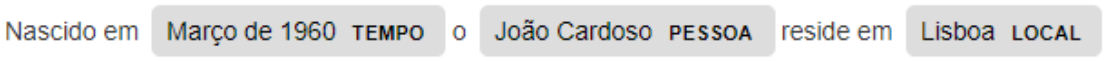

FIGURE 4.8: Text HTML tagging output with SpaCy display

Chapter 5

Metrics and Resources

This chapter introduces some evaluation metrics and the datasets used in work. The performance evaluation was carried out using the metrics introduced in this chapter. These metrics aim to understand the results of the developed work, and it is through them that it is possible to perform a comparison between the different tests made. This chapter not only presents the datasets and corpora used but also describes the new dataset that was created specifically for the evaluation of this work.

5.1 Metrics

Evaluating Natural Language Processing tasks or any other work is an essential part of any project. Metrics evaluate the quality of a model by comparing the model's output (predicted result) to the original annotation (label result). A model that makes a better prediction should yield a higher metric score. However, different problems and models need different metrics and evaluation forms [Marrero et al.2009].

Both the implementations of Part-of-Speech Tagging and Named Entity Recognition were evaluated. The evaluation of the POS Tagging results considers accuracy as the metric. Using this metric, it was possible to draw conclusions by comparing the results to other systems and to other tests we performed. On the other hand, evaluating Named Entity Recognition work allows us to know if new work is evolving in a positive way, getting higher precision and recall. There is a need for systematic evaluations so that all NER systems have the same performance evaluation standards. There are multiple proposed techniques to rank the NER task, but the evaluation is normally based on precision, recall, and f1-score metrics [Sang and De Meulder2003].

The overall value of the metrics (accuracy, precision, recall, and f1-score) can be computed using different averaging strategies [Stubbs et al.2015], such as macro-average and micro-average:

- **Macro-average:** The metrics are calculated from the average results of the whole dataset, giving the same weight to each class;
- **Micro-average:** The metrics are calculated from the average results for each document of the dataset, giving the same weight to each document;

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. But the goal for this dissertation considers the datasets for the desired entity classes and compares the output of this work to the output of the datasets used, through metrics. The classified entities were marked as True Positives(TP), True Negatives(TN), False Positives(FP) and False Negatives(FN). The metrics used are defined as:

- **Accuracy:** The ratio of true results (both TP and TN) among the total number of cases examined;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

- **Precision:** The ratio of correct answers (TP) among the answers produced (Positives). This means checking if the answers marked as positive are truly positive;

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

- **Recall:** The ratio of correct answers (TP) among the total possible correct answers (TP and FN). This means checking if all the positives are marked;

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

- **F1-score:** The average of precision and recall;

$$F1\text{-}scoree = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.4}$$

5.2 Datasets

The datasets are crucial for the success of any Machine Learning work, but as seen in the state-of-the-art, the Named Entity Recognition task for the Portuguese language presents several problems due to the lack of training and testing datasets. The only freely available Portuguese dataset annotated with classes of entities was the one developed for the HAREM events [Santos et al.2006]. There is another annotated corpus, CINTIL[1] dataset, but it is not publicly available. For this reason, the HAREM is the dataset that was used in this dissertation and it

[1] `http://cintil.ul.pt/pt/`

is described in Section 5.2.1. Another Portuguese dataset is the SIGARRA News Corpus, annotated for named entities, consisting of a set of news manually annotated[2]. In the next sections, a description of both datasets is presented, as well as a description of the DataSense NER Corpus built exclusively for this project in order to enable a greater approximation to the type of data and classes of entities that we have as objectives.

5.2.1 HAREM Golden Collection

The HAREM Golden Collection appeared in the 2008 HAREM event [Freitas et al.2010]. The Golden Collection data consists of a set of 129 text documents with 86 000 words covering the Portuguese Language. They are from several genres, such as: News, Interviews, Blogs, Publicity Texts, Web Pages, etc. In these documents, named entities have been manually identified, semantically classified and morphologically tagged. It defines three levels of entity annotations, namely categories, types and subtypes. The categories present in HAREM corpus are the following: PESSOA (Person), LOCAL (Local), TEMPO (Time), ORGANIZACAO (Organization), OBRA (Work), VALOR (Value), COISA (Thing), EVENTO (Event), ABSTRACCAO (Abstraction) and OUTRO (Other).

This dataset is annotated for 10 different categories distributed in an unbalanced way, as we can see in Table 5.1. Apart from the categories, it has a total of 43 types and 21 subtypes that are not used in this dissertation.

Category	Number of Occurrences
PESSOA	2 086
LOCAL	1 453
TEMPO	1 199
ORGANIZACAO	1 084
OBRA	449
VALOR	352
COISA	345
EVENTO	337
ABSTRACCAO	382
OUTRO	79
Total	7 766

Table 5.1: Number of occurrences for each category, in HAREM Golden Collection

Table 5.1 shows that PESSOA is the most common category with 2 086 entities, and OUTRO is the least common. This golden collection has a total of 7 766 category occurrences, and for this dissertation, only the following categories were considered: PESSOA, LOCAL, TEMPO,

[2] `https://hdl.handle.net/10216/106094`

ORGANIZACAO, and VALOR, reducing the classification to only five groups of HAREM Golden Collection categories.

The annotation of the HAREM corpus is made according to the XML format and contains tags and additional information that was not used. All annotations start with the 'EM' tag and end with '/ EM', as well as an ID attribute for easy identification. An example of an annotation is in Listing 5.1.

```
<EM ID="H2-bbb-226" CATEG="PESSOA" TIPO="INDIVIDUAL">Luís Lourenço</EM>
```

Listing 5.1: HAREM golden collection tagging

Since the data in XML format does not serve as input, in order to feed the algorithms the corpus had to be transformed considering only the CATEG tag. It was necessary to transform the data into a CoNLL format with IOB tags, [Straka et al.2016]. The previous example Listing 5.1 can be seen in Listing 5.2 with the CoNNL format. The HAREM golden collection is transformed in the IOB tagging format to be used as input to the algorithms and also in order to facilitate the evaluation of the results obtained.

```
de O
Luís B-PESSOA
Lourenço I-PESSOA
, O
 baseado O
```

Listing 5.2: HAREM example in IOB tagging format

5.2.2 SIGARRA News Corpus

The SIGARRA's News Corpus was built to complement a master thesis [Pires2017] in the field of extraction of Named Entities for the Portuguese language.

This corpus was taken from the SIGARRA information system at the University of Porto[3]. The corpus is a set of 905 news, manually annotated, using the Brat rapid annotation tool [Stenetorp et al.2012], for training named entity recognition models. This corpus presents news for many different domains as it contains information from the different departments of the University, from Medicine to Economics. They selected eight different categories for the entities similar to those already presented in the HAREM corpus. The classes of entities in this corpus are: Hora(Hour), Evento(Event), Organizacao(Organization), Curso(Course), Pessoa(Person), Localizacao(Location), Data(Date) and UnidadeOrganica(Organic Unit). The most common class in this corpus is the Data class, however, the class Pessoa is also quite most common. In Table 5.2 we can see the distribution of the number of entities and conclude that the data

[3] https://sigarra.up.pt/

categories are unbalanced. Overall, there are 12644 entity annotations in this corpus, which makes it much more interesting than HAREM. In this corpus, there are also sets of texts in Brazilian Portuguese, but the percentage is lower than 28 %, which is not enough to compromise the evaluation.

Category	Number of Occurrences
Date	2 811
Organization	2 320
Person	2 159
Organic Unit	1 814
Location	1 593
Hour	1 015
Course	521
Event	411
Total	12 644

Table 5.2: Number of occurrences in SIGARRA News Corpus

Analyzing the values in Table 5.2, the corpus is more than twice the size of the HAREM collection (HAREM with approximately 86 000 tokens, and SIGARRA News with 185 000 tokens), having also double the number of entity annotations (HAREM with 7255, and SIGARRA News with 12644). Nevertheless, for this dissertation, there are some entities that were not used, such as Hora, Curso, and Evento. The entities 'Organizacao' and 'UnidadeOrganica' were merged into a single entity called 'Organizacao', this new entity having 4134 occurrences. Like the HAREM corpus, this one was annotated in a format that was later changed to the IOB tag format. The corpus is available[4] in the standoff format, where the category is represented always with the offset that represents the entity corresponding to the original text, as we saw in Listing 5.3.

```
T1   Data 52 75  22 de fevereiro de 2017
T2   Localizacao 86 110   Auditrio Alberto Amaral
T3   UnidadeOrganica 114 120 FADEUP
T4   Data 330 345   22 de fevereiro
T5   Localizacao 369 403  Auditrio da Faculdade de Desporto
T6   Pessoa 431 445 Walter Osswald
```

Listing 5.3: SIGARRA News Corpus tagging

In Listing 5.3, we can see the representation is made by **T1** which represents the first tokens, followed by the entity **Data**, followed by the offset **52 75** in which the entity is present and finally the identified entity, **22 de Fevereiro de 2017**. This type of representation is heavier since we always need to have the original text file and the corresponding annotated text file. As in the previous case, the entire corpus was transformed into the CoNLL format before being used.

[4] https://rdm.inesctec.pt/dataset/cs-2017-004

5.2.3 DataSense NER Corpus

From the two datasets presented in the sections above, none of them presents all the classes of entities of this work, and mainly in the two corpora, neither of them respects the context of the work, being the great majority of documents News or Web pages text. For this reason, one of the key points of this work was the construction of the test corpus, the DataSense NER Corpus. This corpus was built with the aim of understanding the results obtained when applied to the real context of the DataSense Project. For this purpose, a set of documents was collected and annotated according to the desired entity classes through the Prodigy tool [5].

