Universidade de Évora - Instituto de Investigação e Formação Avançada

Programa de Doutoramento em Informática

Tese de Doutoramento

A Graph-Based Framework for Data Retrieved from Criminal-Related Documents

Gonçalo José Freitas Carnaz

Orientador(es) | Mário João Gonçalves Antunes
Vitor Beires Nogueira

Évora 2021

Universidade de Évora - Instituto de Investigação e Formação Avançada

Programa de Doutoramento em Informática

Tese de Doutoramento

A Graph-Based Framework for Data Retrieved from Criminal-Related Documents

Gonçalo José Freitas Carnaz

Orientador(es) | Mário João Gonçalves Antunes
Vitor Beires Nogueira

Évora 2021

A tese de doutoramento foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor do Instituto de Investigação e Formação Avançada:

Presidente	Irene Pimenta Rodrigues (Universidade de Évora)
Vogais	Catarina Helena Branco Simões Silva (Universidade de Coimbra) Dora Regina Oliveira Melo (Instituto Politécnico de Coimbra) José Saias (Universidade de Évora) Maria Luísa Torres Ribeiro Marques da Silva Coheur (Universidade Técnica de Lisboa - Instituto Superior Técnico) Vitor Beires Nogueira (Universidade de Évora) (Orientador)

Évora 2021

Dedico à voz surda, ao olhar ausente, mas à presença eterna.

Acknowledgments

"Era pelo bem comum. E por isso aconteceu."

Imperador Marco Aurélio (121-180 d.C.),
Meditações, Livro 4

When I began this doctoral program, I had no idea of the difficulties, the resilience and strength that was need to complete the program. Along the way, I started to "compare" to an *Tour de France*, and the famous climbs to the Alpe d'Huez. Therefore, this PhD was and always will be my *Tour de France*.

The humble accomplishment of this thesis would not have been possible without the contribution of many individuals, to whom I express my appreciation and gratitude. First and foremost I want to thank my advisers Professor Vitor Beires Nogueira and Professor Mário Antunes, for the patience, encouragement, and for being present with their knowledge and guidance in all steps of my PhD. Also, I want to thank to criminal domain experts Rui Santos, Adolfo Santos and Jorge Cordeiro for the guidance related to criminal investigations. I could not forget, the people who contributed in various ways to the fulfillment of this journey, to Professor Nuno Ferreira, Professor Frutuoso Silve, Professor Salviano Soares for his encouragement.

There were several people, during my doctoral program, whose friendship and support gave me the strength and motivation to carry on, therefore, to my friends Telmo Lino, Leonel Cardoso, Pedro Couceiro and Ricardo Dinis. An at last, but not least, my Family - with all support and not realizing what I was really doing.

Contents

List of Figures

List of Tables

List of Acronyms

AD	Lying Trees
CSV	Comma-Separated Values
CRFs	Conditional Random Fields
DM	Data Mining
ETL	Extract, Transform, and Load
GDB	Graph Databases
HTML	Hypertext Markup Language
HMMs	Hidden Markov Models
IATE	Interactive Terminology for Europe
IE	Information Extraction
KR	Knowledge Representation
KBs	Knowledge Bases
KB	Knowledge Base
MaxEnt	Maximum Entropy Model
MUC-6	Sixth Message Understanding Conference
ML	Machine Learning
NER	Named-Entity Recognition
NE	named-entity
NEs	named-entities
NLP	Natural Language Processing

NERC	Named-Entity Recognition and Classification
OPC	Órgãos de Polícia Criminal
OIE	Open Information Extraction
OSINT	Open Source Intelligence
POS	Part-Of-Speech Tagging
PDF	Portable Document Format
RDF	Resource Description Framework
RE	Relation Extraction
SIIC	Sistema Integrado de Informação Criminal
SRL	Semantic Role Labeling
SVM	Support Vector Machines
SEM	Simple Event Model
SNA	Social Network Analysis
SQL	Structured Query Language
UML	Unifed Modelling Language
URL	Uniform Resource Locator
TBX	TermBase eXchange
WWW	World Wide Web
XML	Extended Markup Language

Resumo

Uma framework baseada em grafos para dados recuperados de documentos relacionados com crimes

A digitalização das empresas e dos serviços tem potenciado o tratamento e análise de um crescente volume de dados provenientes de fontes heterogeneas, com desafios emergentes, nomeadamente ao nível da representação do conhecimento. Também os Órgãos de Polícia Criminal (OPC) enfrentam o mesmo desafio, tendo em conta o volume de dados não estruturados, provenientes de relatórios policiais, sendo analisados manualmente pelo investigadores criminais, consumindo tempo e recursos.

Assim, a necessidade de extrair e representar os dados não estruturados existentes em documentos relacionados com o crime, de uma forma automática, permitindo a redução da análise manual efetuada pelos investigadores criminais. Apresenta-se como um desafio para a ciência dos computadores, dando a possibilidade de propor uma alternativa computacional que permita extrair e representar os dados, adaptando ou propondo métodos computacionais novos.

Actualmente existem vários métodos computacionais aplicados ao domínio criminal, nomeadamente a identificação e classificação de entidades nomeadas, por exemplo narcóticos, ou a extracção de relações entre entidades relevantes para a investigação criminal. Estes métodos são maioritariamente aplicadas à lingua inglesa, e em Portugal não há muita atenção à investigação nesta área, inviabilizando a sua aplicação no contexto da investigação criminal.

Esta tese propõe uma solução integrada para a representação dos dados não estruturados existentes em documentos, usando um conjunto de métodos computacionais: *Preprocessamento de Documentos*, que agrupa uma tarefa de Extracção, Transformação e Carregamento adaptado aos documentos relacionados com o crime, seguido por um pipeline de Processamento de Linguagem Natural aplicado à lingua portuguesa, para uma análise sintática e semântica dos dados textuais; *Método de Extracção de Informação 5W1H* que agrupa métodos de Reconhecimento de Entidades Nomeadas, a detecção da função semântica e a extracção de termos criminais; *Preenchimento da Base de Dados de Grafos e Enriquecimento*, permitindo a representação dos dados obtidos numa base de dados de grafos Neo4j.

Globalmente a solução integrada apresenta resultados promissores, cujos resultados foram validados usando protótipos desemvolvidos para o efeito. Demonstrou-se ainda a viabilidade da extracção dos dados não estruturados, a sua interpretação sintática e semântica, bem como a representação na base de dados de grafos.

Palavras chave: Representação do Conhecimento, Domínio Criminal, Processamento de Linguagem Natural, Relatórios Policiais, Base de Dados de Grafos, 5W1H

Abstract

The digitalization of companies processes has enhanced the treatment and analysis of a growing volume of data from heterogeneous sources, with emerging challenges, namely those related to knowledge representation. The Criminal Police has similar challenges, considering the amount of unstructured data from police reports manually analyzed by criminal investigators, with the corresponding time and resources.

There is a need to automatically extract and represent the unstructured data existing in criminal-related documents and reduce the manual analysis by criminal investigators. Computer science faces a challenge to apply emergent computational models that can be an alternative to extract and represent the data using new or existing methods.

A broad set of computational methods have been applied to the criminal domain, such as the identification and classification named-entities (NEs) or extraction of relations between the entities that are relevant for the criminal investigation, like narcotics. However, these methods have mainly been used in the English language. In Portugal, the research on this domain, applying computational methods, lacks related works, making its application in criminal investigation unfeasible.

This thesis proposes an integrated solution for the representation of unstructured data retrieved from documents, using a set of computational methods, such as *Preprocessing Criminal-Related Documents* module. This module is supported by Extraction, Transformation, and Loading tasks. Followed by a Natural Language Processing pipeline applied to the Portuguese language, for syntactic and semantic analysis of textual data. Next, the *5W1H Information Extraction Method* combines the Named-Entity Recognition, Semantic Role Labelling, and Criminal Terms Extraction tasks. Finally, the *Graph Database Population and Enrichment* allows us the representation of data retrieved into a *Neo4j* graph database.

Globally, the framework presents promising results that were validated using prototypes developed for this purpose. In addition, the feasibility of extracting unstructured data, its syntactic and semantic interpretation, and the graph database representation has also been demonstrated.

Keywords: Knowledge Representation, Criminal Domain, Natural Language Processing, Criminal Investigation Reports, Graph Databases, 5W1H

1

Introduction

"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."

J.R.R. Tolkien, The Lord of the Rings

The data retrieved from criminal-related documents need to be represented as an organized collection of data. However, this can only be executed if we understand the domain and the written language, along with the computational methods that may fulfill the requirements to understand each required challenge. This chapter introduces the motivations for this thesis, the challenge that led to this work, the research questions elaborated, the main goals, and the expected outcomes

1.1 Motivation

The interaction between computer science and the criminal domain is an emerging topic motivated by several factors, namely the necessity of processing the growing volume of data from different sources to help criminal analysis. Other domains have the same issue, where the volume of data is produced through organizations and services, reflecting the increase of computer systems and the network interconnections supported by the World Wide Web (WWW). IDC [1] refers that the data volume will increase by a factor of 300 (from 130 exabytes to 40.000 exabytes) from 2005 to 2020. This data comes from heterogeneous sources in different forms, such as structured, semi-structured, and unstructured data that computational methods could process, such as Natural Language Processing (NLP) (Gleick and Calil, 2013) (Oussous et al., 2018). The data retrieved and represented is an asset for organizations, with potential for diverse sectors, such as academic research, industrial, government, or sports (Cavanillas et al., 2016).

Under the umbrella of this digital overgrowing, the criminal-related information from different police department sources raises problems of information sharing and collaboration between departments and agencies, and the promotion of crime intelligence analysis and knowledge management by each department (Chen et al., 2003). It is associated with the fact that *"the treatment of data and information related to criminal activities has gained drastically in importance"* (Albertetti and Stoffel, 2012). This domain needs a way to understand the investigation process performed by experienced police officers that capture the investigative knowledge (Dean et al., 2008) and the modeling using a computational approach for the Law Enforcement (Seidler12 et al., 2012), or data science approaches applied to criminal analysis (Qazi et al., 2017).

When considering criminal activities related to the Portuguese context, we may identify a challenge and a necessity to automatically process the data overgrowing. This data is usually stored on databases (paper and digital) or in criminal investigation reports. To have a perspective of data volume, the Portuguese Internal Security Report [2] gathers all the crimes reported by the OPC that compose the Internal Security System [3]. According to the 2018 report, an amount of 347.204 crimes have been reported. Table 1.1 enumerates the number of crimes, divided by general or violent and dangerous, between $2017 - 2018$.

Criminality	Reported Crimes (#Count)
general	333223
violent and serious	13981

Table 1.1: General and Violent/Dangerous Criminality (2017-2018).

Considering that several reports and forensic evidence are produced in textual form in each reported crime, the amount of data produced overgrows to numbers that challenge human and digital processing. Thus, an automatic process to retrieve unstructured data and its knowledge representation is a challenge for the research community.

The literature presents different forms to process, understand, and represent unstructured data, divided into three main fields, the Extract, Transform, and Load (ETL), NLP and Knowledge Representation (KR), to address some of the issues, such as identification and classification of NEs. To our knowledge, the NLP applied to the Portuguese language has introduced computational methods that allow the application to different domains, such as the medical domain. However, it is still a field untapped in the criminal domain,

[1] See www.idc.com [Accessed: 1 May 2020].

[2] See www.portugal.gov.pt [Accessed: 1 Jul 2019].

[3] See legislacao.mai-gov.info [Accessed: 1 October 2019].

with needs regarding the syntactic and semantic analysis of criminal-related documents. The systems that support the KR for the Portuguese police, like the Sistema Integrado de Informação Criminal (SIIC), use a relational approach for data representation, with obvious limitations in the representation of unstructured data retrieved from documents.

Our motivation arises from the Portuguese police department's daily work, like the manual analysis of criminal-related documents and the lack of approaches in the Portuguese language applied to the criminal domain. The introduction of a framework to extract and represent data based on computational methods from the ETL, NLP, Machine Learning (ML) or KR fields is an add-on to the computational field applied to the criminal domain using the Portuguese language. This framework carries benefits for the police departments, and more directly to the criminal investigators, such as:

- decreases the time-consuming on manual analysis of the criminal-related documents;
- an easy way to query the data.

This thesis aims to deal with data retrieval from criminal-related documents and the representation into a graph database, enabling data querying.

1.2 Problem Description

The challenge arose from the interviews with domain experts (criminal investigators), and it starts when a criminal event originates several documents from different sources, like SMS, chat logs, emails, taping transcripts, newspapers, and others. This criminal-related information is synthesized into a criminal investigation report that describes the actors, the pieces of evidence, and the description of the crime. These reports follow a narrative form, written in a template divided by sections (Galante, 2016). When a criminal investigation ends, they are sent to a Court Law and rest as sources of knowledge for further investigations. The analysis is performed in IBM™ i2 Analyst's Notebook tool, where domain experts perform a manual annotation of the actors and their relations, generates a network of entities and relations. However, as the volume of reports increases, the need for a tool to automate the process increases.

The figure 1.1 shows an diagram that synthesized the problem.

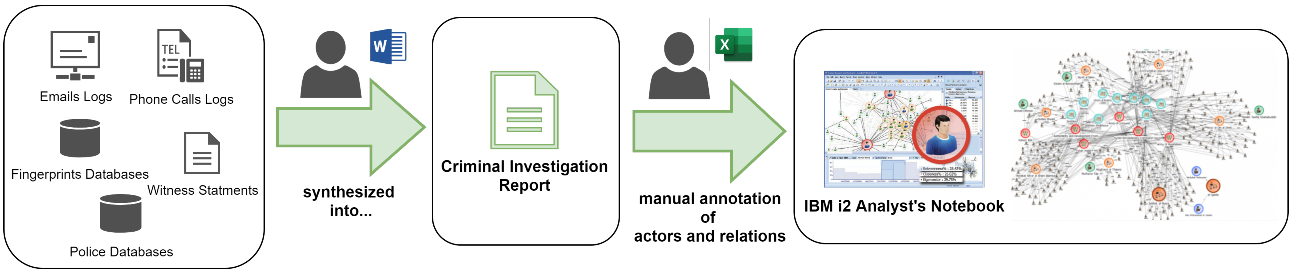


Figure 1.1: Workflow for the Criminal Investigation Reports Analysis.

The analysis of documents is conditioned by the personal vision (interpretation) of each domain expert. The steps performed by domain experts until IBM™I2 Analyst's Notebook representation are:

- the documents with criminal information are generated from different sources and formats;

- the information of these documents are synthesized into a criminal investigation report, using the Microsoft™ Word;
- the domain experts execute a manual annotation of actors and relations, and create a Microsoft™ Excel sheet with results;
- the Microsoft™ Excel sheet is introduced into IBM™ i2 Analyst's Notebook, and a network of actors and their relations are created for analysis.

1.2.1 Research Questions

In this subsection, we define the research questions for our research and pinpoint what we want to achieve, with the objective of providing a more precise focus and a better understanding of the matter in question. The contribution of this PhD thesis is thus, to answer the following research questions:

> **Research Question 1**: *"Is it possible to understand the unstructured data concerning the criminal domain by applying computational methods?"*
>
> **Research Question 2**: *"Can we identify relevant entities related to the criminal domain?"*
>
> **Research Question 3**: *"Can we extract information from criminal-related documents by using an 5W1H approach?"*
>
> **Research Question 4**: *"Is it possible to apply a graph-based framework represent information extracted from documents related to criminal investigations?"*

This work will be supported by a detailed study of the related work published in scientific journals, books and conferences, and applying technological and scientific knowledge to answer the research questions described above.

1.2.2 Contributions

This work provides an end-to-end framework based on a graph database that provides a unified approach to represent the criminal-related documents. We named the framework as **SEMCrime**, which means the fusion of two words *"**SEMantic**"* and *"**Crime**"*. *SEMCrime* provides a way of representing the data retrieved from criminal-related documents in a semantic form and populating a *Neo4j* graph database with the retrieved data.

We need to consider the contributions and applicability that can eventually arise from this study. In this sense, the expected outputs provided by this work are as follow:

- to give an approach that ties together the criminal and the computer science domains focused on the Portuguese criminal environment;
- to developed an end-to-end framework, from data extraction to data knowledge representation into the selected graph database;

- to develop an information extraction method based on an *5W1H* approach for understanding the semantics in each criminal-related documents;
- to train models that identify and classify named-entities related to the criminal domain;
- to identify and classify criminal terms in criminal-related documents;
- a graph database that enables end-user queries.

To our knowledge, this framework is the first approach applied to the Portuguese criminal domain.

1.3 Thesis Structure

The remaining chapters of this thesis are organized as follows:

- *Theoretical Foundation*: chapter 2 describes an overview of the theoretical foundations of researched areas involved in this work, such as data concepts, ETL, NLP, ML and Graph Databases (GDB), providing the building blocks that will support our proposals;
- *State of the Art*: chapter 3 describes the state of the art applied to the criminal domain and related work that could help us achieve the proposed outcomes;
- *SEMCrime Framework and Preprocessing*: chapter 4 describes an overview of the *SEMCrime* architecture; introduces the criminal-related documents used as dataset; and finally, discussed the module that enables the transformation of unstructured into semi-structured data by performing a group of tasks that allows extraction, cleaning, and loading activities. Also, includes the description of syntactic analysis of documents by using a NLP pipeline
- *Recognizing Entities in Criminal-Related Documents*: chapter 5 an unified approach to a Named-Entity Recognition (NER) module applied to the criminal domain in Portuguese language;
- *Criminal Related Data Representation in a Graph Database*: chapter 6 proposes a new approach for knowledge representation supported by a *Neo4j* graph database tailored to Portuguese criminal-related information retrieved from written documents, and used by police departments or investigative journalism. It is supported by a *Neo4j Criminal-Related Documents Representation Module*, which is divided into the following modules: first, a *Criminal Information Extraction Module*, which enables the extraction of semantic knowledge of criminal-related documents, by combining the Semantic Role Labeling (SRL), NER, Criminal Term Extraction, into the *5W1H* Information Extraction Method; second, a *Graph Database Population and Enrichment* module to populate and enrich the graph database with the data retrieved;
- *Final Remarks*: chapter 7 explores the conclusions regarding the proposed work, the obtained results, and future work directions.