This new corpus largely focused on the type of documents for which this dissertation was intended, i.e. contracts, CVs, personal data forms, and other documents present in the companies' documentary databases. The purpose of this corpus is to test all classes of sensitive data, in particular, Personal Identification Numbers that are not annotated in any type of corpus, Portuguese or other languages.

The creation of this corpus followed a set of predefined rules for the DataSense Project. These rules were intended to respect the entities defined as sensitive and remove some errors that existed in the previous corpus. We adopted rules such as: for the entity 'Pessoa' only the name should be considered and not professions or positions that precede the name. For example, the sentence 'O Presidente Dr. Cavaco Silva esteve ontem presente em Espanha.' was tagged as a name only, 'Cavaco Silva', and did not include 'Presidente' or 'Dr.'.

This corpus consists of 78 documents, 1722 paragraphs and 5593 entities. In Table 5.3, we can see the distribution of the annotated entities by category. It is possible to verify that the entity with the greatest number of occurrences is the tag 'Organizacao', followed by the tag 'Local' and 'Pessoa'. This is due to the fact that the annotated documents were mostly Cv's, contracts and minutes.

[5] `https://prodi.gy/features/named-entity-recognition`

Category	Number of Occurrences
NumIdentificacaoCivil	130
IdentificacaoBancaria	233
NumCartaoDeCredito	8
NumIdentificacaoFiscal	352
NumPassaport	10
NumSegSocial	40
NumUtenteDeSaude	11
ContactoTelefonico	342
NumCartaConducao	5
Pessoa	648
Local	792
Organizacao	1296
Tempo	467
Valor	216
Med	29
Profissao	532
CodigoPostal	214
EnderecoEletronico	268
Total	5593

Table 5.3: Number of occurrences in DataSense NER Corpus

Chapter 6

Evaluation and Results

This chapter describes the evaluation methods used for each experiment and shows the results achieved, it is divided into two main sections. The first is related to the Preprocessing techniques, and experiments carried out at the Part-of-Speech Tagging level explained in sections 4.2.1. of this document. The second section corresponds to the evaluation and results achieved for the task of Named Entity Recognition, in chapter 4.3. Each section describes the methods used to carry out the evaluation and results obtained. The results of performance and accuracy, obtained in the first trial carried out by INOV-INESC Inovação for the NER component, are also demonstrated.

All implementations and tests performed were carried out on a single computer, with the following specifications: NVIDIA GeForce GTX 1060 Ti GPU, 16GB of RAM and an Intel Core i5-7600K CPU.

6.1 Part-of-Speech Tagging Evaluation and Results

The evaluation of the experiments for the morphosyntactic analysis component was performed using the Floresta Sintática Corpus, annotated with Part-of-Speech tags. The three experiments evaluated in this chapter are those presented in section 4.2.1 of the Preprocessing tasks used in the development of this work. The main goal was to evaluate the different tests performed at the level of the Part-of-Speech tagging task:

- 1st Experiment - Direct application of the POS Tagging NLTK library model, to corpus;
- 2nd Experiment - Re-training of the NLTK model from the 1st experiment, with a Portuguese annotated corpus;

- 3rd Experiment - Direct application of the POS Tagging SpaCy library model, to the corpus;

For these experiments, 30% of the Floresta Sintatica corpus was used, which corresponds to approximately 7269 sentences and approximately 16870 morphosyntactically annotated words. The remaining 70% was used for the retraining of the model in the 2nd experiment. Before carrying out the evaluation, it was necessary to make a transformation to the corpus at the level of the Part-of-Speech tags, so that the three experiments could be evaluated and compared equally. The corpus was transformed based on in Table 7.1, in the Appendix section. The final corpus used for the evaluation contains eleven different tags, which are: ADJ (adjective), ADP (adposition), ADV (adverb), CONJ (conjunction), DET (determiner), NOUN (noun), NUM (numeral), PRON (pronoun), PUNCT (punctuation), VERB (verb) and X (other).

Finally, after having the Floresta corpus with a uniform set of tags, we moved on to the evaluation. The experiments were performed, and the output results of each model were saved. In the 1st and 3rd experiments, after the models were executed and the test dataset was annotated with NLTK and SpaCy tags, all data was processed and treated again so that the tags corresponded to the eleven presented above. That is, applying a set of rules to the annotations made by NLTK and SpaCy, the tag names were changed. For example in the case of SpaCy the tag 'AUX' was transformed into 'VERB'. In the 2nd experiment, this additional processing is not necessary since the data is already annotated with the correct tags to evaluate the model. The focus of this preprocessing was to achieve the best accuracy mainly for the NOUN tag which is the most relevant when it comes to personal or sensitive data.

Given the methodology described previously, the tests with the Floresta Sintáctica were performed for the three experiments with a single execution for each model. After the transformation of the results, the accuracy was calculated for each experiment. We present the results obtained in Table 6.1. All results presented in this dissertation were rounded to one decimal place.

Experiments	Accuracy Results
1st Experiment - NLTK model	79.1%
2nd Experiment – NLTK retrained model	83.9%
3rd Experiment – SpaCy model	86.4%

Table 6.1: Accuracy results for Part-of-Speech Tagging Experiments

Analyzing the results, we can see that the 3rd experiment in which the SpaCy library model was used is the one that produces the best results. Through the analysis of the results, we were also able to realize that the retraining of the NLTK model with the corpus in Portuguese produced significant improvements, specifically 4.8%. We can conclude that the worst result of the three experiments was obtained by the model provided in the NLTK library. This result was

predictable since the model was not developed directly for the Portuguese language. With this, we can conclude that the model with the best behavior was the last one in the SpaCy library, which consists of an implementation based on statistical models and the use of multi-task CNN [Arnold2017]. It achieved an accuracy of 86.4% for the transformed Floresta Sintáctica corpus.

Since the objective of these experiments was to achieve the best possible result for the task of Named Entity Recognition, each tag was evaluated individually. The individual analysis enabled us to perceive which classes have more errors and if the errors produced by the morphosyntactic analysis models could harm the NER results. In Figure 6.1, we can see for each one of the experiments the accuracy results for each POS tag.

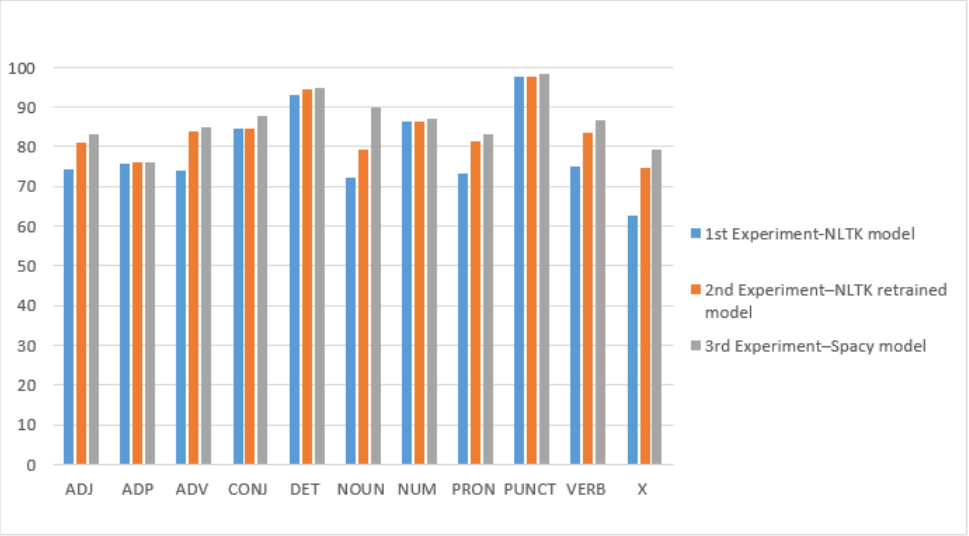

FIGURE 6.1: POS Tagging accuracy organized by experiment

Through this representation, it is possible to have a more detailed comparison for each of the used models. It is possible to conclude that the 3rd experiment achieved superior results in almost all Part-of-Speech tags. It is also possible to conclude that there is an abrupt difference in the ADJ(adjective), ADV(adverb), NOUN(noun), PRON(pronoun), VERB(verb) and X(other) tags, in the 1st experiment, relative to the other two. The improvements in the 3rd experiment for the tags NOUN and VERB are even more notable.

Since the final goal of this work is the recognition of sensitive entities, due to the need to choose one of the POS Tagging models to integrate into the processing chain of the NER component, an additional assessment on the 3rd Experiment was carried out. For this extra evaluation, the Floresta Sintatica corpus was again transformed maintaining the 'prop', 'nprop', and 'vaux' tags, in which the first two correspond to the PROPN(proper noun) tag and the last one to AUX(auxiliary verb). This transformation was carried out to perceive the behavior of the SpaCy model at the level of Proper Names. For this work, the correct identification of this tag is very important since it corresponds to the names of People and Places that we want to identify

as sensitive data. Adding these two new POS tags, the model was again executed, now with thirteen tags. In this evaluation, the SpaCy model used obtained an accuracy of 86.8%, higher than previously achieved. In Figure 6.2, it is possible to see the accuracy for each tag.

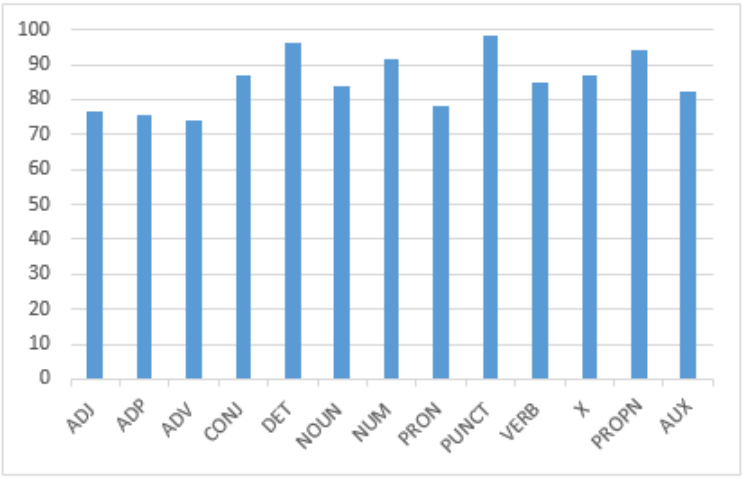

FIGURE 6.2: POS Tagging accuracy evaluation by tag

In the figure, it is possible to see that the percentage achieved on this corpus for the PROPN tag is higher than 90%. These values allow us to have confidence in the results obtained, and this makes it possible to use the model provided by this library as a model for the morphosyntactic analysis component in this work. In addition to the higher accuracy values obtained, this experiment also obtained a better performance. It runs the entire set of tests for the Floresta Sintatica corpus with 7256 phrases in 22.46 seconds.

Comparing the results obtained with the SpaCy library model to other POS Tagging work, both for the Portuguese language and for the same corpus, we were able to draw very satisfactory conclusions. The existing work for the Floresta Sintáctica corpus [Afonso et al.2002] using Maximum-Entropy models and a set of additional features [Aires et al.2000], got accuracy results of 87.73%, very close to the 86.8% achieved by the model used without any study of additional parameters. Other works on the same corpus but using Neural Nets [Marques and Lopes1996], obtained results of 87.8%, differing only 1% relative to the model used. However, for the POS tagging task for the Portuguese language, there are also several works using the PAROLE Copus [Ribeiro2003]. With this corpus, superior results are achieved, for example with the use of HMM was achieved results of 95% [Ribeiro et al.2003]. Therefore we can conclude that this model has a good performance without any kind of parameterization, compared to the current state-of-the-art results, and that it is a viable option in cases where the focus is not on Part-of-Speech Tagging.

6.2 Named Entity Recognition Evaluation and Results

This section presents the evaluation of Named Entities Recognition carried out in this dissertation. All the results obtained by the different approaches used are shown, and the results are compared in order to understand which is the best technique or set of techniques to use in the Named Entity Recognition module of the NER component. All the tests were based on the previous preprocessing steps and the SpaCy model used for the Part-of-Speech tagging annotations.

As we saw in chapter 4, different methodologies were applied to the NER task: Ruled Based Models, Lexicon Based Models, and Machine Learning Models. This division is a consequence of not having any corpus for training and teste that contains all the classes of entities that were worked on in this dissertation, the different techniques were used for different classes of entities. In Table 6.2, we can see a simpler representation of the methodologies used and wich corpus was used for each evaluation.

Methodology	Named Entity Class	Evaluation Corpus
Ruled Based Methods	Personal Identification Numbers Tempo CodigoPostal EnderecoElectronico	DataSense NER Corpus
Lexicon Based Methods	Profissao Med	
	Pessoa Local Tempo	HAREM Golden Collection SIGARRA News Corpus
	Valor	HAREM Golden Collection
Machine Learning Methods		
	Pessoa Local Tempo Organizacao	HAREM Golden Collection SIGARRA News Corpus

Table 6.2: Methodology used by a class of entity and evaluation corpus

All classes discovered through Ruled Based Models as well as the Profissao(Job) and Med(Medical data) classes can only be evaluated by the DataSense NER Corpus. The class Valor(Value), which is present in both Lexicon Based and Machine Learning methods used, was only trained with and analyzed on the corpus HAREM, since the corpus SIGARRA does not have this entity class annotated. All other classes of entities could be evaluated with both the HAREM and SIGARRA corpora, which also allowed a comparison between the results obtained for both corpora. The evaluation of the task of Named Entity Recognition is based on the following metrics, presented in the previous chapter: precision, recall, and f1-score. The evaluation was carried out using the

conlleval[1] script, which requires a file in the CoNLL format, with both the output of the models and the input file. In this file, each line contains a token, the input corpus tag, and the predicted tag. This script accepts files with or without IOB tags, but for this dissertation, the results were evaluated with IOB tags. After each run, we merge the outputs with the input corpus tokens and tags, so that it could be used as input for the evaluation script.

In the following sections of this chapter, we describe in detail the evaluation methods used and the achieved results. The evaluation of the Named Entity Recognition Methods used was divided into four categories that are presented in chapter 4: Lexicon-Based Model, NER Tools, Statistical Models, and Neural Network Model.

6.2.1 Lexicon Based Models Evaluation and Results

The recognition of NE based on lexicons was one of the methods used in this dissertation, mainly for the classes of entities for which there is no annotated corpus. It is the case of the classes Profissao(Job) and Med(Medical data) that can only be evaluated with the corpus Datasense NER. But since these experiments were also performed with lexicons for the classes Pessoa(Person), Local, Tempo(Time) and Valor(Value), they were evaluated using the corpus HAREM and SIGARRA. The NER based on lexicons, as explained in chapter 4, was done by means of comparison and the use of Part-of-Speech tagging analysis. In this case, the total of the two corpora was used for the evaluation, and the evaluation metrics used where precision, recall, and f1-score. Table 6.3 shows the results obtained.

	HAREM Golden Collection	SIGARRA News Corpus
Precision	71.00%	51.32%
Recall	55.60%	74.10%
F1-score	62.36%	60.64%

Table 6.3: Lexicon Based Models results

From Table 6.3, we can not draw proper conclusions since there is no explicit difference between the two corpora. However, in Figure 6.3 we can see the detailed f1-score results for each class of entities. This allows us to conclude that the classes of entities PESSOA and LOCAL cannot be used to achieve satisfactory results. This is due to the inexistence of many names and places in the lexicons used and the fact that there is a great deal of confusion between the two entities. On the other hand, the class TEMPO had f1-score results higher than 90% for both cases.

[1] `https://github.com/sighsmile/conlleval`

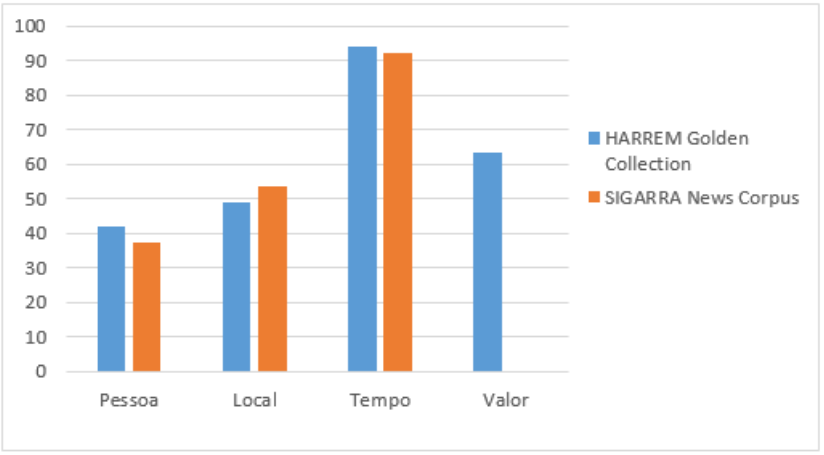

FIGURE 6.3: Lexicon Based Models f1-score results by entity class, by the corpus

These results serve as a baseline for comparison with the other implemented methodologies. Comparing the results obtained to other works with the same lexicon-based approach and with the same corpus, we can draw some conclusions. For the HAREM Golden Collection [Mota and Santos2008] the results obtained for class entities PESSOA and LOCAL are very close to those obtained with the REMMA system [Ferreira et al.2008], but when comparing the same system for TEMPO and VALOR classes [Santos et al.2006], we get results with an f1-score 20% higher on average. Another system with a similar approach that used the same corpus is Rembrandt [Cardoso2012], this system got its results for the TEMPO class with an f1-score of 33.07%, much lower than ours. For the remaining classes of entities, the results are similar, except for the PESSOA class where the Rembrandt system achieves results of 47.40%, slightly higher than ours. In terms of the use of lexicon-based approaches, we were able to outperform the existing state-of-the-art results for the class TEMPO, maintaining the results for the remaining classes of entities. For the SIGARRA corpus, there is only one work [Pires2017], presenting a proposal based on NER tools, which in this case achieves higher average results than those obtained with lexical-based methodologies.

6.2.2 NER Tools Evaluation and Results

The first models evaluated were the models of Entities Recognition using the - SpaCy and Stanford NER tools. These tools were trained with the HAREM and SIGARRA corpora and later evaluated. For both tools, the default settings of the models were used, and no additional hyperparameters were chosen. In order to ensure the robustness of the evaluation metrics, we used 10-fold cross-validation and calculated the mean precision, mean recall, and the macro-average of the f1-score. The corpora were divided into 10 subsets with sentences of equal size, where one subset was used for testing and the rest for training. Finally, after training the models,

and running on the test set, the output results were evaluated using the CoNLLL script, and the mean for each level was calculated, resulting in an average for each tool and each entity class.

The results of 10-fold cross-validation with the HAREM and SIGARRA corpus for the SpaCy and Stanford NER tools are presented in Tables 6.4 and 6.5. There is a clear distinction between the tools and also between the corpus used.