2

Theoretical Foundation

"It is strange that only extraordinary men make the discoveries, which later appear so easy and simple."

Georg Lichtenberg

This chapter aims to support the theoretical foundations of the building blocks that touch the different computer science fields and the criminal domain. We start by reviewing the criminal domain, namely the concepts, definitions, and a focus on Portuguese criminal investigation. We then followed by an introduction of basic concepts of data and the ETL approach. A brief explanation of the NLP field, from the syntactic to the semantic analysis. In addition, an introduction to the ML methods used in our approach. Finally, graph database concepts and technologies are introduced.

2.1 The Criminal Domain Context

The International Association of Crime Analysts (IACA) introduces a definition for **criminal analysis** (also referred to as crime analysis) *"...process in which a set of quantitative and qualitative techniques are used to analyze data valuable to police agencies and their communities. It includes analyzing crime and criminals, crime victims, disorder, quality of life issues, traffic issues, and internal police operations. Its results support criminal investigation and prosecution, patrol activities, crime prevention and reduction strategies, problem-solving, and the evaluation of police efforts." (IACA et al., 2014).* The criminal analysis is recognized by the IACA in four categories (Albertetti and Stoffel, 2012) (IACA et al., 2014):

- **Crime intelligence** analysis: focuses on data about people involved in crimes and about the authors' information (intelligence);
- **Tactical crime** analysis: investigates space, time, offenders, victims, and "modus operands" from police databases, for example, the crime patterns and crime linkage are two analysis conducted in this category;
- **Strategic crime** analysis: retrieved from information sources, such as trend analysis and hot spot analysis;
- **Administrative crime** analysis: is linked to the global organization of police agencies regarding communication, budget, statistics, or employment.

Other concepts are common in this field. **Intelligence** refers to the definitions described above and allows the understanding of all processes, persons, and activities involved in a crime analysis or criminal investigation (Atkin, 2011) described as information acquired, exploited, analyzed, and protected by police departments' activities support criminal investigations. During this process, **evidence** is identified during a criminal investigation protected from disclosure during criminal proceedings. Criminal police departments are focused on the **criminal investigation**, which is the discipline that involves the collection and study of information related to a crime. Kind (1994) defines criminal investigation as a process of starting an investigation of a crime and ending up in a Court Law, divided into a problem to find, the decision to charge, and the problem.

In criminal police departments, data is gathered from different sources during a criminal investigation (Kind, 1994) from three identifying types of sources: open, closed, and classified sources. **Open source** is the one that is publicly available (Open Source Intelligence (OSINT)), such as online newspapers, which can be legitimately obtained (Akhgar et al., 2017). However, some issues could arise by using these sources. For example, public domain information is frequently considered limited or inaccurate. **Closed source** refers to information collected for a specific issue with limited access. In most cases, this information is in the form of structured databases. **Classified** data is collected for a specific task using by means, like human and technical resources, but with restrictions on its dissemination and accessed levels.

2.1.1 5W1H information

Questions like *"What Happened?"*, *"Who or What participated in the event?"*, *"Where did it happen?"* and *"When did it happen?"* are intrinsically connected to a criminal investigation. Each actor played some role (temporary or permanent) assigned to a particular crime event. Thus, events provide a way to describe relationships between people, objects, locations, actions, or time. Numerous events could be found in criminal-related documents; some of these events were linked to a specific domain, such as homicide, organ trafficking, robbery, or others. There are also events not explicitly linked to the criminal domain; however, "all information is relevant.".

The *5W1H* is an abbreviation for six questions, namely **What?, Who?, Where?, When?, Why? and How?**. This approach consists of answering these questions systematically to collect all the data necessary to report a situation. By doing this, the problem and the context are described accurately. This approach is used in several fields, such as investigative journalism and criminal analysis. Our objective is to apply *5W1H* in the criminal context. Thus, the next paragraphs analyzed each question, the ***5W's***: What?, Who?, Where?, When?, and Why?, and the ***1H*** (How?).

What

In the criminal context, the confirmation of the existence of a crime triggers a criminal action that ends (typically) with a criminal investigation that is finished with a criminal investigation report. This question focuses on characterizing all the facts, actions, situations, directly and indirectly, related to the criminal investigation (Braz, 2019). In our context, the "What" question describes a state, fact, action, or situation that will lead to a criminal action from participants in the event. To simplify, we have focused on an event or action "caused" by a verb, which is used as a trigger to answer other questions being answered by using the arguments (participants of the event) that play different roles in it. The following sentence displays underlined terms that identify the answer(s) to the questions.

> *"O Rui Silva e o Pedro Silva* ***assaltaram*** *o Banco de Portugal em Coimbra, pelas 14 horas."*
>
> **In English**: "Rui Silva and Pedro Silva **robbed** Bank of Portugal in Coimbra, at 2 pm".

Although not all verbs are related to the criminal domain, we still need to extract them because a simple word, such as "talk" could lead to a "to steal something or someone". However, some verbs and nouns are related to the criminal domain with a wide range of synonyms.

Who (and Whom)

These questions seek to identify all suspects or agents who, directly or indirectly, may or may not have to do with the investigation. Thus, the importance of determining and identifying the victims, witnesses, deponents, informers, police, and others involved (Braz, 2019). The identification of these entities involved can give rise to originate other questions:

- Who directs the operation?;
- Who committed the crime?;
- Which organization is behind the crime?;
- Who authorizes it?

Identifying people, gangs, or organizations allows us to answer the previously listed questions. However, identification is not always distinct, given that the omission of the name of the person or organization (for example, the use of terms such as witnesses, police, or informants) is present in everyday criminal writing. Other authors, like Wunderwald (2011) determined that the subject of the sentence identifies the "who". For the "Whom" identification, the sentence's object could be used as the portion of the text with the same answer. The terms underlined in the sentences identify the answer(s) to the questions in this section.

> *"O **Rui Silva** e o **Pedro Silva** assaltaram o **Banco de Portugal** em Coimbra, pelas 14 horas."*
>
> **In English**: "**Rui Silva** and **Pedro Silva** robbed **Bank of Portugal** in Coimbra, at 2 pm".

In our context, the NEs are used to identify the persons and organizations, among others. Those are marked to identify the "Who" and the "Whom".

When

Crime events are always associated with time and date periods. Therefore, finding NEs, like date or time within the text, is useful to mark crime events and associated events. Furthermore, time and date are indispensable elements for excluding facts and evidence production from a criminal analysis perspective. Thus, the answer to the question "When" allows us to identify the event's date by combining it with other elements that can produce the crime evidence (Braz, 2019). The terms underlined in the sentences identify the answer(s) for the questions in this section.

> *"O Rui Silva e o Pedro Silva assaltaram o Banco de Portugal em Coimbra, pelas **14 horas**."*
>
> **In English**: "Rui Silva and Pedro Silva robbed Bank of Portugal in Coimbra, at **2 pm**."

Extracting terms related to date/time and identifying with others is a difficult task, and in many cases, it is not possible. For example, some often NEs are labeled as to date/time type; however, sometimes, it is difficult to identify their relationship with the event.

Where

Events are associated with a particular place. Therefore, an essential task for producing criminal evidence is geo-localization. However, it is challenging to represent this kind of knowledge because it may involve different property types, such as street name, door number, city name, or even rooms in a house (like kitchen or bedroom). The terms underlined in these sentences identify the answer(s) to the questions in this section.

> *"O Rui Silva e o Pedro Silva assaltaram o Banco de Portugal em **Coimbra**, pelas 14 horas."*
>
> **In English**: "Rui Silva and Pedro Silva robbed Bank of Portugal in **Coimbra**, at2 pm."

In our context, the NEs are used to identify the locations.

Why

This issue is associated with motivation or the cause of a given event. The motivation is based on the event's author and can be linearly understandable, or sometimes, it goes beyond human or rational understanding (Braz, 2019). From a computational perspective, representation is not linear because it can involve entities (relatively easy to identify). However, it also can involve concepts that, given their semantic nature, are complex to identify, such as psychiatry justifications.

How

The criminal event needs to be reconstructed for judicial proof or strictly the technical nature reasons. This process is named as *"modus operandi"*. Thus, the "how" question allows us to understand how a particular crime was committed. This process is detailed in (Braz, 2019).

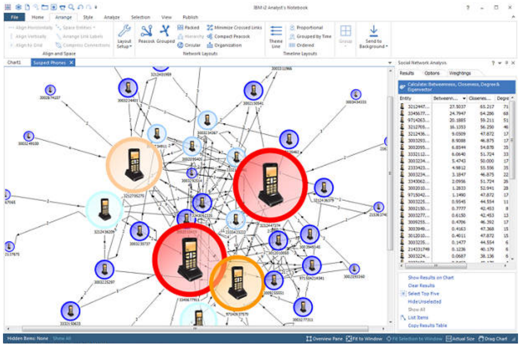
2.1.2 Criminal Investigation in Portugal

The concepts introduced earlier are also applied to the criminal domain in Portugal. However, in the Portuguese perspective, the criminal investigation is defined by Law 49/2008 of 27th August [1], named as *Lei de Organização da Investigação Criminal*, which defines it as *"... the set of measures that, under the terms of the criminal procedural law, are designed to ascertain the existence of a crime, determine its agents and their responsibility and discover and collect evidence as part of the process."*. Thus, the criminal investigation is like a complex system, where information and their collection plays a decisive role. To understand the origin of the data analyzed throughout the thesis, we need to introduce some concepts, processes, documents, and software used by criminal police.

In Portugal, the police departments investigate crimes reported by individuals or the crimes detected by the **Orgãos de Polícia Criminal**, known as OPC (Costa, 2012) which are composed, for example, by the Policia Judiciária [2], Policia de Segurança Pública [3], and the Guarda Nacional Republicana [4] and among others. These policies are responsible for undertaking the criminal investigation, retrieving pieces of evidence from multiple sources, and formats (Pereira, 2013).

Software to Support Criminal Investigation

The Portuguese criminal police use the **IBM™ i2 Analyst's Notebook** [5], which is a software product that helps to turn input data into structured knowledge, by displaying connected networks, spacial information, social network analysis, and or temporal views to help us to uncover hidden connections or patterns in data.



Entity	Betweenn...	Closenes...	Degre
3212447...	27.5037	65.217	71
3345677...	24.7947	64.286	68
9714263...	20.1885	59.211	51
3212709...	16.1353	56.250	46
3212436...	9.0509	47.872	17
3003293...	8.9088	46.875	17
3002095...	6.8544	54.878	35
3332112...	6.0640	51.724	33
3003234...	5.4743	50.000	17
2333423...	4.9812	55.556	35
3003234...	3.1847	46.875	22
3343062...	2.0956	51.724	26
3002010...	1.2833	52.941	28
9715042...	1.1490	47.872	17
3003226...	0.9545	44.554	11
3002150...	0.7777	42.453	8
3003277...	0.6150	42.453	13
3009255...	0.4706	46.392	17
3003949...	0.4163	47.368	15
3012010...	0.4011	47.872	15
3003235...	0.1477	44.554	6
214331749	0.1236	40.179	6
3003224...	0.0687	38.136	6

Figure 2.1: IBM™ i2 Analyst's Notebook Screenshot.

[1] See www.pgdlisboa.pt [Accessed: 1 May 2020].
[2] See www.policiajudiciaria.pt [Accessed: 1 May 2020].
[3] See www.psp.pt [Accessed: 1 May 2020].
[4] See www.gnr.pt [Accessed: 1 May 2020].
[5] See www.ibm.com/products/analysts-notebook [Accessed: 1 May 2020].

Another tool associated with this product is the **IBM i2 Text Chart** that enables text extraction and visualization from unstructured data by automatically discovering and highlighting entities, such as organizations, persons, or events, in documents. However, the **Text Chart Auto Mark** feature is developed in English. Therefore, it is not available for other languages, and **Text Chart Auto Mark** supports terms found in documents in the following languages, such as English, French, German, Spanish, Italian, and Dutch (for further information read [6]).

The Criminal Integrated Information System (in Portuguese: SIIC) aims to ensure a high level of security in the exchange of criminal information between OPC's, to carry out preventive and criminal investigation actions, with the objective to strengthening crime prevention and repression, by performing basic operations of CRUD (Create, Read, Update and Delete) over the information system.

2.2 Dealing With Data

Regarding the reference to unstructured data, we need to introduce theoretical support to concepts, techniques, or methods frequently used in the data environment. Therefore, we summarize several foundations that support our work.

The DIKW Pyramid (Frické, 2019), known as the knowledge pyramid, introduces several concepts like data, information, knowledge, and wisdom. The figure 2.2 represents the knowledge pyramid:

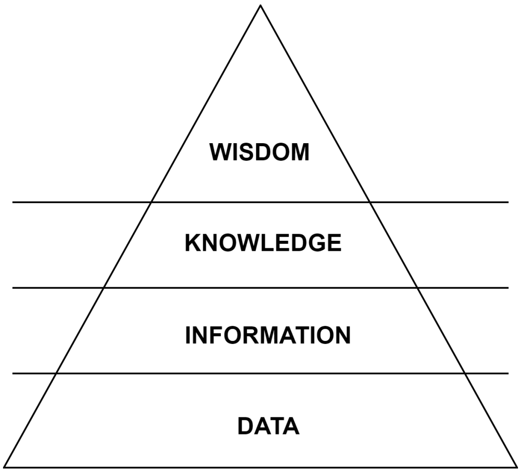


Figure 2.2: DIKW Pyramid.

The followed items describe the pyramid elements:

- **Data**: are symbols and signals produced, retrieved and stored to represent properties of objects, events and their environments (Frické, 2019) (Zins, 2007);
- **Information**: is *"defined as a message that contains relevant meaning, implication, or input for decision and/or action."* (Liew, 2007), that arises from real-time and historical sources. A question that starts with "who", "what", "where", "when" or "how many" and answer are inferred from data, were data becomes information (Frické, 2019);
- **Knowledge**: encloses the terms that revels cognition or recognition (know-what), capacity to act (know-how), and understanding (know-why) (Liew, 2007) (Frické, 2019), *"is an extrapolative and*

[6] See www.ibm.com/support/pages/ibm-i2-text-chart-release-notes [Accessed: 1 May 2020].

non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.)" (Liew, 2013);

- **Wisdom**: when individuals adds value to the information extracted (Frické, 2019).

However, in an environment with different data sources, we can found data assuming different forms, such as unstructured, structured, and semi-structured data. These concepts are described below:

- **Unstructured data**: have an internal structure, capable of becoming understandable by humans, but without a pre-defined data model or scheme. This data may be textual or non-textual and human- or machine-generated. The automatic processing of this kind of data requires advanced techniques to analyze the text regarding its context effectively. Typical human-generated unstructured data include text files, email messages, Social Media posts, Website data, Mobile data, Communications streams, Media and Business applications outputs (Kanimozhi and Venkatesan, 2015);
- **Structured data**: the data included in relational database systems, such as database tables, objects, tags, reports, indexes, and other database concepts. Therefore, the data have a structured data scheme (Kanimozhi and Venkatesan, 2015);
- **Semi-structured data**: equivalent to structured data with a lack of data model structure that does not follow a formal structure. A scheme definition is optional and contains tags or other markers to separate semantic elements, enabling information grouping and hierarchies. The languages, like Extended Markup Language (XML) or another mark-up language, Javascript object notation (JSON), are used to manage semi-structured data (Kanimozhi and Venkatesan, 2015).

The data is submitted in different phases. Which starts with the **Data Generation** phase pointed out as the generation of data from heterogeneous data sources, such as mobile, satellites, laboratories, supercomputers, search entries, chatting records, blog messages, social media posts, videos, sounds, sensors, and all other available data sources (Hu et al., 2014).

Followed by the **Data Acquisition** phase is divided into three steps, such as data collection, data transmission, and data pre-processing. **Data Collection** introduces a process of retrieving raw data from sources, eliminating "dirty data" to remove impact to subsequent data analysis. This phase is conditioned by the physical characteristics of the data sources. Acquiring data from heterogeneous sources is also a time-related issue, divided by Hu et al. (2014) into two paradigms:

- **Streaming processing** enables data analysis as soon as possible to achieve faster results. Arriving in the form of a stream, continuous and carrying an enormous volume of data to be stored, thus processed by small portions in limited memory;
- **Batch processing** data is stored first, and posterior is analyzed. A well-defined batch processing model is MapReduce (Dean and Ghemawat, 2008), where data is divided into small chunks where systems performed parallel processing over each chunk. Then the results are aggregated together into a single result. As examples of this application, we have the bioinformatics, web mining, and ML fields.

After we have collected the data, a transfer process must be performed, called by **data transmission**, and then data is sent to a repository, such as in a data center or even a flat table. Finally, **data pre-processing** validates the data quality, e.g., noise, redundancy, consistency. Thus, some tasks must be performed to improve the data's quality: Integration, Cleansing, and Redundancy Elimination (Hu et al.,

2014) (Khan et al., 2014). Once the data has been generated and acquired, the **data storage** phase organizes data into a single format and stores it for posterior analysis (Hu et al., 2014). Following, the **data analysis** aims to extract useful and pertinent information regarding the data retrieved from the other phases, regarding the subjects under study, such as for decision-making purposes, interpreting and extrapolating data and determining how to use it, verifying if the data has its legitimacy, to diagnose and infer data faults and predicting future under-analyzed circumstances (Hu et al., 2014). Therefore, the data analysis techniques are needed to extract relevant information, with every technique being applied over data depending on data types, such as association rule learning (Maltby, 2011), data mining (Hu et al., 2014), cluster analysis(Manyika et al., 2011), ML (Manyika et al., 2011), text and statistical analysis (Hu et al., 2014). Finally, the **data destruction** is the process of destroying data stored physically on the devices.