	HAREM Golden Collection	SIGARRA News Corpus
Precision	57.26%	89.01%
Recall	52.89%	83.97%
F1-score	54.99%	86.41%

Table 6.4: Stanford NER tool evaluation results

	HAREM Golden Collection	SIGARRA News Corpus
Precision	50.13%	85.10%
Recall	41.70%	76.27%
F1-score	45.53%	80.44%

Table 6.5: SpaCy tool evaluation results

The Stanford NER model obtained the best results, with the f1-score values being 7.72% higher on average, compared to the SpaCy model. However, the training and testing time in the case of Stanford NER is much longer. To complete all the training and 10-fold cross-evaluation, the Stanford NER model took 87.06 minutes for HAREM and 141.75 minutes for SIGARRA. In the case of SpaCy, this time was much shorter taking 6.84 minutes to complete the processing for HAREM and 11.14 minutes for SIGARRA. Despite the significant differences between both tools, it is also possible to notice a large difference between the results obtained for the two corpora: the HAREM Golden Collection resulted in clearly lower results compared to the SIGARRA News Corpus. This difference is due to the fact that HAREM is an outdated collection and smaller than SIGARRA, with fewer samples from each entity.

To better understand the behavior of each tool and each corpus, we analyzed each entity individually. Figure 6.4 shows the graph of the f1-score obtained by tool and by entities. In the case of the entity VALOR we can see that it does not exist in the SIGARRA corpus, so the result has only presented for HAREM. It is also possible to see that the results for the classes of entities PESSOA, TEMPO and ORGANIZACAO are above the average obtained for the other entities. We can also conclude that the results obtained for the class VALOR are much lower than the results achieved with the lexicon-based model. This value was already expected to take into account the number of samples in the training set, as we saw in chapter 5.

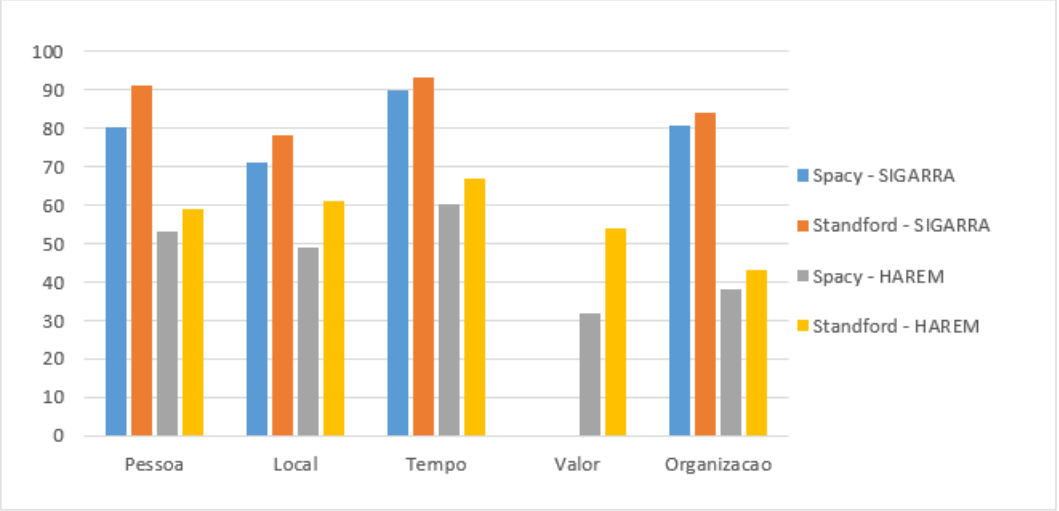

FIGURE 6.4: Named Entities Recognition tools f1-score results by class, by corpus

Comparing the work done with the existing state-of-the-art for NER tools [Atdağ and Labatut2013], we were able to conclude that SpaCy and Stanford achieve better results than other existing tools. For the HAREM Golden Collection, the existing results are 30.97% with the NLTK library model and 53.63% with the OpenNLP library [Pires et al.], which results are lower than the f1-score results of 54.99% that we achieved with the Stanford NER library model. Comparing the SIGARRA News Corpus with the same tools, SpaCy and Stanford NER, we realize that by studying the hyperparameters of the models it is possible to achieve a small improvement in the models [Pires2017]. The improvements in both tools were about 1% for the same corpora. Finally, we concluded that the results obtained with NER tools are superior to the lexical-based test, except for the class VALOR, which presents superior results in the previous tests.

6.2.3 NER with Statistical Models Evaluation and Results

In chapter 4, the implementation of two statistical models was explained - Conditional Random Fields (CRF) and Random Forest. In this chapter, we present the results obtained with these models. As before, the corpus HAREM and SIGARRA were used for the training and testing of both models. The set of features used was explained in previous chapters, and both models were trained with the same set of features, in order to make a more realistic comparison between the two models. These tests were performed in the same format as the previous ones with input and output data annotated in CONLL, IOB tagging format. To perform the evaluation, we used 5-fold cross-validation as an input parameter for the classifier, that is, we divided the corpus into 5 subsets, and the model was trained and tested on them. Some models, such as decision trees [Galathiya et al.2012] and in this case, Random Forest, are often able to obtain high accuracy values in training data but perform much worse in new data. So we train on one subset and

test on the other and repeat for each subset so that the classifier scores correctly on average and the performance estimate is not overly optimistic. In Tables 6.6 and 6.7, we can see the results obtained by the models for the two corpora tested.

	HAREM Golden Collection	SIGARRA News Corpus
Precision	63.48%	73.60%
Recall	44.35%	59.01%
F1-score	52.21%	65.50%

Table 6.6: Conditional Random Fields evaluation results

	HAREM Golden Collection	SIGARRA News Corpus
Precision	49.87%	65.8%
Recall	36.12%	50.1%
F1-score	41.89%	56.89%

Table 6.7: Random Forest evaluation results

By analyzing the results, we concluded that once again, the results obtained for the HAREM Golden Collection corpus are lower and that this corpus is not enough to train a model and have satisfactory results. It is also possible to see that the training results of the statistical models are lower when compared to the NER tools used. Still, we conclude that the Conditional Random Fields model achieves better results when compared to the Random Forest. Even taking into account the better results achieved with the NER tools, the individual analysis of each class of entities was performed in order to try to understand the behavior of the models for each entity. In Figure 6.6, we present the f1-score values for each class individually. We concluded that, despite the average score, the models had a very poor performance. The f1-score results for some of the entity classes were very low. Therefore, the models are basically memorizing words and tags, which is not enough. The context information behind each word needs to be provided to the model as well so that the predictions are more accurate.

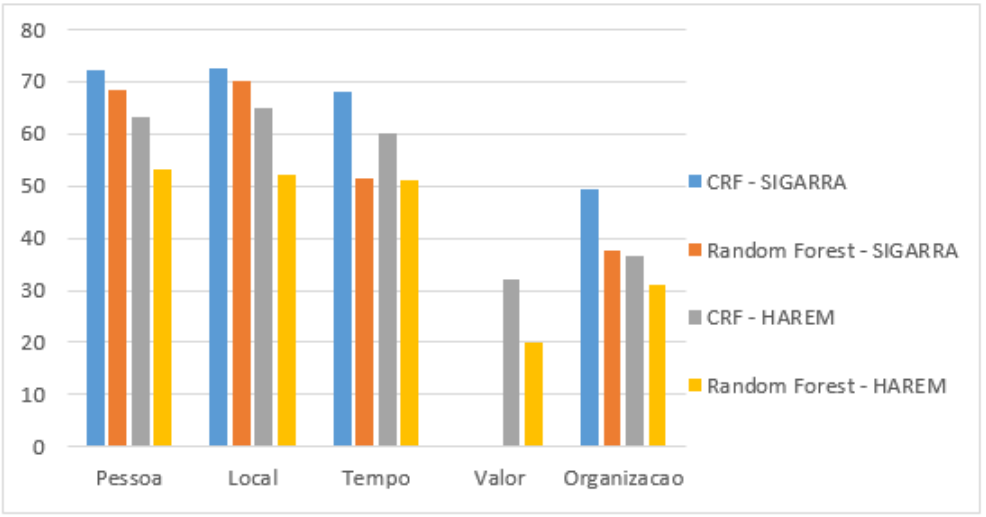

FIGURE 6.5: NER Statistical Models f1-score results by class, by corpus

The results obtained with the CRF model and the HAREM Golden Collection were compared to the results of two NER systems, which conducted experiments under the same conditions of this study. The first NERP-CRF system [Amaral et al.2017] obtained lower results but using the total set of categories available in the corpus. The results of f1-score were 51.57%, 5% lower than the results we obtained in this experiment. Another system that uses the same CRF model is the CRF+LG [Pirovani2019], this system with the use of the CRF model obtained results of 65.33%, higher than the results we obtained. This is due to the use of gazetteers that support the model classification. On the other hand, when comparing the results obtained by both models with the same CRF and Random Forest models applied to the English language, the results obtained with the corpora HAREM and SIGARRA have an f1-score 10% lower on average [Dalianis and Boström2012].

6.2.4 NER with Neural Network Model Evaluation and Results

As we exposed in chapter 4, the final experiment was the implementation of a Bidirectional-LSTM. This model, unlike the others, was trained and tested only with the SIGARRA News Corpus, because of the smaller size of the HAREM Golden Collection. The corpus was previously divided into three sets: training, development, and testing. The embed function that creates word-level embeddings were used to generate an embedding representation for each word from the text. The parameters used for training the model were the ones used in the article for the CoNLL-2003 dataset[Chiu and Nichols2016]: 80 epochs, 0.68 dropouts, 275 LSTM state size and 3 convolutional widths. After training the model and generates the final outputs through the softmax layer, in IOB tagging format, it was possible to perform the evaluation of the model presented in Table 6.8.

Metrics	Results
Precision	81.13%
Recall	75.61%
F1-score	78.27%

TABLE 6.8: Neural Network evaluation results

In Table 6.8, we can see that the results obtained by this model are higher than those obtained by the statistical models. This model obtained an f1-score of 78.25%, about 13% higher compared to the best statistical model implemented. However, as in the other tests, the focus is on the individual analysis of each class of entities. The f1-score results obtained for each class are presented in Figure 6.6.

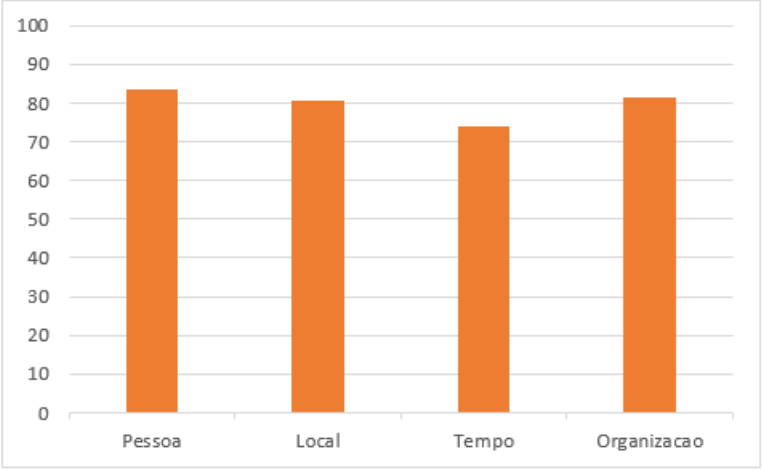

FIGURE 6.6: NER Neural Network f1-score results by class

From Figure 6.6, we can conclude that although we have a good average score, the model performed worse in some classes of entities, specifically in the class VALOR. The result of the f1-score for this class is quite low compared to the other models, although in general, the model achieves good results for the other classes of entities.

In an attempt to compare the state-of-the-art models to the current one tested, we did not find an approach for the same corpus, but we were able to understand that the same Bi-LSTM model applied to the English language has results 12% higher on average [Li et al.2019], [Chiu and Nichols2016]. A similar approach with an LSMT-CRF model [Castro et al.2018], for a corpus in Portuguese, presents f1-score results of 76.03%, lower than the 78.27% we achieved with this model. By analyzing similar models and by improving results with the statistical models tested, we were able to understand that with a larger corpus the results with this type of model greatly improve [Sundermeyer et al.2012].

6.2.5 Summary

In Figures 6.7 and 6.8, it is possible to see for each corpus tested, the comparison results by entity class for each approach. The results achieved with the SIGARRA News Corpus were significantly better in all experiments than the results achieved with the HAREM Golden Collection. This was due to the fact that the SIGARRA corpus is larger than the HAREM, which improves the training process. In addition, the SIGARRA News Corpus contains many documents with the same structure, making it easier to learn. In general, the best global results for HAREM were achieved with the Stanford NER tool and the CRF model. As for SIGARRA corpus, the best results are obtained with the Stanford tool, and with the Bidirectional-LSTM model. However, for some specific categories such as VALOR and TEMPO, the best results were achieved by the lexical-based approach. Overall, the Stanford NER tool was the one with the best performance in the overall set of entity categories, and the differences presented for the Bi-LSTM model for the SIGARRA corpus are not very large.

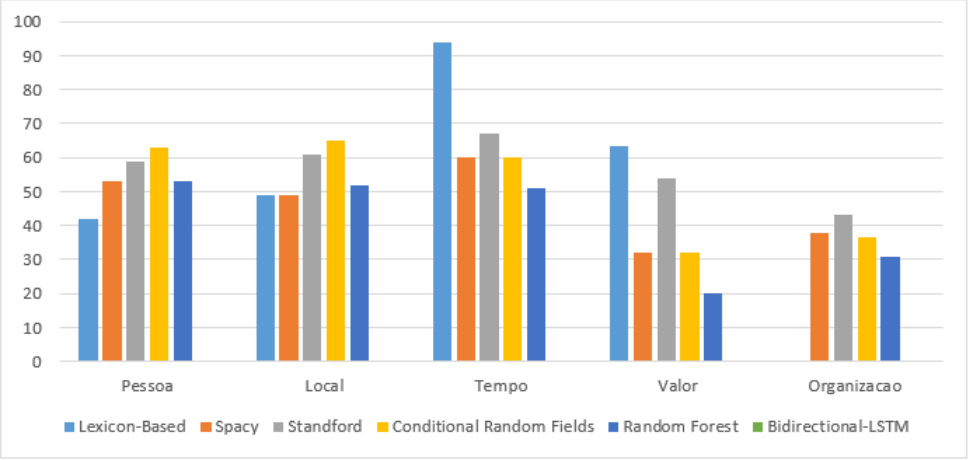

FIGURE 6.7: Comparison of the tests performed for HAREM Golden Collection

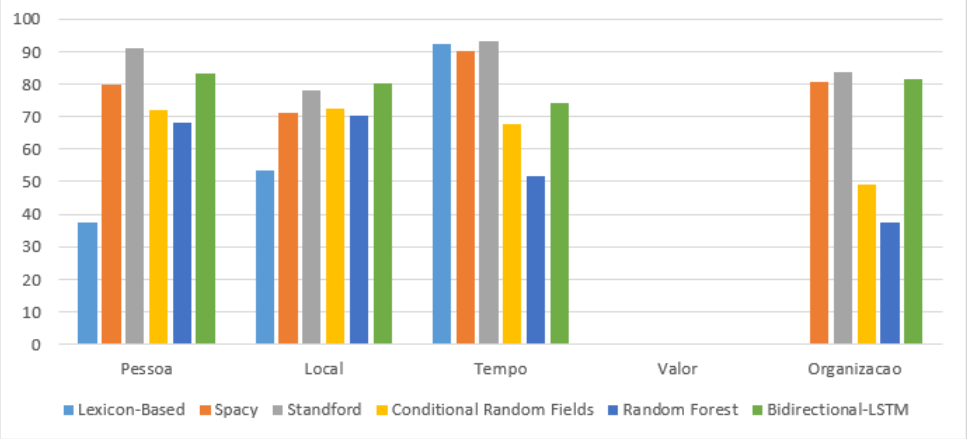

FIGURE 6.8: Comparison of the tests performed for SIGARRA News Corpus

6.3 Named Entity Recognition Component Validation

The last evaluation performed for this dissertation consisted of the validation of all software development, and it was carried out along with all the partners of the DataSense Project, focusing on the Named Entity Recognition Component. The main goal of this evaluation was to assess the performance of the NER component and to assess the quality of the recognition of sensitive entities. The evaluation was carried out on a machine with characteristics identical to those used to develop the models, and it used the DataSense NER Corpus, previously annotated.

Before performing the evaluation tests, there was a set of decisions that were made in order to understand which models would be part of the NER Component (Figure 4.1). At the level of Segmentation and Tokenization tasks, we used the methodologies described in chapter 4. For the Morphosyntactic Analysis component, we used the SpaCy library model, since it was the one that obtained the best results and was used for the training of all NER models. For the Named Entity Recognition Module, considering the results of f1-score and performance for each of the tests before, a set of different tasks was chosen in order to cover all classes of entities required for the DataSense Project. In Table 6.9, we can see the named entity recognition methods chosen for each class of entities.

Entity Classes	Used Models
Personal Identifications Numbers	Ruled Based Models
CodigoPostal	
EnderecoElectronico	
Profissao	Lexicon Based Models
Med	
Tempo	
Valor	
Pessoa	Machine Learning Models - Bidirectional-LSTM
Local	
Organizacao	

Table 6.9: Method of Named Entity Recognition used by a class of entity for NER Component

The choice was made based on the best results of f1-score and performance of all testes that we made before. For this reason, the Bi-LSTM model was used on the NER Component for the classes PESSOA, LOCAL and ORGANIZACAO, although the Stanford NER model had better results, the performance of this model was very low to be used in a real project. With this set of models, the evaluation of DataSense Ner Corpus was performed, and the results are present in Table 6.10.

Metrics	Results
Precision	87.60%
Recall	79.02%
F1-score	83.01%

TABLE 6.10: DataSense NER Corpus evaluation results

The NER Component had an f1-score of 83.01% in the DataSense NER Corpus and took 1716 seconds to complete the processing of all 78 documents. In Figure 6.9, we can see the detailed analysis of the results of the f1-score for each class of entities. From the figure, we can conclude that there are some classes such as Profissao(Job) and Med(Medical data), which have much lower results when compared to the other classes, although in general, all classes presented good results.