Finally, other concepts need to be retained, such as:

- **document**: defined as a *"unit of discrete textual data within a collection that usually, but not necessarily, correlates with some real-world document such as a business report, legal memorandum, e-mail, research paper, manuscript, article, press release, or news story"* (Feldman et al., 2007);
- **characters**: as low-level components that when grouped, can identify semantic features such as words, terms, and concepts;
- **words**: are selected explicitly from documents and describe the necessary level of semantic richness;
- **terms**: are linguistic units extracted from the corpus of a document through term-extraction methodologies;
- **concepts**: is a feature generated for a document through manual, statistical, or methodologies (Feldman et al., 2007).

2.2.1 Extraction, Transformation, and Loading

We have introduced phases to deal with data in the past section until being ready to be used. First, however, there is an approach named as ETL, which is defined as a process of blending data (unstructured and structured) from multiple sources, such as relational databases, text documents, or spreadsheets. Other definitions emerged like *"an ETL process can be thought of as a directed acyclic graph, with activities and record sets being the nodes of the graph and input-output relationships between nodes being the edges of the graph."* (Vassiliadis, 2003); *"Extract-Transform-Load (ETL) activities are software modules responsible for populating a data warehouse with operational data, which have undergone a series of transformations on their way to the warehouse."* (Vassiliadis et al., 2009); *"The ETL process extracts the data from source systems, transforms the data according to business rules, and loads the results into the target data warehouse."* (Kakish and Kraft, 2012).

Therefore, a ETL process is supported by a sequential group of tasks or operations, named as workflow or pipeline. The extraction process extracts sequentially data with different formats (such as web pages and documents), followed by a transformation task. Data is transformed into a common format using a set of rules and a temporary repository called by **data staging area**. Finally, data is uploaded into a repository, namely a database or a XML file. Figure 2.3 shows an diagram with the ETL approach.

The steps described above integrate different tasks in the sense of processing the extracted data. Each ETL process is defined as follows:

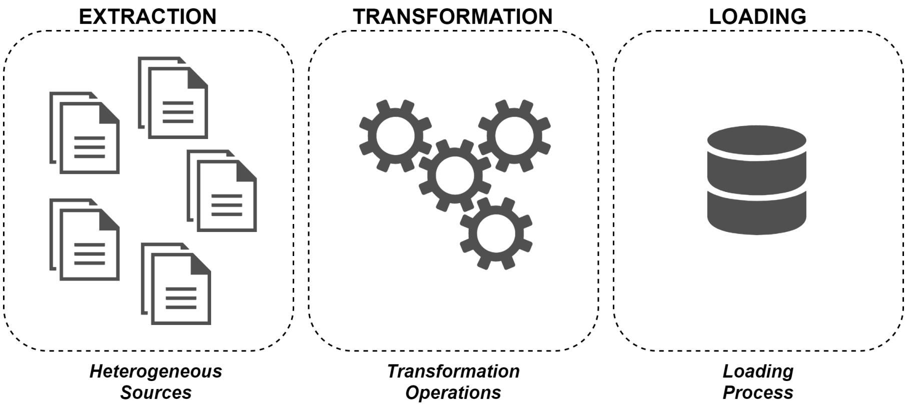


Figure 2.3: ETL Approach.

- *Extraction*: this task aims to identify and collect raw data from heterogeneous sources (such as web pages, spreadsheets, or text documents) to be populated in a target data model. These extraction operations address challenges related to understanding the structure of multiple data sources and semantics of data extracted, data accessibility, and semantic heterogeneity of the data sources (Ong et al., 2017). We need to be aware of (a) which drivers should be used to connect to sources or different file formats, (b) understanding the data scheme/structure of sources, and (c) handling the sources from different nature. Normally, this process is performed overnight (El-sappagh et al., 2011). There are two logical and physical methods regarding the extraction process, the full and incremental extraction (Kakish and Kraft, 2012);

- *Transformation*: this task deals with the mapping and transformation operations of collected data. Some operations should be performed, such as cleaning, filtering, enriching, splitting, or joining. During this operation, data must remain unchanged data by mapping it without information loss (Ong et al., 2017);

- *Loading*: this task involves propagating the transformed data into an endpoint, such as a database or a flat file. Several challenges are associated with the loading process, such as the data quality assessment process, the complexity of incremental data loading, and ETL operation routines (Ong et al., 2017).

ETL have different relationships with inputs (such as Unary, Binary, and N-ary) and outputs (such as routers or filters). In *Unary* activities, is performed a transformation or cleaning data by one input to another output scheme. In *Binary* or *N-ary* activities are executed from multiple inputs and one output (Vassiliadis, 2003). During the data processing, a sort of type used, such as batch, real-time, and near real-time processing, were:

- **Batch processing**: when data processing needs are less time-sensitive, divided into three steps - data is collected, is processed by a separate program, and outputted;

- **Real time processing**: when data processing requires a continuous input, a constant processing, and a steady output of data. Some examples are radar systems, customer service systems, or ATMs banks;

- **Near real-time (NRT) processing**: when data processing occurs periodically, it is focused on data speed processing and can be programmed to process data collection every day, every hour, or every time.

2.3 Natural Language Processing

The language journey began with our birth, from a diminished to an elaborated (speech and written) supporting human interaction. Hence, language communication is completed through words and phrases that humans acquire during their lifetime. Asher and Moseley (2018) enumerated in the *Atlas of the World's Languages*, several languages and dialects from all around the World. Briefly, language is one of the fundamental aspects of human behavior since it allows interaction between individuals and perpetuates knowledge.

Natural language processing is where computer science meets natural languages. Therefore, NLP is a branch of the computer science field, as a multidisciplinary task for analyzing and generating natural languages by computerized methods (Jurafsky and Martin, 2000). This is not considered an easy task, and we need to introduce concepts that somehow divided this task into blocks. Therefore, language could be divided into **knowledge levels** (Indurkhya and Damerau, 2010), which could be used to apply different NLP tasks. Figure 2.4 shows a classical toolkit (Indurkhya and Damerau, 2010) that describes the NLP knowledge levels divided into Lexical or Morphological, Syntactic, Semantics, and Pragmatics Analysis. In addition, there is an initial level, the tokenization (see Section 2.3.1).

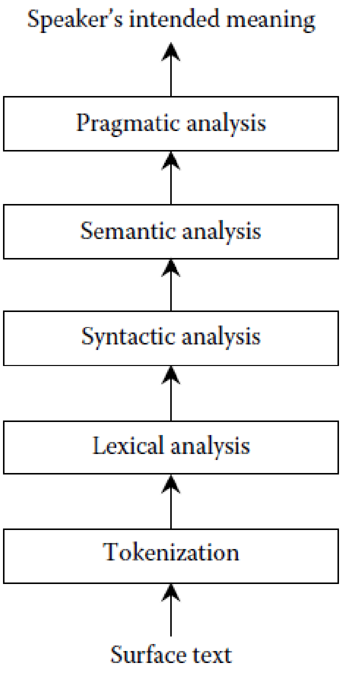


Figure 2.4: Knowledge Levels (Indurkhya and Damerau, 2010).

The **lexical or morphological** level enables the understanding of each word by humans or by NLP methods, by identifying each word as tokens (to have words as tokens as a task, called by tokenization), a classification was assigned to grammatical classes, such as the lemma or the radical. Thus, tasks associated with this level will break down the text into tokens, where each token will be identifying the likelihood of it being a word, a punctuation mark, digits, or other language signs. Then, a grammatical classifies each token's meaning as a verb, pronoun, or name (Indurkhya and Damerau, 2010) (Liddy, 2001).

The **syntactic** level determines the sentence's grammatical structure by establishing the relationship between words in a sentence, such as identifying what a specific adjective is classifying as a given name in a sentence. Next, a grammar or a parser is used to analyze the words in each sentence to discover their grammatical structure. Their output in this level of processing represents the sentence that reveals the structural dependency relationships between the words (Indurkhya and Damerau, 2010) (Liddy, 2001). For instance, the sentences (in the Portuguese language): *"O Inspector Silva persegue o suspeito."* and *"O suspeito persegue o Inspector Silva."* which differ by terms of syntax, yet convey entirely different meanings.

The **semantic** level aims to attribute possible meanings of a sentence by concentrating on the interactions among word-level meanings within a sentence. The tasks associated with this level aim to obtain a text in a formal language that a computer system can comprehend, or in other words, a text without ambiguity, since only then is it possible to represent it in a formal language (Indurkhya and Damerau, 2010) (Liddy, 2001).

Finally, the **pragmatic** level associates the relation between words and the context, which can be related to other approaches regarding language behavior, such as social science, linguistics, or anthropology. This level claims the context and historical knowledge, among other parties, influenced by the text's meaning. However, some issues could arise, such as syntactic ambiguity. Hence, we need to understand that the language does not consist only of isolated words or phrases. These grouped phrases, named by discourse, according to the pragmatics' interpretation, is influenced by its context (Jurafsky and Martin, 2000) (Khan et al., 2016).

2.3.1 NLP Tasks

This section describes the tasks that are included in any major NLP application. In the following paragraphs, we will describe briefly, as possible, all tasks related to to NLP.

Sentence Splitting

This task is focused on breaking up a text into sentences by detecting each sentence's begin and end, by identifying the **sentence boundaries** for selecting the longest meaningful sentences. This is a challenging task because there is not a standard approach to separate sentences in a text. In most cases, sentence splitting boundaries are punctuation marks, such as periods, question marks, or exclamation points. However, the questions and exclamations points are unambiguous markers for sentence boundaries. Other challenges arise from a known character, like the period ".", there is an ambiguity between a sentence boundary marker and a marker of abbreviations like Sr. (in Portuguese: Senhor) or Insp. (in Portuguese: Inspector).

The table 2.1 shows an example using an excerpt of a online news newspaper [7], and after been processed by sentence splitting task, the result is (delimited by tags <S> and </S>):

Tokenization

This task, also known as **word segmentation**, introduced the identification of a word on its atomic level. This identification is related to the variety of human languages and writing systems, like Portuguese

[7] See www.dn.pt [Accessed: 1 April 2020].

Paragraph	Este aumento de circulação - e apreensão de grandes quantidades de cocaína - não é inédito, mas é preciso recuar até aos anos de 2005/6 para encontrar picos semelhantes. Com uma diferença: nessa altura continuavam a circular muito mais outro tipo de drogas, como a heroína, que tem vindo paulatinamente a decrescer.
Sentence Splitting (Sentences)	<S>Este aumento de circulação - e apreensão de grandes quantidades de cocaína - não é inédito, mas é preciso recuar até aos anos de 2005/6 para encontrar picos semelhantes. </S> <S>Com uma diferença: nessa altura continuavam a circular muito mais outro tipo de drogas, como a heroína, que tem paulatinamente a decrescer. </S>

Table 2.1: Sentence Splitting Example

or English, that emphasizes the main types of dependencies, such as language dependence, character-set dependence, application dependence, and corpus dependence (Palmer, 2000). Therefore, this task accomplishes a splitting work over a stream of characters into words into a single "atomic" piece by detecting word boundaries, called **token**. The table 2.2 shows an example of what can be considered as the following sentence being subjected to tokenization that will result in the following *tokens* (delimited by square brackets):

Sentence	O Inspector Silva identificou o suspeito.
Tokenization (Tokens)	[O], [Inspector], [Silva], [identificou], [o], [suspeito], [.]

Table 2.2: Sentence Splitting Example

There are two main approaches to this task:

- **Space-delimited** languages or **segmented languages** approach (e.g., European languages, were words (tokens) and boundaries are indicated by the white-space insertion) (Palmer, 2000);
- **Unsegmented language** approaches, such as Chinese or Thai, where words (tokens) are written in succession with no indication for word boundaries, that require a surplus/an extra amount of additional lexical and morphological information (Palmer, 2000).

The Portuguese language fit on the segmented language approach where most tokens were framed by explicit separators, such as spaces or punctuation. Also, this task could be used as an initial step for identifying and recognizing the processes where entities or other symbols could be identified, as an approach used to perform the identifications that are the regular expressions (Regex):

- Email addresses (such as, email@email.com);
- Ordinal numbers (such as, 1º, 4º);
- Numbers with "." and "," (such as, 22.313,37);
- IP and HTTP addresses (such as, 127.0.0.1, http://www.google.com);
- Integers (such as, 12243);
- Abbreviations with "." (such as, "a.c.", "V.Exa.");
- Sequences of interrogation and exclamation marks;
- Ellipsis (such as, ???, !!!, ?!?!, ...);

- Punctuation marks (such as, ! ? . , : ; ( ) [ ] -);
- Symbols (such as, «, », +, -);
- Roman numerals (such as, LI, MMM, XIV).

Up to this point, each word (token) was detected in processed documents. Therefore, we could have applied several techniques and retrieve information from the analyzed documents, such as *Term Frequency* (*TF*), *Inverse Document Frequency* (*IDF*), *Term Frequency-Inverse Document Frequency* (*TF-IDF*) or using *Binary Representation*.

- **Term Frequency (*TF*)**: when a set of Portuguese documents were analyzed and wished to rank to the most relevant of the query, *"O suspeito Anibal vendeu estufacientes"*. Using an elimination process for the documents, these five words were not contained in these sentences. Therefore, each term's number occurs in each document; the number of times a term occurs in a document is called its term frequency. A definition quoted in (Sanderson, 2010): *"The simplest approach is to assign the weight to be equal to the number of occurrences of term **t** in document **d**...for a document **d**, the set of weights determined by the **tf** weights above (or indeed any weighting function that maps the number of occurrences of **t** in **d** to a positive real value) may be viewed as a quantitative digest of that document."*
- **Inverse Document Frequency (*IDF*)**: In documents, some terms are more common than others (e.g., term "de" is a widespread word). Therefore, giving enough weight to the more essential terms, e.g., "suspeito", "haxixe". Therefore, the term "de" is not the right keyword to distinguish relevant from non-relevant documents and terms, unlike the terms "suspeito", "haxixe". Inverse Document Frequency factor is incorporated, which diminishes the weight of the terms that frequently occur in the document set and increases the weight of rare terms.
- **Term Frequency - Inverse Document Frequency (*TF-IDF*)**: We could combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document, the weight is obtained by multiplying the two measures (Sanderson, 2010).

After applied to a document, the analyzed techniques could lead to a *bag of words* model, where the exact ordering of the terms in a document is ignored, but the number of occurrences of each term is accounted for. We only retain information on the number of occurrences of each term. Thus, the document *"João conhece o Artur"* is, in this view, identical to the document *"Artur conhece o João"*. Nevertheless, it seems intuitive that two documents with a similar bag of word representation are similar in content (Zhang et al., 2010).

Part-Of-Speech Tagging

This task enables the classification by its lexical category and syntactic behavior in a sentence and the semantic field in which the phrase is inserted, assigning appropriated and unique tags that indicate a syntactic role, such as noun, verb, adverb, pronoun, and others. The process that allows the classification of words in terms of part-of-speech is called **tagging** (Adhvaryu and Balani, 2015) (Weston and Karlen, 2011). The resulting outcome of Part-Of-Speech Tagging (POS) is a label, named as **POS Tag**, that is assigned to each token in a corpus which is part of speech and grammatical categories, forming a **tagset**, a group of POS tags (enumerated in `www.universaldependencies.org/u/pos/all.html`) that was used to label a corpus. The use of POS is fundamental in several NLP applications, such as speech recognition, speech synthesis, machine translation, information retrieval, or information extraction. In order to overcome the task, different computational methods were applied to improve classification accuracies, such as Brill Taggers (Brill, 1992), N-Gram Taggers (Prins, 2004), Hidden Markov Models (Ekbal et al., 2007), Braum-Welch Algorithm (Perez-Ortiz and Forcada, 2001) or Viterbi Algorithm (Forney, 1973). Table 2.3 shows an example of POS with the tokens and the corresponding POS tags.

Tokens	O	Luis Silva	assaltou	o	Banco	.
POS Tags	art	prop	v-fin	art	prop	punc

Table 2.3: POS Tagging Example.

There is another task associated to POS, named by stemming (also known as shallow parsing), that involves dividing sentences into non-overlapping segments based on the relatively simple superficial analysis that aims to identify the synthetic parts of a speech or text, labeling sentences as nouns or verbs phrases (NP or VP) as for other synthetic parts (Osborne, 2000) (Akerkar and Joshi, 2003).

Lemmatization

This process groups together different inflected forms of a word, so they can be analyzed as a single item (Brown, 1993). Each word is formed by different forms, for grammatical reasons, such as *"organiza"* or *"organizar"*, or with similar meanings such as *"democracia"*, *"democrático"*, and *"democratização"*. Therefore, the process detects relevant words or candidate terms, transforming the verbs (into their infinite form) and the names (into their normative form singular), by using vocabulary and the morphological analysis of words, supported by algorithms that determine the **lemma** of a word, based on its meaning, by using a POS task. Several lemmatization algorithms are proposed to obtain these results, though being considered time-consuming, error-prone, and somewhat tricky.

Tokens	O	Luis Silva	assaltou	o	Banco	.
POS Tags	art	prop	v-fin	art	prop	punc
Lemmas	o	luis silva	assaltar	o	banco	.

Table 2.4: Lemmatization Example.