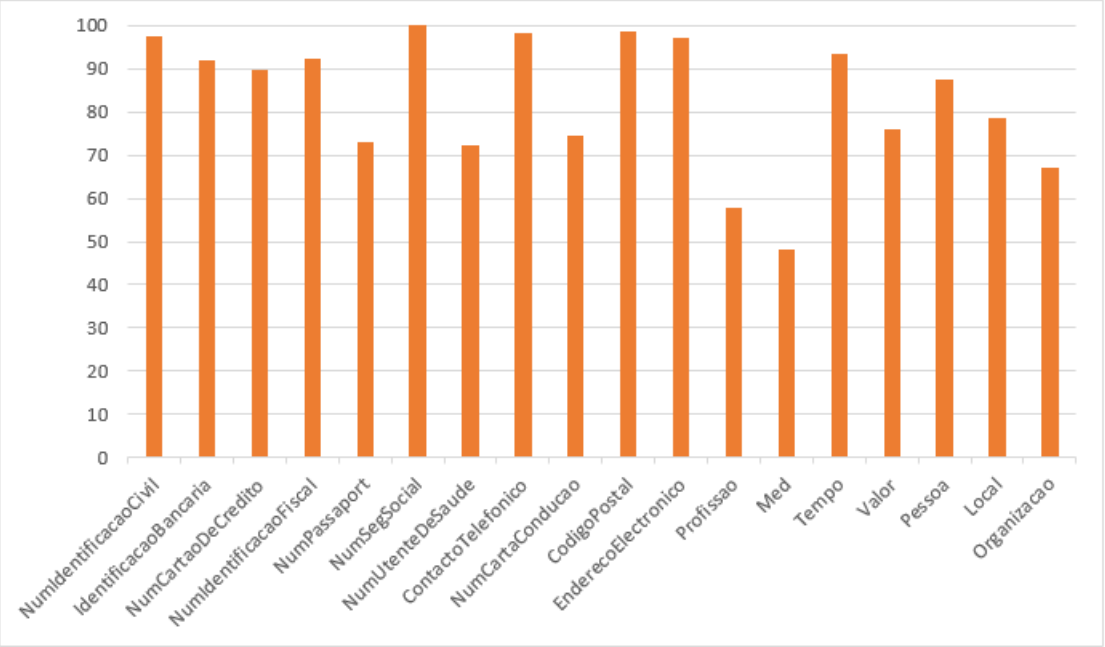

FIGURE 6.9: NER Component f1-score results by class

Chapter 7

Conclusion and Future Work

This chapter presents the conclusions regarding the results achieved with the study for Named Entities Recognition for the Portuguese language. We present the achieved goals, as well as the main contributions achieved by this dissertation. In addition, we address some possible future work, namely the further study of the processing of the Portuguese language and the implementation of the remaining modules of the DataSense Project.

7.1 Contributions

Through this dissertation, it was possible to make various contributions to the scientific community and to several projects, particularly in the Named Entity Recognition area, and sensitive data topic. The most significant contributions were:

- A detailed study of entity recognition for the Portuguese language with different techniques and tools, based on the HAREM Golden Collection corpus and SIGARRA News Corpus. Different approaches were studied and compared: Ruled-Based Models, Lexicon-Based Model, Machine Learning Models, and Deep Learning Models.
- The study of natural language processing tools, at the level of Part-of-Speech Tagging annotation for Portuguese.
- The application of entities' recognition to a larger number of categories, such as, to all sensitive data covered by the GDPR European directive. All categories of personal and sensitive data were covered with different NER techniques/tools.
- The recognition of sensitive data for unstructured text information and in textual documents of companies' repositories.

- A functional prototype of the NER module, using the DataSense Project already implemented in the Portuguese market and the SocialOpinion research project (Appendix 7.1).
- Two publications in conferences regarding the NER module inserted in the DataSense project [Mariana Dias2019a] and the NLP methodologies used for the corpus HAREM [Mariana Dias2019b].

7.2 Conclusions

The main goal of this dissertation was to develop a functional prototype of named entity recognition for the Portuguese language. The focus of the developed prototype was the recognition of sensitive data in unstructured texts, according to all categories covered by GDPR. This prototype was validated, under the Portugal2020 DataSense Project, through the tests of efficiency and performance. This validation was carried out with the project stakeholders, and the developed prototype had a f1-score of 83.01% in the NER task. Thus, the prototype is currently being used in the project in production. In addition, the work developed was also be integrated into other projects as it happened with SocialOpinion Project, which demonstrates an essential contribution in this area.

We studied the Part-of-Speech Tagging task, which helped with the entities' recognition. The experiments were carried out using Floresta Sintáctica Corpus and applying natural language processing tools. This study allowing classifying our training and test data before proceeding to the NER task. This morphosyntactic analysis allowed us to help the hand-coded techniques used for the recognition of entities. It also allowed considering an additional feature in all trained models, always considering the POS tag, important mainly for the entities Person, Place, and Organization.

The development and testing of the work in Named Entity Recognition were done using three different corpora: HAREM Golden Collection, SIGARRA News Corpus and DataSense NER creating exclusively for validation of the work performed. For this task, different approaches were used, and several experiments were done in order to achieve the best results for each entity class. A rule-based model and morphological analysis were implemented, achieving the best results for entities with well-defined formats and that follow strict rules. Models based on lexicons were also implemented for a reduced set of entities, achieving a f1-score result of 62.36% for HAREM and 60.64% for SIGARRA. Although the global results when using lexicon-based models are lower than the current state-of-the-art, for TEMPO and VALOR entities the results were higher than those achieved with other methodologies, and they were a way of solving the PROFISSAO and MED entities for which there was no labeled data in Portuguese but were necessary for the proposed prototype.

For the remaining classes of entities, different experiments were carried out, including the performance study of language processing tools, the implementation of statistical machine learning models and the implementation of a Bidirectional-LSTM neural network. Firstly, the SpaCy and Stanford Core NLP tools were tested, and the best f1-score results were 86.41% for the SIGARRA corpus with the Stanford model. Demonstrating that multi-language tools also achieve good performance when used in the Portuguese language. To have greater control in data training and in the construction of the models used, we implemented three models. Two statistical models, a Conditional Random Fields and a Random Forest, allowed us to conclude that the CRF model achieved better results than the second model, but not better than the SpaCy and Stanford NLP tools. With these two models, we were also able to understand that the HAREM corpus is not enough for training more complex models due to its size and the reduced number of annotated entities. Finally, the third model implemented was the Bi-LSTM, which ended up being used in the prototype that resulted from this dissertation. The Bi-LSTM model, although it did not achieve the highest percentage of f1-score, was the model that obtained the best global results and performance. Since Stanford's NLP tool model, which was the one that obtained the best f1-score results in NER, had a very high processing time, and this aspect is important according to the ultimate goal of having a functional prototype. Furthermore, the Bi-LSTM model allows for greater freedom to study new parameters and improvements in the future.

The work was developed made possible through a hybrid approach, and the use of different methodologies covered all sets of entities that represent sensitive data. We also conclude that it is possible for the Portuguese language to have valid results for named entities recognition tasks, and it can be used in real scenarios with a value in the Portuguese market.

7.3 Future Work

The work carried out in this dissertation presented promising results, and showed that it is possible to integrate natural language processing techniques into real scenarios to solve problems of recognition of sensitive data. Through our experiments, we realized that it is possible to obtain good results with a corpus in the Portuguese language. However, the number of existing resources should be increased, namely the DataSense NER corpus should continue to grow in the future in order to train models with this corpus and provide even better results, mainly on the topic of sensitive data. In the future, it would be interesting to carry out an in-depth study of the parameters and hyperparameters of the used models, in order to understand where the errors in the trained models are occurring and to study how these errors can be solved. We believe future work should include a more in-depth study of the models and tools for named entities recognition. In addition, future work will involve the implementation of the remaining modules of the DataSense Project, in the context of sensitive data, Coreference Resolution, and

Evaluation and Human Feedback, which is based on the work of this dissertation. In other words, in the very near future, the goal will be to discover relationships between entities and to aggregate the sensitive entities present in the documents.

Bibliography

[Adèr2008] Adèr, H. J. (2008). *Advising on research methods: A consultant's companion*. Johannes van Kessel Publishing.

[Afonso et al.2002] Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá (c) tica: a treebank for portuguese. In *quot; In Manuel González Rodrigues; Carmen Paz Suarez Araujo (ed) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002) Paris: ELRA.* ELRA.

[Agarwal et al.2011] Agarwal, M., Goutam, R., Jain, A., Kesidi, S. R., Kosaraju, P., Muktyar, S., Ambati, B., and Sangal, R. (2011). Comparative analysis of the performance of crf, hmm and maxent for part-of-speech tagging, chunking and named entity recognition for a morphologically rich language. *Proc. of Pacific Association For Computational Lingustics*, pages 3–6.

[Aires et al.2000] Aires, R. V. X., Aluísio, S. M., Kuhn, D. C., Andreeta, M. L., and Oliveira Jr, O. N. (2000). Combining multiple classifiers to improve part of speech tagging: A case study for brazilian portuguese. In *the Proceedings of the Brazilian AI Symposium (SBIA'2000)*, pages 20–22.

[Albrecht2016] Albrecht, J. P. (2016). How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287.

[Amaral et al.2008] Amaral, C., Figueira, H., Mendes, A., Mendes, P., Pinto, C., and Veiga, T. (2008). Adaptação do sistema de reconhecimento de entidades mencionadas da priberam ao harem. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca*.

[Amaral and Vieira2013] Amaral, D. O. and Vieira, R. (2013). O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa (named entity recognition with conditional random fields for the portuguese language)[in portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

[Amaral and Vieira2014] Amaral, D. O. F. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

[Amaral et al.2013] Amaral, D. O. F. d. et al. (2013). O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa.

[Amaral et al.2017] Amaral, D. O. F. d. et al. (2017). Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras.

[Appelt et al.1993] Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., and Tyson, M. (1993). Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178.

[Arnold2017] Arnold, T. (2017). A tidy data model for natural language processing using cleannlp. *arXiv preprint arXiv:1703.09570*.

[Atdağ and Labatut2013] Atdağ, S. and Labatut, V. (2013). A comparison of named entity recognition tools applied to biographical texts. In *2nd International conference on systems and computer science*, pages 228–233. IEEE.

[Athiwaratkun et al.2018] Athiwaratkun, B., Wilson, A. G., and Anandkumar, A. (2018). Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*.

[Baevski et al.2019] Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

[Baldwin et al.2015] Baldwin, T., de Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

[Bellot et al.2003] Bellot, P., Crestan, E., El-Bèze, M., Gillard, L., and de Loupy, C. (2003). Coupling named entity recognition, vector-space model and knowledge bases for trec 11 question answering track. *NIST SPECIAL PUBLICATION SP*, (251):398–406.