Dependency Parsing

This task seeks the syntactic structure of a sentence described in terms of the words (or lemmas) in a sentence with an associated set of directed binary grammatical relations that hold among the words. To enable this task, **dependency parsers** were introduced to enable sentence syntactic analysis and built a syntactic structure according to a given formal grammar (Jurafsky and Martin, 2000). Among the relations established between words, a defined set of **dependency relations** is defined in the Universal Dependencies project (Nivre et al., 2016) that provides a list of dependency relations. Direct and labeled arcs connect **heads** to **dependents** among relations between words. This is called **dependency structure** because the labels are assigned from a list of grammatical relations. It includes a **root node** that explicitly marks the **root of the tree** and the head of the entire structure.

For the development and evaluation of dependency parsers, **dependency treebanks** were introduced by having a human factor and the manual annotation of the dependency structures for a given corpus. Therefore, several dependency treebanks were proposed, such as Prague Dependency Treebank (Hajic et al., 2001) or Penn Treebank (Taylor et al., 2003).

Figure 2.5 shows the tree, named as **dependency tree**, for the sentence *"A 35 year old woman was found murdered in a farm near Kosamada village in Kamrej district of Surat on Monday"* (Sunkara, 2019).

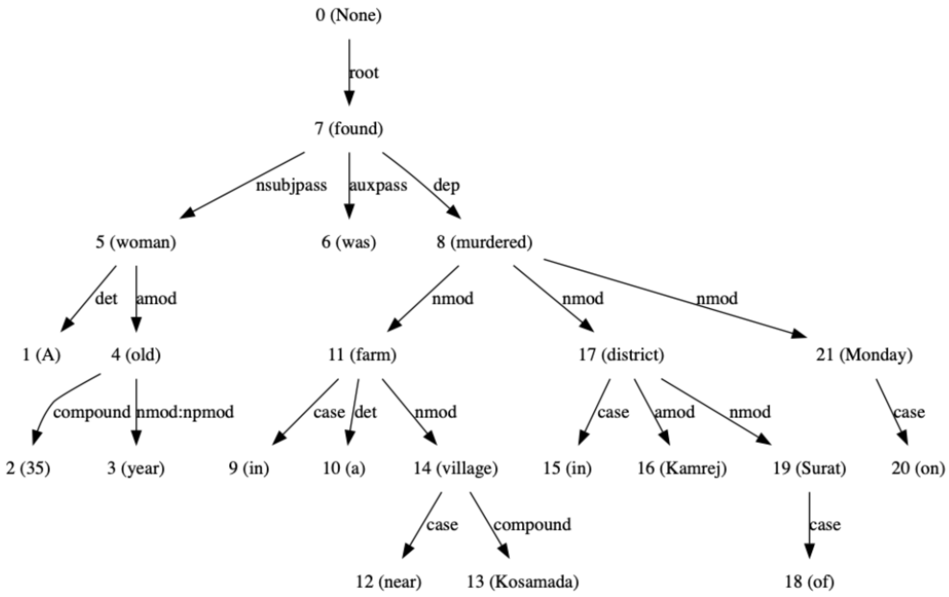


Figure 2.5: Dependency Tree Example (Sunkara, 2019).

Named-Entity Recognition

The NER is an essential element in any system whose purpose is to retrieve information from textual data. Therefore, Nadeau and Sekine (2006) defined NER as *"a subproblem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations, and companies."*. This task is used in other NLP domains, such as Sentiment Analysis, Machine Translation, Question Answering, Text Summarization, or Language Modelling.

In 1995, the Sixth Message Understanding Conference (MUC-6) coined a new concept, called NER, also known as Named-Entity Recognition and Classification (NERC), that aims to automatically recognize mentions of **named-entities**, such as persons, places, date and time or other entities extracted from structured and unstructured documents (Nadeau and Sekine, 2006). Since MUC-6, the NER task can be distinguished by three types of named-entities: ENAMEX (person, organization, and location), TIMEX (date and time), and NUMEX (money and percent).

The **named-entity** term identifies a word or phrase with common or similar attributes to describe textual references to real-world entities (McNamee et al., 2011). The word named refers to a distinct entity, typically by proper names or domain-dependent, such as crime-related entities, such as weapons.

Take a look to the following examples in Portuguese: *"Durante a investigação, a Policia Judiciária identificou duas suspeitos, João Pires e Alberto Cardoso , de assaltarem o Banco de Portugal. O assalto foi executado em Setembro."* when introduced in a NER system it should be able to identify five NEs. The following sentences has five NEs annotated.

> Durante a investigação, a Policia Judiciária $_{\textbf{Organization}}$ identificou dois suspeitos, João Pires $_{\textbf{Person}}$ e Alberto Cardoso $_{\textbf{Person}}$, de assaltarem o Banco de Portugal $_{\textbf{Organization}}$. O assalto foi executado em Setembro $_{\textbf{Date/Time}}$.

To develop the NER task, a broad spectrum of approaches were used, such as dictionary-based, rule-based, and machine learning approaches. When approaches have training samples, the learning methods are borrowed from supervised machine learning, such as Hidden Markov Models (Zhou and Su, 2002), Maximum Entropy Models (Curran and Clark, 2003), and Conditional Random Fields (Finkel et al., 2005) among others. But when we don't have training samples, we can take approaches like dictionary-based methods (Aronson, 2001), work by matching the text phrases with concepts that exist in the dictionaries; and rule-based methods (Eftimov et al., 2017), that apply regular expressions in a combination of information from external resources and characteristics of useful entities. For further information about the addressed topic, see (Nadeau and Sekine, 2006) (Wen et al., 2019).

Semantic Role Labeling

This task automatically labels the syntactic arguments, known as **semantic roles**, in a sentence. Thus, it was taking an **predicate** that refers to the main verb of the sentence in each predicate to distinguish arguments that are present and are semantically related to the predicate.

The semantic roles fit on the notion of lead *"Who did What to Whom, How, When and Where"* (Palmer et al., 2010). The arguments are Agent, Patient, and Instrument, and the adjunctive arguments of the predicate such as Locative, Temporal, and Manner (Jurafsky and Martin, 2000). The semantic arguments can be grouped into two categories: core arguments (Arg) and adjunctive arguments (ArgM), as shown in table 2.5.

Core	Adjunctive
V (Verb)	ArgM-DIR (Direction)
A0 (Subject)	ArgM-MNR (Manner)
A1 (Object)	ArgM-LOC (Location)
A2 (Indirect Object)	ArgM-TMP (Temporal marker)
A3 (Start point)	ArgM-PRP (Purpose)
A4 (End point)	ArgM-NEG (Negation)
A5 (Direction)	ArgM-REC (Reciprocal)
	AM-DIS (Discourse marker)

Table 2.5: Core and Adjunctive Arguments.

The semantic roles are the core of semantic role labeling. A predicate, represented by names, adjectives, adverbs, or more commonly verbs, establishes a meaningful relationship with its subject and its complements. That meaning can derive from the action, state, mental and relational events, among others.

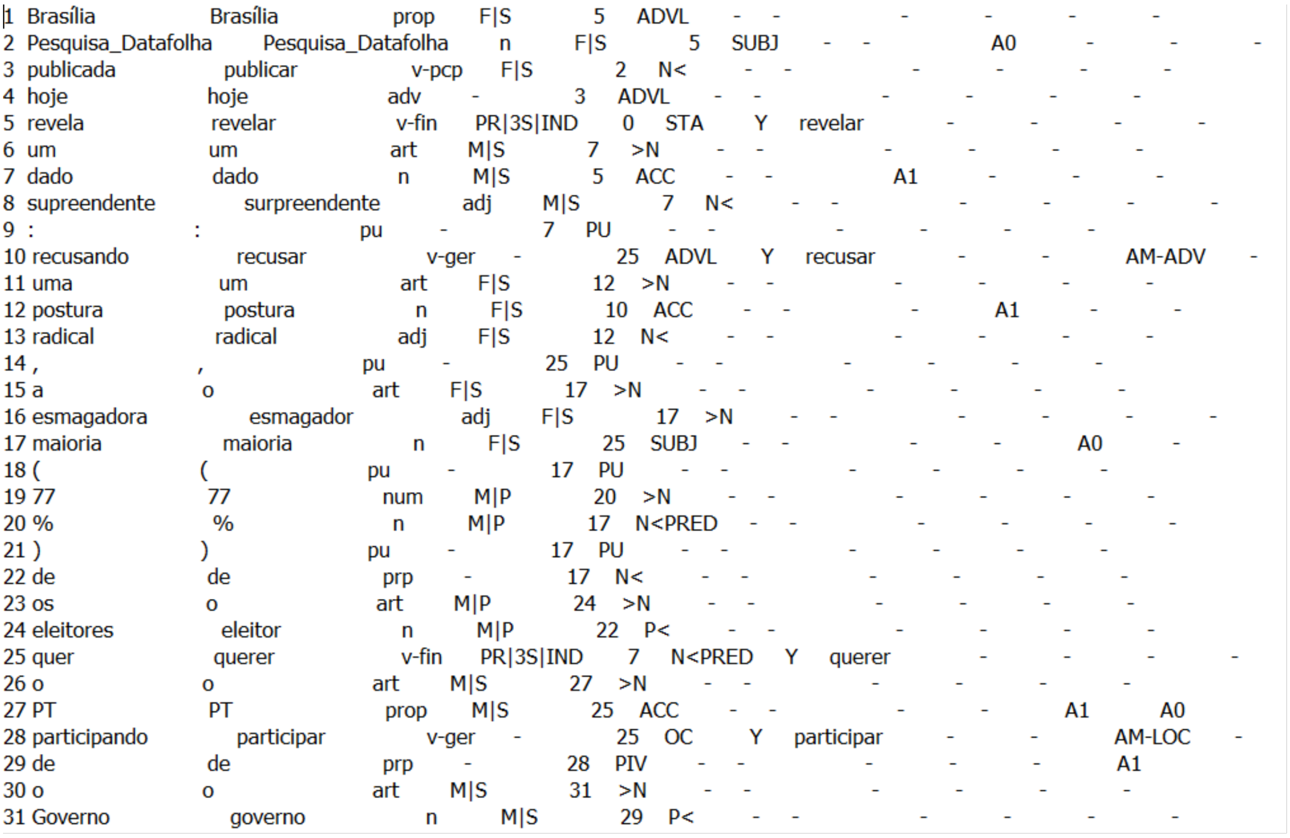
Resources like VerbNet [8], PropBank (Palmer et al., 2005), PropBank.Br (Duran and Aluísio, 2012) and FrameNet [9] identifies the verbs that realize semantic roles for their arguments in multiple ways, such as the case of the **VerbNet** groups verbs into semantic classes. **PropBank** contains sentences annotated with

[8] See verbs.colorado.edu/verbnet/ [Accessed: 1 July 2020].

[9] See framenet.icsi.berkeley.edu/fndrupal/ [Accessed: 1 July 2020].

verb-specific semantic roles, enumerated in table 2.5. **FrameNet** defined in terms of frames rather than verbs, where in every frame, the core roles and non-core roles are defined.

PropBank-Br is a Brazilian Portuguese treebank with semantic role labels following the Propbank guidelines, i.e., each verb has several core arguments like **Arg0-5** and there are **12 ArgM** tags, which include expressions like defining location, time, and others. Figure 2.6 shows an excerpt (retrieved from 143.107.183.175:21380/portlex/index.php/pt/projetos/propbankbr) from PropBank-Br, where we can see the semantic role tags.



```
1  Brasília           Brasília            prop   F|S         5   ADVL     -  -  -          -          -        -
2  Pesquisa_Datafolha Pesquisa_Datafolha  n      F|S         5   SUBJ     -  -  A0         -          -        -
3  publicada          publicar            v-pcp  F|S         2   N<       -  -  -          -          -        -
4  hoje               hoje                adv    -           3   ADVL     -  -  -          -          -        -
5  revela             revelar             v-fin  PR|3S|IND   0   STA      Y  revelar    -          -          -        -
6  um                 um                  art    M|S         7   >N       -  -  -          -          -        -
7  dado               dado                n      M|S         5   ACC      -  -  A1         -          -        -
8  supreendente       surpreendente       adj    M|S         7   N<       -  -  -          -          -        -
9  :                  :                   pu     -           7   PU       -  -  -          -          -        -
10 recusando          recusar             v-ger  -           25  ADVL     Y  recusar    -          -          AM-ADV   -
11 uma                um                  art    F|S         12  >N       -  -  -          -          -        -
12 postura            postura             n      F|S         10  ACC      -  -  -          A1         -        -
13 radical            radical             adj    F|S         12  N<       -  -  -          -          -        -
14 ,                  ,                   pu     -           25  PU       -  -  -          -          -        -
15 a                  o                   art    F|S         17  >N       -  -  -          -          -        -
16 esmagadora         esmagador           adj    F|S         17  >N       -  -  -          -          -        -
17 maioria            maioria             n      F|S         25  SUBJ     -  -  -          -          A0       -
18 (                  (                   pu     -           17  PU       -  -  -          -          -        -
19 77                 77                  num    M|P         20  >N       -  -  -          -          -        -
20 %                  %                   n      M|P         17  N<PRED   -  -  -          -          -        -
21 )                  )                   pu     -           17  PU       -  -  -          -          -        -
22 de                 de                  prp    -           17  N<       -  -  -          -          -        -
23 os                 o                   art    M|P         24  >N       -  -  -          -          -        -
24 eleitores          eleitor             n      M|P         22  P<       -  -  -          -          -        -
25 quer               querer              v-fin  PR|3S|IND   7   N<PRED   Y  querer     -          -          -        -
26 o                  o                   art    M|S         27  >N       -  -  -          -          -        -
27 PT                 PT                  prop   M|S         25  ACC      -  -  -          -          A1       A0
28 participando       participar          v-ger  -           25  OC       Y  participar -          -          AM-LOC   -
29 de                 de                  prp    -           28  PIV      -  -  -          -          -        A1
30 o                  o                   art    M|S         31  >N       -  -  -          -          -        -
31 Governo            governo             n      M|S         29  P<       -  -  -          -          -        -
```

Figure 2.6: Propbank.Br Excerpt.

Table 2.6 describes the annotations of PropBank-Br label each of the arguments of a predicate verb with argument numbers, where the core arguments are labeled Arg_0, Arg_1, ... Arg_4. In addition, adjunctive arguments are labeled AM-... and of these, this system used *AM-TMP* (temporal), *AM-LOC* (locative), or *AM-NEG* (negation).

The SRL is useful for diverse applications, such as machine translation, information extraction, text summarization, and question and answering systems.

2.3.2 Information Extraction

Information extraction points to the *"automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources"* (Sarawagi et al., 2008). Therefore, obtain structured information from unstructured text has been an exciting challenge

Label	Arguments	Tokens
A0	3.064	11.225
A1	5.181	34.378
A2	1.103	6.585
A3	114	571
A4	75	362
AM-ADV	369	1.675
AM-CAU	156	1.244
AM-DIR	15	47
AM-DIS	294	843
AM-EXT	81	184
AM-LOC	764	3,648
AM-MNR	408	1,955
AM-NEG	337	358
AM-PNC	171	1.377
AM-PRD	192	1.293
AM-REC	65	73
AM-TMP	1.125	4.318
X	–	59.939

Table 2.6: PropBank-Br statistic. "X" for non-verb, non-argument tokens (Fonseca and Rosa, 2013).

to the research community in the last decades. The main objective of **information extraction** is to automatically extract useful information from unstructured data by performing different techniques, such as **term or topic** extraction, extraction of NEs and relations among these entities, and **events** extraction in which these entities participate. We remained above a short introduction to the most common methods used for Information Extraction (IE).

- **Term or topic**: extraction has the goal of automatically extracting relevant terms from a given corpus (Alrehamy and Walker, 2017). This extraction is important to the model of a specific knowledge domain to collect a vocabulary of domain-relevant terms, constituting the linguistic surface manifestation of domain concepts. Different approaches were applied to automatically extract term from domain-specific documents, making use of NLP tasks, such as POS or phrase chunking, to extract term candidates. The evaluated terms are noun phrases that include compounds, like "credit card";

- **Events**: extraction from unstructured data such as the news from online newspapers is a common task in IE systems. This extraction seeks to answer the *"Who did What to Whom, How, When and Where"* questions, finding events with their participants. Also, time and location are extracted. For extraction, different approaches were taken by using data-driven, knowledge-driven, and hybrid event approaches. Event extraction has many applications, such as link analysis, geospatial analysis, and biographical information, useful for the criminal domain;

- **Relationships** extraction: can be defined among multiple entities (*n-ary*). Of course, binary relationships are also consider, they identify relationships between two entities, in a form of triple $<ent_1$, *relation*, $ent_2>$, where ent_1 and ent_2 are NEs or noun-phrases in a sentence, and **relation** is a relationships type that connect the arguments, such as the affiliation relationship between a person and organization - *"Peter **works for** Google"*.

2.3.3 Linguistic Corpus

The development from scratch or the purposing of an existing linguistic corpus is crucial to study the human (natural) language combined with computer methods by selecting or grouping a well-defined set of texts regarding a specific domain. Thus, Sinclair (Sinclair, 2005) defined corpus as *"...a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research"*. Therefore, a corpus building's focus is to define a collection of machine-readable texts in a natural language that can be used as a representative sample with particular factors, such as newspaper articles, literary fiction, or other relevant documents. The corpus content must be "representative of a language variety" (Leech, 1992). In the criminal domain, if the corpus content represents the specifications of linguistic phenomena, if we can extrapolate to a more significant population from which it is taken, then we can say that it "represents that language variety.". Atkins et al. (1992) propose elaborating attributes that can be used to define the different text types, and contribute to creating a balanced corpus.

In a typical automatic classification task, like NER, accurate NE detection requires a good training set. In a training set, all the terms that represent desired entities should be labeled appropriately. Usually, the labeling process is a laboring and error-prone task.

The annotated (manually) corpus is often divided into three distinct parts:

- **Training corpus**: a portion of annotated corpus set aside to train the system;
- **Development corpus**: a portion of annotated corpus set aside to tune the training configuration, to used for error analysis and parameter tuning;
- **Test corpus**: a portion of the annotated corpus set aside to the system be evaluated.

The *Floresta Sintatica* corpora, is a set of the corpus (morph) syntactically analyzed and structured in a format that resembles a tree, named by Lying Trees (AD), like the ones we see in the figure 5.1. Each tree set constitutes a treebank. The analyzed corpora are as described below:

- the **Amazónia** corpus contains 485 million words (about 275000 sentences) extracted from a collaborative website Overmundo, in Portuguese-Brazilian language. All texts were collected in September 2008;
- the **Selva** treebank contains about 400000 thousand words (about 28100 sentences) divided between different textual genres and the Portuguese and Portuguese-Brazilian language variants. The treebank is subdivided into three corpora: Selva Falada, Selva Científica, and Selva Literária;
- the **Floresta Virgem** treebank is composed of 95000 sentences (about 1640000 words), joining the two corpora: CETENFolha corpus, built from texts of the Brazilian newspaper "Folha de São Paulo" (1994); and the CETEMPublico corpus, taken from texts of the Portuguese newspaper "Publico" (1991 to 1998);

The described corpora are available in other formats, like AD in VISL format, CG (automatic), CG (v7.4, manually review), PennTreebank, Tiger-XML, SQL, SimTreeML, and Perl. All the project material (Bosque, Selva, Amazônia e Floresta Virgem) was analyzed automatically by the PALAVRAS Bick (2000) system (discussed in chapter 3). However, the Selva corpus has partially reviewed by linguistics da Silva Teixeira et al. (2008). Table 2.7 resumes the analyzed corpus.

Corpus	Genre	Source	Words	Sentences	Language
Floresta Virgem	Jornalistic	Folha de São Paulo and Público newspapers	1640000	96000	Portuguese Brazilian
Amazónia	Opinion	Overmundo blog	4580000	275000	Brazilian
Selva	Spoken Scientific Literacy	Interviews, books and scientific articles	400000	281000	Portuguese Brazilian

Table 2.7: Floresta Sintática Corpora (exclude the Selva Corpus).

2.3.4 Language Resources

The **language resources** refer to sets of electronic speech or language data in machine-readable form, used to support research and applications in the NLP field. Typically, such data are annotated with linguistic information such as morpho-syntactic categories, syntactic, or discourse structure. These resources are used for building, improving, or evaluating NLP and speech algorithms or systems (Lezcano et al., 2013).

There are official sources of open data [10] with a list of language resources available for the Portuguese language, being part of a task across the European Union to connect legal and other document types. The linguistic information has several resources:

- **Glossary**: is a lexical list of specialized terms and meanings, such as *Glossary of Abbreviations proposed by UNESCO* [11];
- **Lexicon**: are networks of information about words and their contexts, usual groups of synonyms that provide definitions and relations between them, such as *openWordnet-PT* [12];
- **Thesaurus**: it is a controlled list of semantically and generically related terms that cover a specific domain of knowledge hierarchically structured, such as *Eurovoc* [13];
- **Dictionary**: a record that organizes a language in the form of lemmas an alphabetical order and where their meaning is explained;
- **Gazetteer**: its a set of lists containing names of entities such as persons, locations, or organizations. These lists are used to find similar terms in a corpus, applied in tasks like NER.

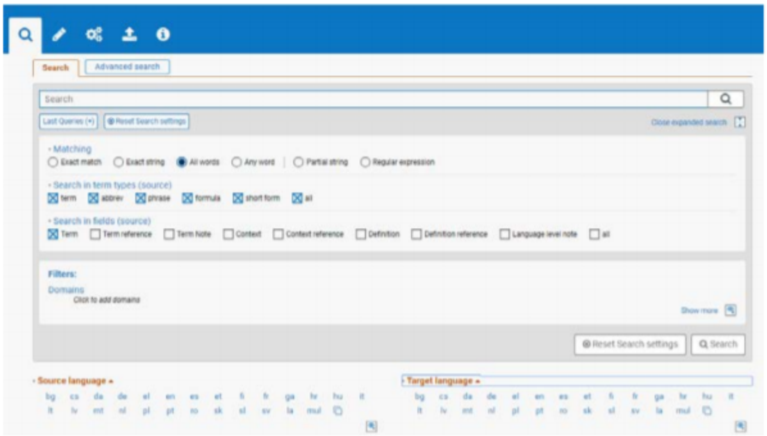
Interactive Terminology for Europe (IATE)

IATE is an multilingual terminological database of the European Union [14], available through an open-access platform [15] and it is a terminological reference for translators and language users in Europe. This database is maintained by translators and terminologists of the language services of the European Union institutions. However, IATE shared responsibility of all EU institutions and bodies involved in the project (European Parliament, Council of the EU, Commission, Court of Justice, Court of Auditors, European Economic

[10] See data.europa.eu/euodp/pt/home [Accessed: 1 July 2020].
[11] See en.unesco.org/ [Accessed: 1 July 2020].
[12] See github.com/own-pt/openWordnet-PT [Accessed: 1 July 2020].
[13] See data.europa.eu/euodp/pt/data/dataset/eurovoc [Accessed: 1 July 2020].
[14] See europa.eu/european-union/index_pt [Accessed: 1 July 2020].
[15] See iate.europa.eu/ [Accessed: 1 July 2020].

and Social Committee, Committee of the Regions, European Central Bank, European Investment Bank, Translation Centre), hosted in European Commission at Luxembourg. The database is available in 26 languages (including Portuguese language), and is exported to TermBase eXchange (TBX) [16] and Comma-Separated Values (CSV) [17] formats.

The IATE database is divided into domains, such as the financial crisis, the environment, fisheries, and migration. For domain classification system uses the *EuroVoc* thesaurus. Figure 2.7 displays the IATE search interface for browsing terms.



Matching
○ Exact match ○ Exact string ● All words ○ Any word | ○ Partial string ○ Regular expression

Search in term types (source)
☒ term ☒ abbrev ☒ phrase ☒ formula ☒ short form ☒ all

Search in fields (source)
☒ Term □ Term reference □ Term Note □ Context □ Context reference □ Definition □ Definition reference □ Language level note □ all

Filters:
Domains
Click to add domains

Search
Advanced search
Search
Last Queries (•)
Reset Search settings
Close expanded search
Show more
Reset Search settings
Search
Source language
bg cs da de el en es et fi fr ga hr hu it
lt lv mt nl pl pt ro sk sl sv la mul
Target language
bg cs da de el en es et fi fr ga hr hu it
lt lv mt nl pl pt ro sk sl sv la mul

Figure 2.7: IATE Search Interface.

IATE is accessible by search API that ensures the access to data through programming tools. It is also possible to download a subset of data related to a specific request, such as linguistic research.

Listing 2.1: IATE file in CSV format (excerpt).

```
E_ID|E_DOMAINS|L_CODE|T_TERM|T_TYPE|T_RELIABILITY|T_EVALUATION

73180|LAW|en|retract|Term|Reliable|
75181|LAW|en|special relationship|Term|Reliable|
1131014|LAW|pt|caixa principal|Term|Reliable|
1131014|LAW|en|chief cashier|Term|Reliable|
1390716|LAW|pt|direito comercial|Term|Reliable|
1390716|LAW|en|commercial law|Term|Reliable|
```

Listing 2.1 shows an excerpt of IATE download file, divided into several columns, such as:

- E_ID: identifies using a unique ID the term selected;
- E_DOMAINS: identifies the domain, or several domains, with its sub domains;
- L_CODE: defines the language code, such as “pt” for Portuguese language;
- T_TERM: term name;
- T_TYPE: identifies the term type, such as term, abbreviation, appellation, short form, phrase, formula, and lookup;

[16] See www.iso.org/standard/62510.html [Accessed: 1 July 2020].
[17] See tools.ietf.org/html/rfc4180 [Accessed: 1 July 2020].

- T_RELIABILITY: ensures the reliability of the term, such as 1 to 4 stars;
- T_EVALUATION: ensures the evaluation, such as preferred, admitted, obsolete, deprecated, proposed.

2.3.5 NLP Conferences

In order to test the applications developed within the scope of the NLP, over the years, evaluation conferences have been introduced, such as Message Understanding Conference (MUC) (Hirschman, 1998), Automatic Content Extraction (ACE) [18] and Conference on Natural Language Learning (CoNLL) [19]. These evaluations permit a common task definition, data sets, and evaluation metrics; and are open to any interested research group participation. Table 2.8 enumerated the major international evaluations conferences of NER systems.

Conference Name	Year	Language(s)
MUC-6	1995	English
MUC-7	1997	English
CoNLL	2002	Dutch, Spanish
CoNLL	2003	English, German
ACE	2005	Arabic, Chinese, English
HAREM 1	2006	Portuguese
ACE	2007	Arabic, Chinese, English, Spanish
ACE	2008	Arabic, English
HAREM 2	2008	Portuguese

Table 2.8: NER Evaluations Conferences [adapted from (McNamee et al., 2011)].

The *HAREM* is an evaluation campaign of the NER systems organized by Linguateca [20] applied to the Portuguese language (Santos et al., 2006). To evaluate the works presented to *HAREM*, a set of directives was established, called by *HAREM Evaluation Directives*. These guidelines represent a set of scores, rules, and measures used to compare the systems' outputs against a text formally prepared for comparison and extract evaluation measures. The text is called *Golden Collection* (built by the community). Over the years, two major joint assessment events have been held:

- First HAREM: started in September of 2004, and joint evaluation in February of 2005. The *Golden Collection* consists of 89241 words, and within this group, there were recognized around 3851 named entities;
- Second HAREM (Carvalho et al., 2008): started in September 2007 and ended in September 2008. The *Golden Collection* were built from 129 documents. In the evaluation, it was possible to choose the categories, types, and subtypes or other attributes (see figure 2.8) that the systems intend to label and be evaluated according to the interest that these labels may have in the context of other developed applications.

[18] See www.ldc.upenn.edu/collaborations/past-projects/ace [Accessed: 1 April 2020].

[19] See www.conll.org/ [Accessed: 1 May 2020].

[20] See www.linguateca.pt/ [Accessed: 1 May 2020].

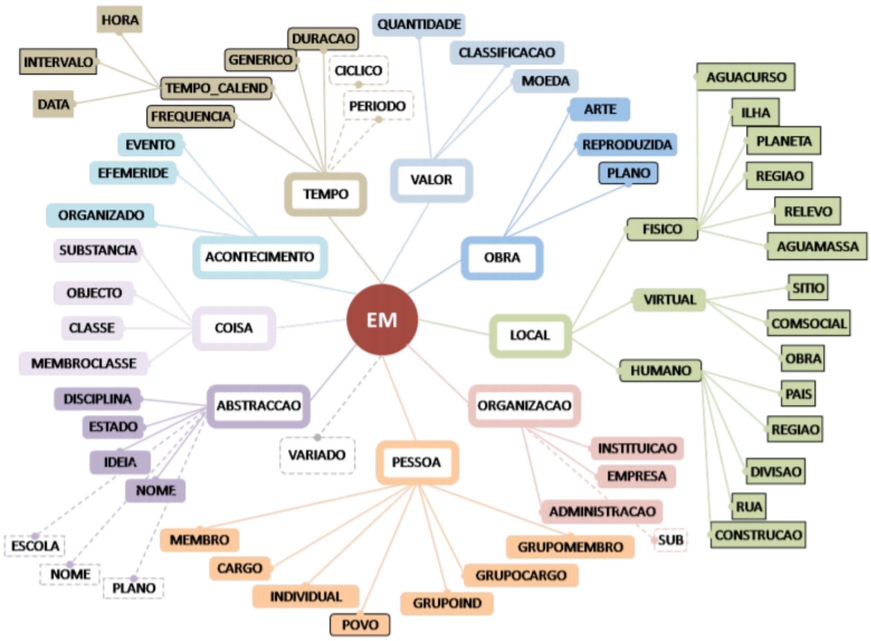


Figure 2.8: Second HAREM Categories Tree (Carvalho et al., 2008).

2.4 Evaluation Measures

This section introduces the metrics to analyzed the NLP tasks in terms of performance as an important step to benchmark the performance results of new proposals. To evaluate the results obtained by an NER system (for example), several performance measures where established, namely *accuracy*, *Precision*, *Recall* and *F-Measure*.

The term *accuracy* (*A*) refers to the percentage of correct labels out of the total label amount. Sometimes, we are also interested in performance improvements for a specific label y. For a given label y, we can thus measure the *precision* and *recall*, where *precision* indicates, for all of the contexts automatically labeled with y, the percentage where y was the correct label in the test corpus, and *recall* indicates, for all of the contexts where y is the correct label in the test corpus and the percentage of that was automatically labeled with y (Robertson, 2000).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Therefore, **precision** (*P*) is defined by the ratio of correct answers (True Positives) among the total answers produced (Positives), where *TP* - *True Positive*, a predicted value was positive, and the actual value was positive and *FP* - *False Positive*, predicted value was positive, and the actual value was negative.

$$P = \frac{TP}{TP + FP} \quad (2.2) \qquad R = \frac{TP}{TP + FN} \quad (2.3) \qquad F1 = 2 * \frac{P * R}{P + R} \quad (2.4)$$

The **recall** is defined as a ratio of correct answers (True Positives) among the total possible correct answers (True Positives and False Negatives), where *FN* - *False Negative*, a predicted value was negative, and the actual value was positive. Finally, the **f-measure**, also know as *F1*, is the harmonic mean of precision and recall. The figure 2.9 shows the Instances used to solve the functions listed above, and their connection with the classes (relevant and retrieved).

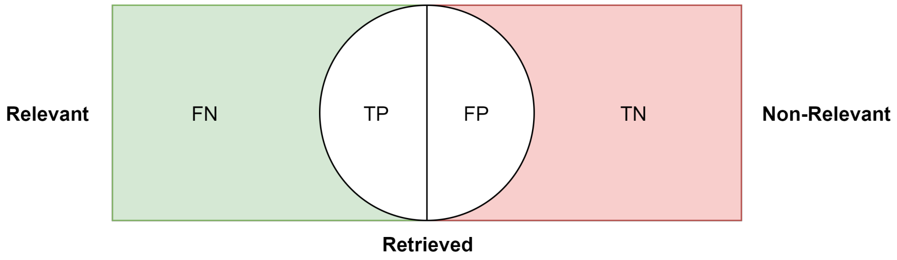


Figure 2.9: Evaluation Measures - Instances (FN, TP, FP, and TN).

The **confusion matrix** is a specific table layout that allows visualization of the performance of an algorithm, such as supervised learning. Each row of the matrix represents the instances in a relevant class, while each column represents the instances in a retrieved class (or vice versa). The table 2.9 shows the instances matched with relevant/non-relevant and retrieved/non-retrieved classes. The **Relevant** means all instances that are important to the domain, **Non-Relevant** means all instances that are not important for the domain, and **Retrieved** means all instances that are retrieved during an information extraction task.

	Relevant	**Non-Relevant**
Retrieved	true positive (TP)	false positive (FP)
Not-Retrieved	false negative (FN)	true negative (TN)

Table 2.9: Confusion matrix

2.5 Machine Learning Methods

Nowadays, the **machine learning** is a "fashionable" concept with references on social media, with multiple definitions. For example, Mitchell et al. (1997) defined machine learning as *"the study of computer algorithms that improve automatically through experience"*. In NLP, several tasks used these methods, like POS, NER, or dependency parsing, where sentences are modeled as sequences of tokens, and usually, a **sequence classifier** is used. This classifier is defined as a model used as input, a sequence of sentences, and outputs the sequence of labels that best classifies each input. POS tags, NE classes, and chunks are examples of the labels used in sequence classifiers.

We have different learning methods, such as supervised, semi-supervised, unsupervised learning, and distantly supervised. These methods differ in approach, data type, input and output, and task type or problem that intended to solve. The **supervised learning** models are a set of data containing the inputs and the desired outputs, known as **training data**. These training data are used to classify an input text, depending on the NLP task at hand (Russell and Norvig, 2002). The **unsupervised learning** uses a set of data that are not labeled or classified and tries to find a structure of the observed data, like a clustering of data points. This approach is based on the existence or non-existence of a pattern on each data set. Thus, this approach excludes a training phase, and it is applied to text data without manual effort being necessary. The **semi-supervised** is a blended approach to supervised (without labeled training data) and unsupervised learning (with labeled training data). The **distantly supervised** relies on a Knowledge Base (KB) to collect examples of the relationship we want to extract. This approach is often used in extraction tasks (Lockard et al., 2018). The rule-based or dictionary-based approaches used in machine learning methods apply hand-crafted or curated rules to manipulate the input data as a human intelligence mimicking process. Identifying and utilizing a set of rules that represent the domain knowledge captured by the system can be considered necessary to enable this. It is necessary for a set of rules or knowledge bases that will set up the prediction model.

The ML have a vast number of learning methods, we give a brief description of the methods used on this thesis, the statistical (see section 2.5.1), and the neural networks (see section 2.5.2) methods.

2.5.1 Statistical Methods

In order to review the statistical methods that during the thesis are referenced and used in the elaboration of the same, we summarize in figure 2.10 some of the methods.