[Bikel et al.1998] Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1998). Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

[Bikel et al.1999] Bikel, D. M., Schwartz, R., and Weischedel, R. M. (1999). An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231.

[Borthwick et al.1998] Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*.

[Branco and Silva2004] Branco, A. and Silva, J. R. (2004). Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In *LREC*.

[Brill and Mooney1997] Brill, E. and Mooney, R. J. (1997). An overview of empirical natural language processing. *AI magazine*, 18(4):13–13.

[Cardie1997] Cardie, C. (1997). Empirical methods in information extraction. *AI magazine*, 18(4):65–65.

[Cardoso2008] Cardoso, N. (2008). Rembrandt-reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. *quot; Encontro do Segundo HAREM (Universidade de Aveiro Portugal 7 de Setembro de 2008)*.

[Cardoso2012] Cardoso, N. (2012). Rembrandt-a named-entity recognition framework. In *quot; In Nicoletta Calzolari; Khalid Choukri; Thierry Declerck; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis (ed) Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)(Istambul 23-25 de Maio de 2012; Maio de 2012)*.

[Carreras et al.2002] Carreras, X., Màrquez, L., and Padró, L. (2002). Named entity extraction using adaboost. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

[Carvalho2012] Carvalho, W. S. (2012). *Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina*. PhD thesis, Universidade de Sã Paulo.

[Castro et al.2018] Castro, P. V. Q., da Silva, N. F. F., and da Silva Soares, A. (2018). Portuguese named entity recognition using lstm-crf. In *International Conference on Computational Processing of the Portuguese Language*, pages 83–92. Springer.

[Chen et al.2014] Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, 19(2):171–209.

[Cheung2008] Cheung, S. (2008). Proof of hammersley-clifford theorem. *Unpublished, February*.

[Chiu and Nichols2016] Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

[Ciravegna2001] Ciravegna, F. (2001). 2, an adaptive algorithm for information extraction from web-related texts. In *In Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*.

[Clough2005] Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 25–30. ACM.

[Colic and Rinaldi2019] Colic, N. and Rinaldi, F. (2019). Improving spacy dependency annotation and pos tagging web service using independent ner services. *Genomics & Informatics*, 17(2).

[Collins2002] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

[Collobert and Weston2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

[Cui et al.2018] Cui, Z., Ke, R., and Wang, Y. (2018). Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*.

[Dale et al.2000] Dale, R., Moisl, H., and Somers, H. (2000). *Handbook of natural language processing*. CRC Press.

[Dalianis and Boström2012] Dalianis, H. and Boström, H. (2012). Releasing a swedish clinical corpus after removing all words–de-identification experiments with conditional random fields and random forests. In *the Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC*, pages 45–48.

[de Freitas et al.2005] de Freitas, M. C., Uzeda-Garrão, M., Oliveira, C., dos Santos, C. N., and Silveira, M. C. (2005). A anotaçao de um corpus para o aprendizado supervisionado de um modelo de sn. In *Proceedings of the III TIL/XXV Congresso da SBC*.

[Derczynski et al.2017] Derczynski, L., Nichols, E., van Erp, M., and Limsopatham, N. (2017). Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

[Doddington et al.2004] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, page 1. Lisbon.

[Fernandes2018] Fernandes, I. A. D. (2018). A deep learning approach to named entity recognition in portuguese texts.

[Ferrández et al.2007] Ferrández, Ó., Kozareva, Z., Toral, A., Muñoz, R., and Montoyo, A. (2007). Tackling harem's portuguese named entity recognition task with spanish resources. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, capítulo*, 11:137–144.

[Ferreira et al.2008] Ferreira, L., Teixeira, A., and Cunha, J. P. S. (2008). Remma-reconhecimento de entidades mencionadas do medalert. *Desafios na avaliaçao conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, Aveiro, Portugal*, 7.

[Finkel et al.2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

[Florian et al.2003] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada.

[Freitas et al.2010] Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., and Santos, D. (2010). 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.

[Freitas et al.2008] Freitas, C., Rocha, P., and Bick, E. (2008). Um mundo novo na floresta sintá (c) tica–o treebank do português. *Calidoscópio*, 6(3):142–148.

[Fresko et al.2005] Fresko, M., Rosenfeld, B., and Feldman, R. (2005). A hybrid approach to ner by memm and manual rules. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 361–362. ACM.

[Galathiya et al.2012] Galathiya, A., Ganatra, A., and Bhensdadia, C. (2012). Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning. *International Journal of Computer Science and Information Technologies*, 3(2):3427–3431.

[Gattani et al.2013] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. (2013). Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137.

[Ghaddar and Langlais2017] Ghaddar, A. and Langlais, P. (2017). Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.

[Graves and Schmidhuber2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

[Grishman1995] Grishman, R. (1995). The nyu system for muc-6 or where's the syntax? Technical report, NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE.

[Grishman and Sundheim1995] Grishman, R. and Sundheim, B. (1995). Design of the muc-6 evaluation. In *Proceedings of the 6th conference on Message understanding*, pages 1–11. Association for Computational Linguistics.

[Grishman and Sundheim1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

[Hagège et al.2008] Hagège, C., Baptista, J., and Mamede, N. (2008). Reconhecimento de entidades mencionadas com o xip: Uma colaboração entre a xerox e o l2f do inesc-id lisboa. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca*.

[Heuss et al.2014] Heuss, T., Humm, B., Henninger, C., and Rippl, T. (2014). A comparison of ner tools wrt a domain-specific vocabulary. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 100–107. ACM.

[Hochreiter and Schmidhuber1997] Hochreiter, S. and Schmidhuber, J. (1997). Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

[Jiang et al.2007] Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). Mipred: classification of real and pseudo microrna precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(suppl_2):W339–W344.

[Jiang et al.2016] Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27.

[Jin2015] Jin, N. (2015). Ncsu-sas-ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 87–92.

[Johnson2019] Johnson, J. A. (2019). Twenty-first century insurance-cyber insurance. *The Computer & Internet Lawyer*, 36(8).

[Kannan and Gurusamy2014] Kannan, S. and Gurusamy, V. (2014). Preprocessing techniques for text mining.

[Kazama and Torisawa2007] Kazama, J. and Torisawa, K. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 698–707.

[Kazama and Torisawa2008] Kazama, J. and Torisawa, K. (2008). Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *proceedings of ACL-08: HLT*, pages 407–415.

[Klein et al.2003] Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. (2003). Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 180–183. Association for Computational Linguistics.

[Korba et al.2008] Korba, L., Wang, Y., Geng, L., Song, R., Yee, G., Patrick, A. S., Buffett, S., Liu, H., and You, Y. (2008). Private data discovery for privacy compliance in collaborative environments. In *International Conference on Cooperative Design, Visualization and Engineering*, pages 142–150. Springer.

[Lafferty et al.2001] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Lample et al.2016] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

[LeCun et al.2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Li et al.2019] Li, P.-H., Fu, T.-J., and Ma, W.-Y. (2019). Remedying bi-lstm-cnn deficiency in modeling cross-context for ner. *arXiv preprint arXiv:1908.11046*.

[Liaw et al.2002] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[Limketkai et al.2007] Limketkai, B., Fox, D., and Liao, L. (2007). Crf-filters: Discriminative particle filters for sequential state estimation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3142–3147. IEEE.

[Lin and Wu2009] Lin, D. and Wu, X. (2009). Phrase clustering for discriminative learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1030–1038. Association for Computational Linguistics.

[Loper and Bird2002] Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

[Magge et al.2018] Magge, A., Scotch, M., and Gonzalez-Hernandez, G. (2018). Clinical ner and relation extraction using bi-char-lstms and random forest classifiers. In *International Workshop on Medication and Adverse Drug Event Detection*, pages 25–30.

[Malecha and Smith2010] Malecha, G. and Smith, I. (2010). Maximum entropy part-of-speech tagging in nltk.

[Mamede et al.2016] Mamede, N., Baptista, J., and Dias, F. (2016). Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294. IEEE.

[Mariana Dias2019a] Mariana Dias, Rui Maia, F. J. R. R. . M. A. (2019a). Data sense platform. *In IASTEM - 586th International Conference on Science Technology and Management (ICSTM).*

[Mariana Dias2019b] Mariana Dias, João C. Ferreira, R. M. P. S. R. R. (2019b). Privacy in text documents. *33rd IBIMA Conference,Granada, Spain.*

[Marques and Lopes1996] Marques, N. C. and Lopes, G. P. (1996). Using neural nets for portuguese part-of-speech tagging. In *In Proc. Of the 5th CSNLP Conference*. Citeseer.

[Marrero et al.2009] Marrero, M., Sánchez-Cuadrado, S., Lara, J. M., and Andreadakis, G. (2009). Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

[McCallum and Li2003] McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

[McDonald1993] McDonald, D. (1993). Internal and external evidence in the identification and semantic categorization of proper names. In *Acquisition of Lexical Knowledge from Text.*

[Merchant et al.1996] Merchant, R., Okurowski, M. E., and Chinchor, N. (1996). The multilingual entity task (met) overview. Technical report, DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD.

[Mikheev et al.1999] Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

[Milidiú et al.2007] Milidiú, R. L., Duarte, J. C., and Cavalcante, R. (2007). Machine learning algorithms for portuguese named entity recognition. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 11(36):67–75.

[Mota2008] Mota, C. (2008). R3m, uma participação minimalista no segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008.*

[Mota and Santos2008] Mota, C. and Santos, D. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

[Mota et al.2007] Mota, C., Santos, D., and Ranchhod, E. (2007). Avaliação de reconhecimento de entidades mencionadas: princípio de arem. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pages 161–175.

[Nadeau and Sekine2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Neter et al.1996] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

[Nie and Fan2006] Nie, Y. and Fan, Y. (2006). Arriving-on-time problem: discrete algorithm that ensures convergence. *Transportation Research Record*, 1964(1):193–200.

[Nothman et al.2013] Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

[Olah2015] Olah, C. (2015). Understanding lstm networks.

[Olney et al.2016] Olney, W., Hill, E., Thurber, C., and Lemma, B. (2016). Part of speech tagging java method names. In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 483–487. IEEE.

[Peters et al.2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[Pinto et al.2003] Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM.

[Pires et al.] Pires, A., Devezas, J., and Nunes, S. Benchmarking named entity recognition tools for portuguese. *Proceedings of the Ninth INForum: Simpósio de Informática*, pages 111–121.

[Pires2017] Pires, A. R. O. (2017). Named entity extraction from portuguese web text.

[Pirovani2019] Pirovani, J. P. C. (2019). Crf+ lg: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português.

[Ponomareva et al.2007] Ponomareva, N., Rosso, P., Pla, F., and Molina, A. (2007). Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In *Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP*, volume 479, page 483.

[Rademaker et al.2017] Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.

[Ralph1995] Ralph, G. (1995). The new york university system muc-6 or where's the syntax. In *Proceedings of the Sixth Message Understanding Conference*, pages 167–175.

[Ramshaw and Marcus1999] Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

[Ratinov and Roth2009] Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.

[Rau1991] Rau, L. F. (1991). Extracting company names from text. In *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, volume 1, pages 29–32. IEEE.

[Ribeiro2003] Ribeiro, R. (2003). *Anotação morfossintàctica desambiguada do português*. PhD thesis, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[Ribeiro et al.2003] Ribeiro, R., Oliveira, L. C., and Trancoso, I. (2003). Using morphossyntactic information in tts systems: Comparing strategies for european portuguese. In Nuno J. Mamede, Isabel Trancoso, J. B. M. d. G. V. N., editor, *Computational Processing of the Portuguese Language*, pages 143–150, Faro. Springer.

[Ritter et al.2011] Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

[Rocha et al.2016] Rocha, C., Jorge, A., Sionara, R., Brito, P., Pimenta, C., and Rezende, S. (2016). Pampo: using pattern matching and pos-tagging for effective named entities recognition in portuguese. *arXiv preprint arXiv:1612.09535*.

[Sang and De Meulder2003] Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

[Santorini1990] Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. University of Pennsylvania, School of Engineering and Applied Science ....

[Santos and Cardoso2006] Santos, D. and Cardoso, N. (2006). A golden resource for named entity recognition in portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 69–79. Springer.

[Santos and Cardoso2007] Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

[Santos et al.] Santos, D., Seco, N., Cardoso, N., and Vilela, R. An advanced ner evaluation contest for portuguese.

[Santos et al.2006] Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006).*

[Sarmento2006] Sarmento, L. (2006). Siemês–a named-entity recognizer for portuguese relying on similarity rules. In *International Workshop on Computational Processing of the Portuguese Language*, pages 90–99. Springer.

[Schuster and Paliwal1997] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

[Shishtla et al.2008] Shishtla, P. M., Gali, K., Pingali, P., and Varma, V. (2008). Experiments in telugu ner: A conditional random field approach. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.*

[Solorio2007] Solorio, T. (2007). Malinche: A ner system for portuguese that reuses knowledge from spanish. *Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo*, 10:123–136.

[Stenetorp et al.2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

[Straka et al.2016] Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*, pages 4290–4297.

[Stubbs et al.2015] Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

[Sundermeyer et al.2012] Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association.*

[Sundheim1995] Sundheim, B. M. (1995). Overview of results of the muc-6 evaluation. In *Proceedings of the 6th conference on Message understanding*, pages 13–31. Association for Computational Linguistics.

[Sureka et al.2009] Sureka, A., Goyal, V., Correa, D., and Mondal, A. (2009). Polarity classification of subjective words using common-sense knowledge-base. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pages 486–493. Springer.

[Takatsuka et al.2007] Takatsuka, M., Tada, M., and Sasaki, R. (2007). Proposal of the e-discovery system for sanitizing disclosure information and for securing evidence. In *Future Generation Communication and Networking (FGCN 2007)*, volume 2, pages 102–107. IEEE.

[Teixeira et al.2011] Teixeira, J., Sarmento, L., and Oliveira, E. (2011). A bootstrapping approach for training a ner with conditional random fields. In *Portuguese Conference on Artificial Intelligence*, pages 664–678. Springer.

[Tjong Kim Sang2002a] Tjong Kim Sang, E. F. (2002a). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

[Tjong Kim Sang2002b] Tjong Kim Sang, E. F. (2002b). Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

[Todorovic et al.2008] Todorovic, B. T., Rancic, S. R., Markovic, I. M., Mulalic, E. H., and Ilic, V. M. (2008). Named entity recognition and classification using context hidden markov model. In *2008 9th Symposium on Neural Network Applications in Electrical Engineering*, pages 43–46. IEEE.

[Toral and Munoz2006] Toral, A. and Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

[Tran et al.2015] Tran, K.-N., Christen, P., Sanner, S., and Xie, L. (2015). Context-aware detection of sneaky vandalism on wikipedia across multiple languages. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–391. Springer.

[Vijayarani et al.2015] Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

[Yadav and Bethard2018] Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

[Zhou and Su2002] Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.

Appendixes

Social Opinion Project

The Social Opinion Project is a research project developed by INOV INESC Inovação. This project aims to understand the population sentiment through social networks, regarding a certain theme or news. The main focus is the text processing in the Portuguese language and the analysis of twitts sentiment published on social networks. As we can see in Figure 7.1, the first phase consists of processing and extract entities from the news. These entities are used to search for related information on social networks, in this case, Twitter. With the information of the twitts extracted we are able to perceive: The areas of greatest geographical affluence where the theme is being commented on; The associated twitts sentiment on social networks (being classified as Positive, Neutral or Negative); The number of publications evolution over the last few days;

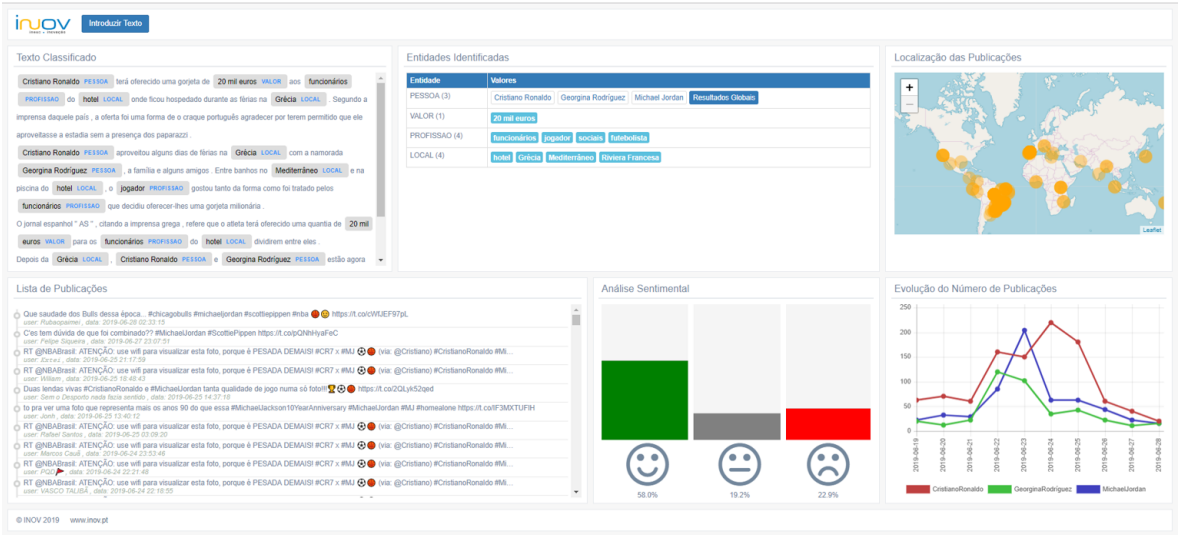

FIGURE 7.1: Social Opinion Project overview

The module used for entities extraction in this project consists of the work developed in this dissertation. The NER module developed was fully integrated without changes or modifications. In this case, were not only considered the annotations of the entities corresponding to the Personal Identification Numbers.

Part-of-Speech Tagging classes

Table 7.1 presents the conversion applied to the Part-of-Speech tags of the Floresta Sintáctica corpus. In the left column are represented the original corpus tags, in the right column are represented the corresponding tags used in this dissertation for the training of models.

Original Tag	**Corresponding Tag**
n	"NOUN",
num	"NUM",
v-fin	"VERB",
v-inf	"VERB",
v-ger	"VERB",
v-pcp	"VERB",
pron-det	"PRON",
pron-indp	"PRON",
pron-pers	"PRON",
art	"DET",
adv	"ADV",
conj-s	"CONJ",
conj-c	"CONJ",
conj-p	"CONJ",
adj	"ADJ",
ec	"PRT",
pp	"ADP",
prp	"ADP",
prop	"NOUN",
pro-ks-rel	"PRON",
proadj	"PRON",
prep	"ADP",
nprop	"NOUN",
vaux	"VERB",
propess	"PRON",
v	"VERB",
vp	"VERB",
in	"X",
prp-	"ADP",
adv-ks	"ADV",
dad	"NUM",
prosub	"PRON",
tel	"NUM",
ap	"NUM",
est	"NOUN",
cur	"X",
pcp	"VERB",
pro-ks	"PRON",
hor	"NUM",
pden	"ADV",
dat	"NUM",
kc	"ADP",
ks	"ADP",
adv-ks-rel	"ADV",
npro	"NOUN"

Table 7.1: Conversion of Part-of-Speech Tagging classes