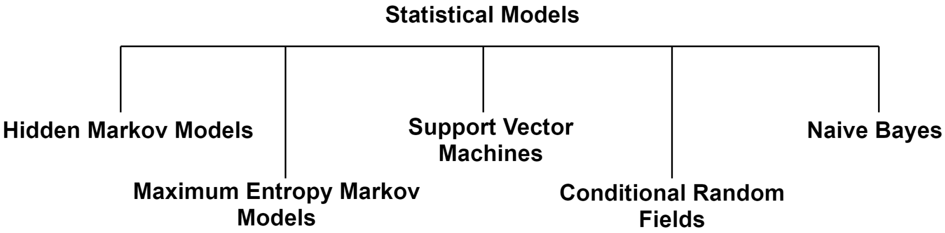


Figure 2.10: Statistical Methods (Summary)

We start by presenting the **Hidden Markov Models (HMMs)** as a probabilistic model for sequential data modeling, which can be used to decode labels of a sentence, used to calculate and assign the hidden events, such as tags, from observable events or as the words of a sentence (Eddy, 1996) (Rabiner, 1989). Therefore, the model is based on two concepts, observations, and hidden states. They take a sentence as observed data and the states by representing semantic information related to the sentence. The HMMs parameters are configured to achieve the log-likelihood maximization between the sentence and the semantic information. The obtained states will be mapped into the semantic tags, which will then generate a learning classifier. This classifier's performance depends on the choice of states and their corresponding observations, which can be evaluated by the best state sequence using the Viterbi algorithm (Forney, 1973). This model is used in NLP tasks such as POS, text segmentation, and voice recognition.

Maximum Entropy Markov Models (McCallum et al., 2000), also known as the Maximum Entropy Model (MaxEnt) is a discriminative model that aims to maximize the data, as to generalize as much as possible, for the training data (Sharnagat, 2014), by combining features of HMMs and **Maximum Entropy Models** (Van Campenhout and Cover, 1981). This model observes input data and calculates the probability of the possible sequences according to the observable events.

Support Vector Machines (SVM) Models introduced by (Cortes and Vapnik, 1995), based on the idea of learning a linear hyperplane that separates the positive examples from negative examples by a large margin. Furthermore, a large margin suggests the distance between the hyperplane and the point from either instance at its maximum. This model is used in handwritten digit recognition, object recognition, image face detection, and text categorization.

Conditional Random Fields (CRFs) Models was introduced by (Lafferty et al., 2014), as a statistical modeling tool for pattern recognition and ML using structured prediction, like the MaxEnt. In most classifiers, the prediction of a label for a single sample, without considering the "neighboring" samples, must take this model into account within its context. The prediction is modeled as a graphical model that implements dependencies between the predictions to perform this task.

The **Naïve Bayes** classifier is based on the Bayes theorem (Stone, 2013), as a Bayesian classifier (statistical classifier), by predicting the probability of class membership, such as the probability that any sample belongs to any given class. Furthermore, a naïve bayesian classifier assumes that the effect of an attribute value on a given class operates independently of the values of other attributes. This assumption is referred to as class conditional independence.

2.5.2 Neural Networks

Neural Networks are based on an imitation of human neurological function, like neurons. They use a training dataset for learning and then apply it to generalize patterns for classification and prediction. Therefore, a neural network should predict new observations from other observations (the training dataset).

The **Perceptron** algorithm (Rosenblatt, 1958) is a single layer neural network used in supervised learning to classify a given input data. For example, in a NER task to identify and classify NEs using a training corpus. It is divided into four parts: Input values or One input layer, Weights and Bias, Net sum, and Activation Function. Figure 2.11 shows the perceptron schema.

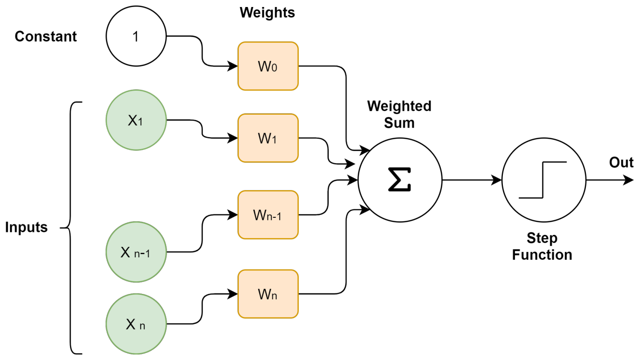


Figure 2.11: Perceptron Diagram.

The algorithm is a step by step evaluation, where the inputs are multiplied with weights *W*, then adds all the multiplied values as *Weighted Sum*, applying the *Weighted Sum* to the correct *Activation Function* (Sharma, 2017) (defines the output of the perceptron neural network like yes or no). At each step, it ensures that the current parameters classify the training example correctly. If done correctly, then it proceeds to the next example. If not, it moves the weight vector and bias closer to the current example. Thus, the algorithm loops over the training data until no further updates are available or a maximum iteration that could be counted and reached.

2.6 Graph Databases

Graphs are a way to represent information by binding their architecture to model situations where relationships among entities hold great relevance. Graph databases inherit concepts from a mathematical field, called **graph theory**. The problem of the seven *Königsberg bridges* was that the initial motivation for the appearance of the graph theory consisted of going through all the bridges passing only once in each one, by taking a better path across the seven bridges (Gribkovskaia et al., 2007). Figure 2.12 shows the Königsberg bridge problem.

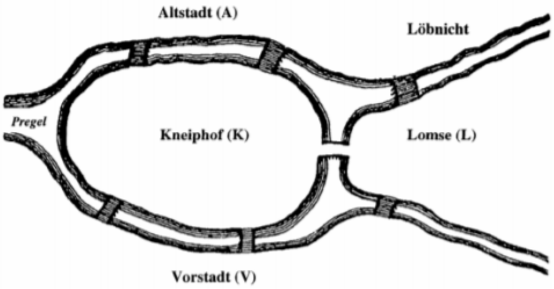


Figure 2.12: The *Königsberg Bridge* Problem (Gribkovskaia et al., 2007).

A **graph database** is a *directed multigraph* [21] $G = (V, E)$ *where every node* $v \in V$ *and every edge* $e \in E$ *is associated with a set of pairs key/value called properties, that are assumed to be finite.*. There are **undirected** graphs where all the edges are bidirectional, and **directed** graphs (or **digraph**) that are defined by edges with a direction associated with them (Ruohonen, 2013) (Jungnickel and Jungnickel, 2005).

A multigraph, also known as **property graph**, is where both nodes and edges are labeled with data in the form of key-value pairs. For example, figure 2.13 represents a property graph as being directed, labeled, attributed, and multigraph. The edges are directed and labeled. Edges have been associated with key/value pair value (i.e., properties).

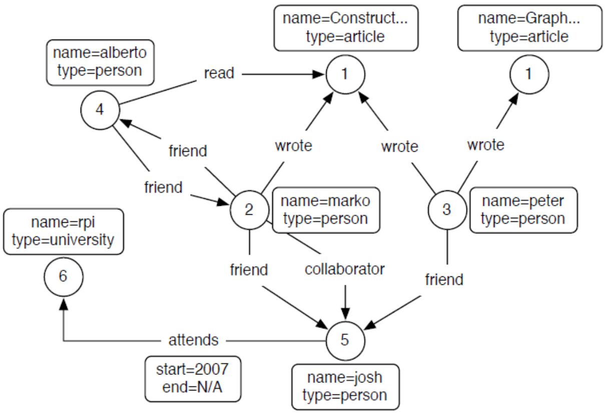


Figure 2.13: Property Graph Representation (Golomb and Taylor, 1984)

This **graph data model** provides us with the following components:

- **Nodes**: is used for the representation of entities information;

[21] is a graph where two nodes can be connected by more than one edge.

- **Relationships**: are used to link nodes and provide a structure to the graph. It is defined by a single name, with a start and end node, with a direction. Therefore, the joint of name and direction provide semantic information to the structure of the nodes;

- **Properties**: are set in the form of name-value pairs. Properties can be added to both nodes and relationships. Adding properties in relationships is used to add metadata and semantics and introduce constraints at runtime queries;

- **Labels**: are used for assigning roles or types to nodes.

In a database context, vertices are referred to as **nodes**, and edges as **relationships**. GDB is an alternative to relational databases (Rodriguez and Neubauer, 2010). The relational model was proposed in 1960s (Codd, 2002) in which rows and columns define tables. A row can be seen as an object, while columns would be attributes/properties of that objects (Rodriguez and Neubauer, 2012). However, the relational model is limited when the need is only to represent the semantics (Hull and King, 1987); when we faced big data problems that could be considered complex, interconnected information increases by bringing issues related to storing, retrieving, and manipulating such data. In this case, it becomes onerous to use relational models for a more detailed comparison between this two approaches (Vicknair et al., 2010). GDB are scheme-free data models, which enable the adaptation of the new data (Miller, 2013). After this, GDB can be considered an effective tool for modeling data when the focus of the relationship between entities is a driving force in designing a data model.

A GDB model has a specific set of values, such as data format and the processing methods showed in figure 2.14.

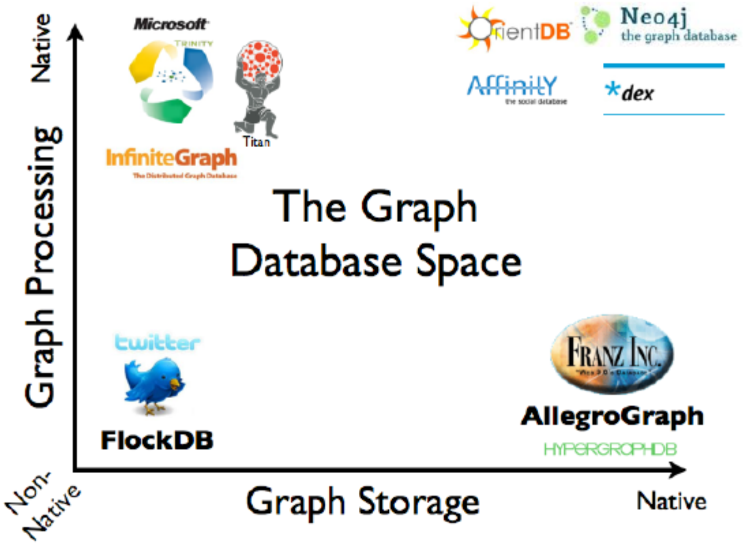


Figure 2.14: Graph Database Spectra (Robinson et al., 2013).

Each graph model has a data format on a spectrum between non-native and native storage by representing a graph. A **non-native storage** converts nodes and edges to relational database tables or another format, such as document-based, where the graph is the storage mechanism. It is useful when a database is extensive . **Triples** is a non-native storage model that consists of a subject, predicate, and object (Vicknair et al., 2010), that have a simplified storage version, namely the Resource Description Framework (RDF) [22]. The **labeled property** graph is a native storage and processing model that have a significant difference between triples, and this model is the ability to store data in nodes, allowing for index-free traversal (Vicknair et al., 2010). The *Neo4j* database is an example. The processing method handles Create, Read, Update, Delete (CRUD) operations (Robinson et al., 2013).

For more in-depth information about of addressed topic in this section, see surveys (Angles and Gutierrez, 2008) (McColl et al., 2014).

Neo4j [23] is an open-source graph database developed by Neo Technology. Released as a commercial variant, licensed under GPLv330 and AGPLv3. *Neo4j* was proposed to enable a data storage that realizes the index-free adjacency property, with full ACID (Atomicity, Consistency, Isolation, and Durability (Haerder and Reuter, 1983)) transaction support, and REST server interfaces. It offers theGo, Groovy, Clojure, Java, JavaScript, PHP, Phyton, Ruby, and Scala programming languages. It is the most widely used product among eBay, LinkedIn, Info jobs, and Walmart.

Figure 2.15 shows the basic concepts in *Neo4j*, in which nodes, relationships, and properties represent the data model. A relation connects two nodes with a mandatory type and can be optionally directed. The **properties** are **key-value pairs** coupled to nodes, and relationships. **Labels** are used to identify the domain by labeling each node, which groups nodes as sets, such as persons or organizations.

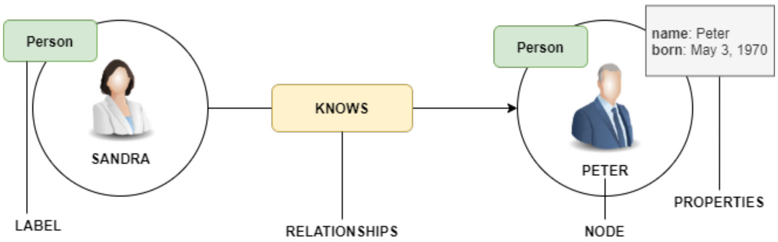


Figure 2.15: Neo4j Basic Concepts.

This database has a **logical data organization** that uses a property graph model that directly represents the basic elements of a graph model (nodes, relationships, and properties). **Physical data organization** is organized as layers, see figure 2.16.

Regarding the **data integrity**, some unique property constraints can be applied to ensure that the property values are unique with a specific label. However, unique constraints do not mean that all nodes have to have a unique value for the properties because nodes without the property is not subject to this rule.

Figure 2.17 shows the *Neo4j* interface that permits to visualize and perform data manipulation operations. It also allows us to query the database and visualize the results and visually navigate through the graph, in the center of figure 2.17, a graph with multiple nodes, relationships, and properties on which the graph databases are built on.

For **data modeling and manipulation**, the *Neo4j* uses **Cypher** as a declarative query language for graph manipulation. It is similar to the Structured Query Language (SQL), and it is a declarative query language that enables users to state actions to perform graph data without complex queries to describe exactly how to do it. Remember that a query language is a collection of operators and inference rules, intending to manipulate and query data in those structures in any combinations, by the user's demand, such as data

[22] See `https://www.w3.org/RDF/` [Accessed: 1 July 2020].

[23] See `neo4j.com/` [Accessed: 1 July 2020].

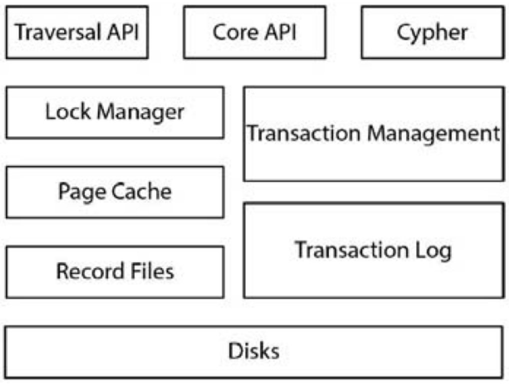
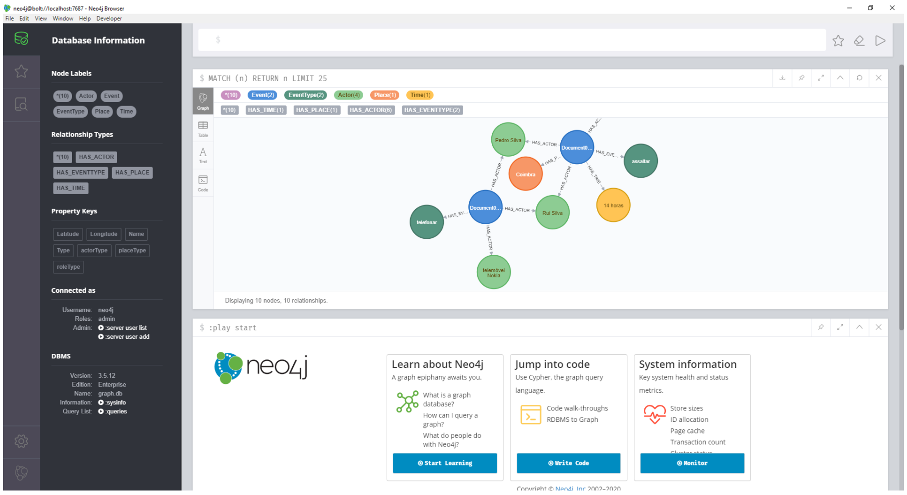


Figure 2.16: Neo4j Physical Data Architecture (Robinson et al., 2013).


neo4j@bolt://localhost:7687 - Neo4j Browser
File Edit View Window Help Developer
Database Information
Node Labels
*(10) Actor Event EventType Place Time
Relationship Types
*(10) HAS_ACTOR HAS_EVENTTYPE HAS_PLACE HAS_TIME
Property Keys
Latitude Longitude Name Type actorType placeType roleType
Graph
Table
Text
Code
*(10) Event(2) EventType(2) Actor(4) Place(1) Time(1)
*(10) HAS_TIME(1) HAS_PLACE(1) HAS_ACTOR(6) HAS_EVENTTYPE(2)
Pedro Silva
Document0...
assaltar
Coimbra
14 horas
Rui Silva
telefonar
telemóvel Nokia
HAS_ACTOR
HAS_EVE...
HAS_TIME
HAS_P...
HAS_EV...
neo4j
Start Learning
Write Code
Monitor

Connected as

Username: neo4j
Roles: admin
Admin: :server user list
:server user add

DBMS

Version: 3.5.12
Edition: Enterprise
Name: graph.db
Information: :sysinfo
Query List: :queries

```
$ MATCH (n) RETURN n LIMIT 25
```

Displaying 10 nodes, 10 relationships.

```
$ :play start
```

Learn about Neo4j
A graph epiphany awaits you.
What is a graph database?
How can I query a graph?
What do people do with Neo4j?

Jump into code
Use Cypher, the graph query language.
Code walk-throughs
RDBMS to Graph

System information
Key system health and status metrics.
Store sizes
ID allocation
Page cache
Transaction count

Copyright © Neo4j, Inc 2002–2020

Figure 2.17: Neo4j Browser Interface.

inserts, update and delete, query, and schema creation and modification (Codd, 2002). For example, in the Listing 2 shows the Cypher query that enables the creation of PETER node as shown in figure 2.15, thus:

```
CREATE (n1:Person {name: "PETER", born: "May, 3, 1970").
```

Listing 2.2: Cypher Example Query.

By default, *Neo4j* has an **graph algorithm library** included as Cypher procedures, in APOC [24] GitHub repository with the following algorithms, such as Centralities (Page Rank, Betweenness Centrality, and Closeness Centrality), Community Detection (Louvain, Label Propagation, Weakly Connected Components, Strongly Connected Components, and Triangle Count/ Clustering Coefficient), and Path Finding (Minimum Weight Spanning Tree, and All Pairs and Single Source Shortest Path).

[24] See github.com/neo4j-contrib/neo4j-apoc-procedures [Accessed: 1 July 2020].

2.7 Concluding Remarks

The topics covered throughout this chapter intended to provide theoretical support by approaching the various computer science fields which seek to answer the problem presented and the research questions asked:

- the concepts and definitions related to the criminal domain were presented, such as *5W1H* information. Also, we have introduced a perspective related to the criminal investigation in Portugal, describing the used tools during an investigation with advantages and gaps;
- our proposal deals with data from its sources, where it is extracted, transformed, and loaded into a structured form. We presented a process named as ETL, as well as the concepts that emerged from data, such as document or concepts;
- the extracted data is in the textual form, therefore, an NLP pipeline must be proposed with the appropriate techniques applied to the Portuguese language. Thus, we introduced concepts that must be understood before being applied. such as sentence splitting or NER;
- we have described the ML approaches, such as statistical methods or neural networks, that were applied in NLP tasks, such as the NER task;
- to enable the data representation, an explanation of graph databases concepts and technologies were introduced.

3

State of the Art

"If I have seen further, it is by standing upon the shoulders of giants"

Sir Isaac Newton (1643-1727)

This chapter summarizes the related work to our proposal since the main goal is to embrace the challenge of proposing a framework covering overlapped topics by performing efforts to describe and discuss the step-forwards and drawbacks performed by different researchers. We analyze the works related to the computational frameworks, NLP, and graph-based approaches applied to the criminal domain and the computational methods that could be applied to the criminal domain to achieve the goal.

A literary review is a way to identify and evaluate the most relevant research available to the fields of NLP, ETL, computational frameworks, GDB and IE applied to the criminal domain, or to support our framework from other domains. Hence, the existing knowledge must be studied and analyzed to avoid mistakes and follow the best-grounded path to achieve higher results (Randolph, 2009).

To base our research and write a literature review, we have established search criteria that support the contributions proposed in this thesis. The search is based on several queries carried out throughout journals and conference papers that were extracted from websites (such as IEEE Xplore Digital Library [1], Springer [2], ACM [3], Google Scholar [4]), and also the university's library catalogs. A concluding remark was introduced to interpret the related work findings or gaps in the research challenge. We organized the approaches order by date, along with a detailed description of the proposed architecture.

The remainder of this chapter is organized as follows: section 3.1 describes the computational frameworks applied to criminal domain; section 3.2 presents the ETL approaches that could help us to process the criminal-related documents; section 3.3 is divided into approaches applied to criminal domain (see section 3.3.1) and other approaches that could be used in our proposal (see section 3.3.2); in section 3.4 describes the approaches related to IE divided into Relation Extraction (see section 3.4.1), Event Extraction (see section 3.4.2), and 5W1H Extraction (see section 3.4.3); finally, in section 3.6 explores the graph-based approaches to criminal domain.

3.1 Computational Frameworks

In the scope of this thesis, the end-to-end computational frameworks were proposed to aim at a set of concepts, techniques, and practices to deal with a specific type of problem in the criminal domain context. They could be used as a reference to help us approach and solve new problems of a similar nature. In the following paragraphs, we describe, by date, each proposed framework applied to the criminal domain.

Chen et al. (2003) proposed the *COPLINK* project, a co-working project between the Tucson Police Department [5] and the University of Arizona [6] - Artificial Intelligence Lab. The objectives of *COPLINK* is to develop several knowledge management technologies supported by a set of modules, like Connect Database, Detect Criminal Intelligence Analysis, and Intelligent Agent Applications. The development faced some challenges, such as information sharing and collaboration between police departments and agencies and promoting crime intelligence analysis and knowledge management by each department. The *COPLINK* architecture is divided into modules: the *Security and Confidentiality Management*; *Collaboration (Data Mining)*; and *Information Access and Monitoring*. Additionally, the system uses a *User Profile Database* and the used *Data Sources*, such as police reports stored in relational police databases. Two relevant features have mentioned:

- *Domain-Specific Detect Concept Space*: identifies documents related to a domain-specific (terms and concepts). Filtering and indexing the terms and perform a co-occurrence analysis to identify the relationships among indexed terms;
- *COPLINK Detect* module: that was designed to recommend similar cases to users and identify police officers with similar information needs.

To implement the features below, several ML and Data Mining (DM) algorithms were applied, such as

[1] See www.ieeexplore.ieee.org/Xplore/home.jsp [Accessed: 1 November 2019].
[2] See www.link.springer.com [Accessed: 1 November 2019].
[3] See www.dl.acm.org [Accessed: 1 November 2019].
[4] See www.scholar.google.com [Accessed: 1 November 2019].
[5] See www.tucsonaz.gov/police [Accessed: 1 March 2020].
[6] See www.arizona.edu [Accessed: 1 March 2020].

ID3 (Grzymala-Busse, 1993), genetic algorithms (Davis, 1991), or relevance feedback (Allan, 1996). The obtained results were stored in a relational database, and a group of end-users validated the proposed system.

Stasko et al. (2007) proposed an NER open-source tool, named by *Jigwaw* [7], to identify and classify NEs like persons, places, objects or actions retrieved from police documents in the English language. The key objective of this tool is to establish relationships between entities across document collections, supported by external NLP tools, such as GATE [8], LingPipe [9], OpenCalais [10] and Ilinois-NER [11]. Authors applied rule-based (Eftimov et al., 2017) and dictionary-based (Deng et al., 2015).

Stampouli et al. (2011) proposed an *Police Intelligence Analysis Framework*, named by *PIAF*, to analyze witnesses statements regarding post-incident investigations by using computational methods, such as fusion algorithms. The architecture has based on a front-end (graphic user interface console) and a back-end (server). The server has divided into Software Architecture; Fusion Algorithms (Castanedo, 2013); Data Access and Analysis. For evidence storage, a database has included on the server. The *PIAF* uses an entity matching (Köpcke and Rahm, 2010) approach to identify incomplete information on witness statements about entities, such as persons. The entity matching with the help of fusion algorithms matches the incomplete information with a known entities dataset. The figure 3.1 shows an example of entity matching and fusion algorithms application.

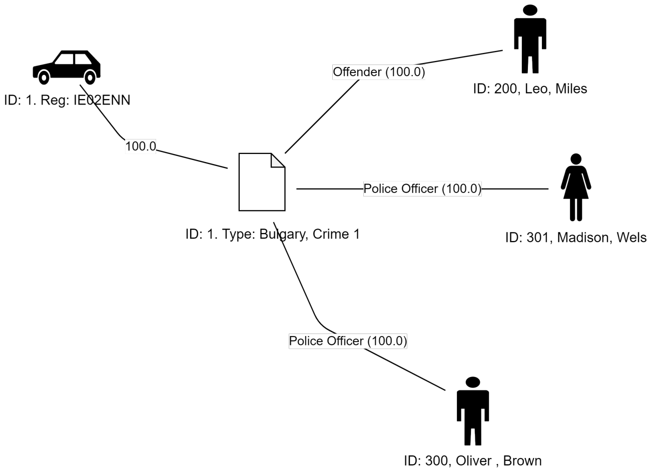


Figure 3.1: Witness's Matching Example [Adapted from (Stampouli et al., 2011)].

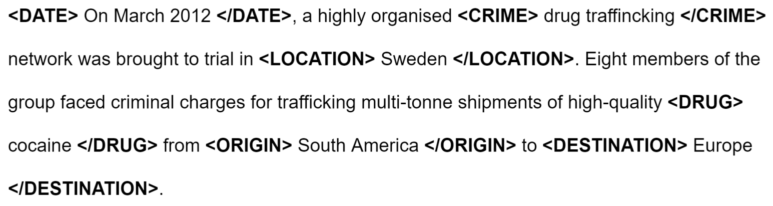
Albertetti and Stoffel (2012) proposed a data marts-based system that extracts data from police reports obtained from heterogeneous sources. The system is supported by a five-step process that proposes to use an operational data structure as a starting point. It ends with the creation of data marts: (1) *Data Identification*, (2) *Business Data Model Design*, (3) *Data Warehouse Model Design*, (4) *Data Mart Models Design*, and (5) *Testing and Analysing*. Authors designed a relational tool for crime analyses, organizing the police reports into a data warehouse. Introduced a *"novel human-centered data mining software"* (Poelmans et al., 2012) to retrieve intelligence from unstructured text, named by CORDIET, supported by use cases provided by the Amsterdam-Amstelland Police, GasthuisZusters Antwerpen (GZA) hospitals, and KU

[7] See www.cc.gatech.edu/gvu/ii/jigsaw [Accessed: 1 March 2019]
[8] See gate.ac.uk [Accessed: 1 March 2019]
[9] See alias-i.com/ [Accessed: 1 March 2019]
[10] See www.opencalais.com/ [Accessed: March 1 2019]
[11] See cogcomp.org/page/software [Accessed: 1 March 2019]

Leuven. For evaluation, authors performed tests with real police datasets using Pentaho^{TM} Data Integration Suite [12] tool. Hosseinkhani et al. (2012) proposed a *Combined Websites and Textual Document Framework*, named by the *CWTDF*, for investigating crime suspects by retrieving and analyzing web pages with the help of web mining techniques to discover crimes in data. The framework inputs are a group of websites related to crime (a list of non-visited websites has defined as a frontier process). All results have been stored in a local repository. This data has processed by the *Processing Repository*, where data has tagged with crime hot-spots, only data already stored. The crime communities are identified and extracted by its entities using text summarizing (Mase and Tsuji, 1999). After these processing activities, community profiles have been built to detect crime hot spots indirectly linked to each other, and criminal networks have been visualized.

Hossain et al. (2012) proposed a system to build stories based on entity networks. The system pipeline uses as input a *Document Corpus*, followed by *Entity Extraction* that uses external tools, such as LingPipe [13], OpenNLP [14], and Stanford NER [15]. The *Coreferencing* task was added to permit to disambiguate pronominal references, and references to the same person. *Entity Modeling* aims to model entities detected in *Entity Extraction* and disambiguate in *Coreferencing*. The *Concept Lattice Generation* enables to construct stories and *Heuristic Search* to find explanations from links between two entities. Finally, stories have been generated and visualized.

Adderley et al. (2014) proposed the *Project Multi-Modal Situation Assessment and Analytics Platform*, named by *MOSAIC*, which is an automatic crime detection and recognition in specific environments. The architecture has based on semantic information and classification of data sources, where the unstructured and structured data are integrated into a standard data format using ontologies. The combination of a NLP pipeline,a KB, and a crime pattern detection (designed to retrieve entities and events from text documents and websites) has included. Casanovas et al. (2014) defined a platform, named by *CAPER*, for detection and prevention applied to organized crime through sharing, exploitation, and analysis of OSINT. The architecture has four components: data harvesting, analysis, semantic storage and retrieval, and visual data analysis. For a semantic representation, the authors used two ontologies (the *European LEAs Interoperability Ontology* and *Multi-lingual Crime Ontology*). Brewster et al. (2014) proposed a system, named *ePOOLICE*, for detecting criminal threats retrieved from OSINT, using an ontology for system support, and knowledge graphs have used to create an *Environmental Knowledge Repository*. The following modules compose the system: the *Environment Crawlers* that extracts data from OSINT; *Filtering & Classification* for detection of NE and further analysis; *Environment Knowledge Repository* stores the relevant information retrieved from source environment and *Further Analysis* that uses several approaches, such as sentiment analysis or conceptual graphs (Sowa, 2008). Figure 3.2 shows an annotation made by *ePOOLICE* concept extraction.



<DATE> On March 2012 **</DATE>**, a highly organised **<CRIME>** drug traffincking **</CRIME>** network was brought to trial in **<LOCATION>** Sweden **</LOCATION>**. Eight members of the group faced criminal charges for trafficking multi-tonne shipments of high-quality **<DRUG>** cocaine **</DRUG>** from **<ORIGIN>** South America **</ORIGIN>** to **<DESTINATION>** Europe **</DESTINATION>**.

Figure 3.2: *ePOOLICE* Concept Extraction (Brewster et al., 2014).

[12] See www.pentaho.com/product/data-integration [Accessed: 1 March 2019]

[13] See http://www.alias-i.com/lingpipe/ [Accessed: 1 March 2019].

[14] See www.opennlp.apache.org [Accessed: 1 March 2019]

[15] See www.nlp.stanford.edu [Accessed: 1 March 2019]

Onnoom et al. (2015) developed an ontology-based framework applied to crime scene investigation documents used by the Forensic Science Police Center of Thailand. The framework has divided into six stages: from a *Data Preparation* that has introduced to sort data from crime scene documents presented in plain text; *Tokenization* to identify each token; *Sentence Patterns Analysis* identify each sentence ranked by frequency, where repeating words has extracted. Finally, the outcome of the last phase enables ontology development and implementation. Wijeratne et al. (2015) suggested architecture to discover criminal gang structure, how they function and operate, by analyzing the social media footprints in Chicago, Illinois (EUA). The architecture has divided into four modules: *Data Collection and Filtering* detects and retrieve tweets associated with gangs from Twitter [16]; the *Slag Term Dictionary* has used for slag terms detection related to Chicago gangs; *Data Processing* allows the data processing using ML algorithms and NLP tasks, such as NER, sentiment analysis, and other methods; *Data Access Tools*, data access tools for semantic processing with tasks-related; and *Data Analysis and Interpretation* a question-answer task, that performs questions like *"Who is the gang user A is affiliated with?"* for more in-depth analysis. Elyezjy and Elhaless (2015) proposed a crime detection framework to discover relationships between offenders and communities by extracting information from police documents in the Arabic language. After these established relationships, results were visualized for analysis with applied tools. The architecture has divided into four modules:

- *Data Gathering*: enables the construction of a police documents corpus by gathering documents from police departments (Gaza Strip);
- *Data Preprocessing*: perform several tasks over corpus, such as NLP tasks like tokenization or sentence splitting;
- *Algorithms Module*: proposed to detect direct and indirect relations between entities;
- *Data Visualization*: uses a graph visualization framework, called Dracula Graph Library [17].

Nouh et al. (2016) proposed the *Cyber Crime Intelligence Framework*, named by *CCINT*, a multipurpose framework to aid law enforcement for detecting, analyzing cybercrime data. The *CCINT* framework can be viewed with a top-down and bottom-up approach, divided into five layers:

- *Users*: individual and collaborative;
- *Front-End*: enables the configuration settings, such as dashboards or visualization methods;
- *Analysis (Functions and Methods)*: uses computational methods for cybercrime accounts for analyses, such as content and sentiment analysis, Social Network Analysis (SNA) the time or geospatial analysis;
- *Data Handling*: uses a crawler to collect data from online sources based on user configurations, followed by pre-processing and cleaning data tasks, search and filtering tasks, and a database for storage;
- *Data Sources*: to retrieve data from OSINT, police reports or evidence.

Mata et al. (2016) proposed a mobile information system based on crowd-sensed and official crime data, like crime reports. This approach seeks to find safe routes using classification methods supported by data retrieved from tweets related to crime events in Mexico City, using a classifier algorithm to collect relevant crime data. The main goal is to integrate crowd-sourcing data (tweets) with official crime reports into a

[16] See `twitter.com` [Accessed: 1 May 2020].
[17] See `www.graphdracula.net` [Accessed: 1 May 2020].

mobile application, with a crime ontology to process data semantically and classify using a Bayes' theorem algorithm. The system has divided into five layers:

- *Retrieval Crime Data Sources*: that permits to extract of tweets related to crime events from the Tweets Database, that enables to define of time and location events;
- *Crime Data Repository*: data collected after been analyzed is integrated with crime data from Official Reports (government institutions);
- *Semantic Processing*: supported by an ontology; *Clustering Approach* uses a Bayes' theorem algorithm to classify data in conjunction with Integrated Crime Database, adding the GeoNames [18] web service to resolve synonymy between names;
- *Generation Safe Routes*: enables the generation of safe routes based on crime rates estimation by applying a Bayes' theorem algorithm; after this, the safe root is visualized on the mobile application.

Wiedemann et al. (2018) suggests an IE pipeline to process collections of unstructured textual data for investigative journalism automatically. The data sources, named by *Hoover*, were developed by the European Investigative Collaborations (EIC) network with a particular focus on large data leaks and heterogeneous datasets. Authors based their work in a configurable pipeline that takes part in the available tools the following modules: preprocessing, dictionaries and regular expressions patterns, temporal expressions, NER, and term extraction.

The table 3.1 summarizes the frameworks by features that we consider relevant to our research.

3.2 Data Processing

The first step in an end-to-end framework is to define a set of tasks that enables data extraction, cleansing, transformation, normalization, and loading into a staging area or by feeding it directly to an NLP pipeline. The enumerated steps are part of an ETL framework, which is a way of resolving some issues related to unstructured data. In the following paragraphs, we describe the relevant research papers and their respective papers relevant and order by date.

Skoutas and Simitsis (2007) proposed an ontology-based approach to support an ETL framework. A graph-based representation supports a data store graph (aggregated with an XML Schema definition) as a common model for data stores, and an ontology has introduced for the application. These proposals have, combined with supporting structured and semi-structured data extraction and representation. Majeed et al. (2010) proposed a technique to data stream handling coupled with a synchronizing task to enable the synchronization between existing data and incoming data streams from sources. The reduction of irrelevant data was made by introducing filters using queries. However, every time that data volume increased, synchronization issues also increased, which led to relevant data to be lost. Song et al. (2010) performing an aggregate operation and querying tasks in a real-time data warehouse has divided into two parts for data synchronization:

- the triggering algorithm that manages the ETL process timings when the data sets are arriving from different sources;
- the scheduling algorithm for resources weighing between queries and updates.

[18] See www.geonames.org [Accessed: 1 March 2019]

References	Data Source (Input)	Computational Methods	Knowledge Bases	Output	Framework Name
Chen et al. (2003)	police databases [structrured data]	DM ML ID3 genetic algorithms relevance feedback	N/D	relational database	**COPLINK**
Stasko et al. (2007)	police documents [unstructured data]	NLP NER	N/D	entities are tagged by their respective category name and stored per document	***JIGSAW***
Stampouli et al. (2011)	witness statements [unstructured data]	entity matching fusion algorithms	Police database	relational database	***PIAF***
Albertetti and Stoffel (2012)	police reports [structured data]	ETL	N/D	relational database	—
Hosseinkhani et al. (2012)	OSINT [unstructured data]	NLP text summarization	N/D	N/D	***CWTDF***
Hossain et al. (2012)	[unstructured data]	NLP NER Concept Lattice Nearest Neighbors Approximation k-Clique Near Neighbor	N/D	N/D	—
Adderley et al. (2014)	[unstructured structured data]	NLP	N/D	ontology	***MOSAIC***
Casanovas et al. (2014)	[unstructured]	NLP	European LEAs Interoperability Ontology Multi-lingual Crime Ontology	ontology	***CAPER***
Brewster et al. (2014)	[unstructured]	NLP	N/D	ontology	***ePOOLICE***
Onnoom et al. (2015)	crime scene documents [unstructured]	NLP	N/D	ontology	—
Wijeratne et al. (2015)	[unstructured]	NLP ML	Slang Term Dictionary	N/D	—
Elyezjy and Elhaless (2015)	police documents [unstructured]	NLP	N/D	N/D	—
Nouh et al. (2016)	OSINT, police reports evidence streams [unstructured]	NLP sentiment analysis SNA time or geospatial analysis	N/D	N/D	***CCINT***
Mata et al. (2016)	crowd-sensed official crime data [unstructured]	NLP	geonames	ontology	—
Wiedemann et al. (2018)	data leaks heterogeneous datasets [unstructured]	NLP	N/D	N/D	—

Table 3.1: Computational Frameworks Summary

Chávez and Li (2011) suggested an ontology-based approach to advance the homogeneity in data sources and solved some drawbacks, such as meta-information extraction. Jiang et al. (2010) employed a domain ontology inside of an ETL process to support the search for data sources and defined data transformation rules, or heterogeneity elimination. Ontology concepts were created from the database schema, using a manual process, and supported by the Entity-Relationship diagram. Data source's findings were semantically analyzed for the transformation to be more efficient. Gorawski and Gorawska (2014) proposed in 2014, a real-time ETL engine architecture to handle data streams in a real-time data warehouse, by using a remote buffer framework for data streams management from different sources, characterized by volume and velocity. Knap et al. (2014) introduced *UnifiedViews*, an ETL framework that enables users to configure the framework, like define, execute, monitor, debug, schedule, or share ETL data processing tasks. This framework enabled users to build their plugins. Li and Mao (2015) suggested a real-time ETL framework that treats historical and real-time data by using different events, with an external dynamic storage area and a dynamic mirror replication technology bypassing the contention between Online Analytic Processing queries and Online Transaction Processing updates. As remarked by the authors, the framework does not handle issues regarding unstructured data. Bansal and Kagemann (2015) proposed a semantic ETL framework that applies semantics to various data fields and allows more efficient data integration. Nath et al. (2017) proposed an ETL framework, called by *SETL*, over the Semantic Web standards and tools, by adding support to semantic data sources in addition to the traditional data sources.

Finally, a fairly large number of tools are available to implement an ETL framework divided into: commercial, trial versions to evaluate our proposals, such as IBM Infosphere [19], Oracle Warehouse Builder [20], Microsoft SQL Server Integration Services [21], and Informatica Powercenter [22]; in the open-source, we can use them for prototyping or end-user applications, like the Talend Open Studio [23], Pentaho Kettle [24], and CloverETL [25], Apache nifi [26], Jaspersoft ETL [27] and GeoKettle [28].

3.3 Natural Language Processing

As part of any pipeline that deals with textual data, NLP tasks are introduced to understand natural languages and adapted to several domains. Thus, we describe the related work against NLP (see section 3.3.1) applied to the criminal domain, which is a computational method to understand the documents that require consideration by police departments as well as the approaches applied to the Portuguese language and not exclusively related to the criminal domain (see section 3.3.2). In the following paragraphs, we describe the relevant research papers and their respective papers relevant and order by date.

3.3.1 Criminal Domain

Chau et al. (2002) proposed a NER approach to extract useful entities from police narrative reports. The entities are identified as person names and narcotic drugs. The system combines lexical lookup, machine learning, and hand-crafted rules to enable identification and classification of NEs.

Ku et al. (2008) proposed an IE system coupled to victims or witnesses online reporting system that enables anonymous crime reporting. The police narrative reports fed the system. The system aims to extract crime-relevant information from police and witness narrative reports using NLP tasks applied to the English language. Based on a pipeline composed of six modules:

- *Narrative Text*: obtain from narrative police and witnesses/victims interview reports;
- *GATE Toolkit* [29]: NLP pipeline using tokenizer, sentence detection, POS, noun chunking, and orthomatching modules. Java Annotations Pattern Engine Rules, named by JAPE, was added to establish a set of rules that uses regular expressions and Gazetteer Lists (built under a lexicon that enables the identification of crime types, weapons, vehicles, scenes, clothes, shoes, and physical features);
- *Phrase Filtering*: use a stop-word list to remove from phrases extracted that are not relevant for IE, such as “a”, “an”, or “the”;
- *Phrase Comparison*: enables the duplicated phrases removal;
- *Output Summary*: generates reports using Microsoft™ Excel, document text and graph representations;

[19] See www.ibm.com [Accessed: 1 July 2020].
[20] See www.oracle.com [Accessed: 1 July 2020].
[21] See www.microsoft.com [Accessed: 1 July 2020].
[22] See www.informatica.com [Accessed: 1 July 2020].
[23] See www.talend.com/products/talend-open-studio/ [Accessed: 1 July 2020].
[24] See www.pentaho.com/product/data-integration [Accessed: 1 July 2020].
[25] See www.cloveretl.com/ [Accessed: 1 July 2020].
[26] See nifi.apache.org/ [Accessed: 1 July 2020].
[27] See community.jaspersoft.com/project/jaspersoft-etl [Accessed: 1 July 2020].
[28] See www.spatialytics.org/projects/geokettle/ [Accessed: 1 July 2020].
[29] See gate.ac.uk [Accessed: 1 January 2020].

- *Online Reporting System*: creates an online crime reporting system used to report crime anonymously.

Pinheiro et al. (2010) proposed a IE system using the Portuguese language, based on a SIM (Semantic Inferential Model) (Pinheiro et al., 2008) that drives semantic analysis has applied on public safety areas, such as police departments. The designed system supports a collaborative web-based system of crime registering, named WikiCrimes (Furtado et al., 2010). This approach performs a morphological and syntactic analysis using NLP tasks that produce syntactic trees as an output; and the Semantic Inferential Model. The Semantic Inferential Model provided a semantic analyzer to understand the relevant information extracted from crime reports and defined a group of generic sentences, called sentence-patterns, such as templates that function as slots where we can block up with terms, e.g., *<X><be murder><by Y>*. Followed by a *Conceptual Base* that adds inferential activity based on concepts found in natural language, defined in the knowledge area. Finally, a *Semantic Inferentialism Analyser (Pinheiro et al., 2009) (SIA)* is responsible for understanding sentences and implementing an inferential network (with pre-conditions and post-conditions of sentences).

Al-Zaidy et al. (2011) used a set of computational methods, such as NLP methods, to discover criminal communities by established their relations after a NLP pipeline applied to extract relevant information from criminal data, such as e-mails, chat logs or any textual data. The purpose of this pipeline is to

- *Extract Person Names*: proposed to identify and extract person names using a Named Entity Recognition (NER), using Stanford Named Entity Tagger [30] to identify English names;
- *Normalization Process*: that performs a process to eliminate duplicated names;
- *Extract Crime Communities*: included to extract frequent communities and to analyze the strength of linkages between individuals (persons);
- *Extract Relevant Information for Police Investigators* such as contact information, e.g., address or phone, and summary topics of documents, using Open Text Summarizer [31]

Technology et al. (2012) designed a NER system applied to Arabic Crime Documents. The system was divided into three phases: *Preprocessing* with a pipeline with sentence splitting, tokenizer, POS and noun phrase chunker; *NER* using a Gazetteer List with terms, like a person, personal properties, location, organizations, and crime terms, and a *Pattern Rules* module that trained the NER classifier to tag crime entities in documents. Wang et al. (2012) proposed an criminal incident system based on prediction features using the Twitter [32] platform as a data source. Built on the automatic semantic understanding and analysis of NLP Twitter posts, by using the dimensionality reduction via Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and prediction via linear modeling. The authors verified their model by predicting hit-and-run crimes. The system has progressively divided into:

- *Data Collection*: that extract events from the main textual content of each tweet using NLP techniques;
- *Semantic Role Labelling*: that identifies and classified the semantic roles, adding the Latent Dirichlet Allocation to identify salient topics within the extracted events;
- *Topic Extraction*: enables the topic extraction;

[30] See nlp.stanford.edu/software/CRF-NER.shtml [Accessed: 1 March 2019]

[31] See github.com/neopunisher/Open-Text-Summarizer/ [Accessed: 1 March 2019]

[32] See twitter.com/ [Accessed: 1 July 2020].

- *Predictive Model*: to facilitate a prediction model based on data;
- *Incident Prediction*: aims to predict hit-and-run crimes from a model trained.

Camara Junior (2013) aimed to create an architecture for indexing information of documents using NLP mechanisms at a semantic level, supported by an ontology. The dataset contains data extracted from cyber-crimes forensics reports produced by Federal Police Forensics Experts in Portuguese-Brazilian language. Helbich et al. (2013) uses text mining approaches, specifically the self-organizing map algorithm (Kohonen, 1990) and its visualization capabilities joint with pattern analysis, to explore hidden information regarding geographical information. This approach has applied to an unsolved homicide series in the city of Jennings, Louisiana.

Arulanandam et al. (2014) proposed a system to extract information from online newspapers focused on the "hidden" information related to theft crime. To enable this extractions, an NER task were applied, like CRFs classifiers. The pipeline has divided into:

- *Building Corpus*: a set of texts were extracted from online newspapers using a web scrapping tool (Mozenda Web Screen Scrapper Tool [33]);
- *Sentence Tokenization*: a list of individual sentences were submitted to tokenization, using Punkt-Tokenizer from NLTK Toolkit [34];
- *Location Identification*: aims to identify locations in each sentence, using NER approaches;
- *Feature Identification* a set of features were defined to assign labels to sentences, such as crime location sentence (CLS) and not a crime location sentence (NO-CLS);
- *Label Assignment*: the labels are manually assigned a CLS - Crime Location Sentence and a NO-CLS - Not a Crime Location Sentence;
- *Training CRFs*: the corpus have been annotated and trained using the CRFs algorithm;
- *Sentence Classification*: the model was created and used to label the sentences in the remaining articles, like text data. The labels obtained are compared with the labels assigned by humans.

Shabat and Omar (2015) proposed to extract crime information from the web, based on machine learning models, and crime NER task, using classification algorithms, e.g., Naiv̈e Bayes, SVM and K-Nearest Neighbor. These classification algorithms were used to feature extraction through a voting combination module that enables feature identification. An indexing module that aids crime type identification by using the same classification algorithms. The system was divided into:

- *Crime DataSet (BERNAMA)*: developed an annotated dataset from crime documents from the Malaysian National News Agency (BERNAMA), the dataset was annotated manually, like crime type, weapons, crime places, and nationality;
- *Pre-Processing phase*: since data was collected from newspapers and social media sites, this phase aims to reduce entropy that exists in data collected, using natural language processing tasks, such as tokenization, stop word removal and stemming;

[33] See www.mozenda.com [Accessed: 1 March 2019])

[34] See www.nltk.org [Accessed: 1 March 2019]

- *Feature Extraction*: the objective is the conversion of each word to a vector of featured values, by building an array of features, such as nationalities, weapons, and crime scene locations, retrieved from online resources;
- *Indexing*: uses a method by converting documents to a document vector has proposed to detected crime types, indexing these documents by crime type;
- *Voting Combination*: the SVM (Suthaharan, 2016), K-Nearest Neighbour (Cunningham and Delany, 2007) and Naiv̈e Bayes (NB) (Murphy et al., 2006) Classifiers have applied in a voting combination, observes the outputs from classifiers before describing them as the input, and select a classifier with the best results to NER and Crime Type Identification;
- *Evaluation*: aims to measure the performance of NER and Crime Type Identification.

Rahem and Omar (2015) propose a rule-based NER model for drug-related crime news documents. Applying heuristic and grammatical rules to extract NEs, such as drug types, drugs amount, drugs price, drug hiding methods, and the suspect nationality.

Bsoul et al. (2016) proposed a system to extract verbs from crime clusters using two datasets, such as a real crime dataset and an industrial dataset. The system added NLP tasks that remove non-relating information by using stop words. Additionally, the Porter stemmer method has been added for word stemming. To identify the verbs, it was used a WordNet [35] identification method. Sheela and Vadivel (2016) proposed an approach to explore the hidden information, supported by text mining techniques, from crime reports and Social Web. This approach deal with the automatic construction of crime-related thesaurus based on domain-related terms and sentences to identify usual patterns. For sentence classification and clustering was used the ANN tool (Krenek et al., 2014).

Das and Das (2017) defined an approach for crime pattern analysis using different text mining techniques, which include NER from crime reports in online newspaper articles, such as The Times of India, The Hindu, and The Indian Express. The two-stage NER approach was defined: *Data Collection and Preparation* uses a web-crawling tool has used for data crawling from papers and websites containing crime reports, like rapes, kidnapping, or child abuse, stored for further crime pattern analysis. The NER module has divided into: *Recognizing the Basic Named Entities* such as area, streets, towns/cities or villages, using different NLP tasks, such as sentence splitting or noun phrase chunking; and *Modus Operandi Features* targets criminals behavior is an essential task for crime pattern analysis, such as crime type, place, reason or victims of crime.

Schraagen (2017) considered evaluating an "out-of-the-box" NER system, named *FROG* [36], for automatically processing crime reports and online criminal complaints by the Dutch police, especially the cyber-related crimes, such as phishing and online trade fraud. Authors used 250 criminal annotated complaints by domain-experts that annotated entities with an entity type, like locations, persons, organizations, events, products, and miscellaneous. An agreement between annotators was measured by Cohen's K was 0.75, under the following conditions, or the exact string or entity type equality. Ejem (2017) proposes on its Master Thesis, an approach regarding relation extraction between NEs, assuming that NEs are already identified and classified in police reports. These police reports are from the Anti-drug Department of the Police of the Czech Republic.

Martin-Rodilla et al. (2019) describes the analysis of 3000 textual reports in São Paulo during the Brazilian dictatorship through unsupervised and supervised approaches, using the *Linguakit Suite*, the *Standford*

[35] See www.wordnet.princeton.edu [Accessed:1 March 2020].

[36] See www.languagemachines.github.io/frog/ [Accessed: 1 January 2020].

CoreNLP [37] tool, *SIEMÊS* (Sarmento, 2006) algorithms, by identifying NEs and relevant terms. This approach tries to set people's information and automate the study of correlations between actors.

Gianola (2020) proposed in its PhD Thesis an adaptation of NLP and IE methods applied to witness interviews in French language using official documents from Gendarmerie Nationale [38], including a comparison of the concepts of entity in criminal analysis and NE in NLP. For NEs identification was based on a rule-based approach, by using Unitex/GramLab [39] tool to support this approach.

In conclusion, the related work described above is summarized in table 3.2, giving a summary of all approaches and solutions taken to achieve the author's primary goals.

References	NLP Tasks or Other Methods	Learning Methods	Knowledge Bases	External Tools	Named-Entities or Terms Extracted	Language
Chau et al. (2002)	NER	hand-crafted rules machine learning	Government Census Data	Arizona Noun Phraser	persons narcotic drugs	EN
Ku et al. (2008)	tokenization, sentence splitting noun chunking, ortho-matching	rule-based	Gazetter Lists	GATE Toolkit	crime types weapons vehicles	EN
Pinheiro et al. (2010)	—	rule-based	Ontology	PALAVRAS	—	BR-PT
Al-Zaidy et al. (2011)	NER	supervised	—	Stanford Named Entity Tagger Open Text Summarizer	persons address or phone summary topics of documents	EN
Technology et al. (2012)	sentence splitting, tokenization POS Tagging, noun phrase chunker	rule-based supervised	Gazetter Lists	—	persons locations organizations crime terms	Arabic
Wang et al. (2012)	SRL Latent Dirichlet Allocation	supervised	—	—	—	EN
Camara Junior (2013)	Sentence Splitting Tokenization	—	Ontology	NLTK Toolkit	—	BR-PT
Helbich et al. (2013)	self-organizing map algorithm					EN
Arulanandam et al. (2014)	tokenization NER	supervised		Mozenda Web Screen Scrapper Tool NLTK Toolkit	places	EN
Shabat and Omar (2015)	NER	supervised	—	—	crime type	EN
Rahem and Omar (2015)	POS Tagging Sentences Classification and clustering	rule-based	—	—	drug types drugs amount drugs price drug hiding methods	EN
Das and Das (2017)	sentence splitting noun phrase chuncking NER		—	—	crime type places victims of crime reason of crime	Hindu
Bsoul et al. (2016)	Stemming	—	—	Porter stemmer	—	EN
Schraagen (2017)	NER	supervised	—	—	locations persons organisations events products miscellaneous	EN
Ejem (2017)	relation extraction	rule-based supervised	—	—	—	Czech
Martin-Rodilla et al. (2019)	NER Term Extraction Sentence Splitting, Tokenization Lemmatization, POS Tagging	unsupervised	Terms List	Linguakit Suite SIEMÊS Standford CoreNLP	persons	BR-PT
Gianola (2020)	NER	rule-based	—	Unitex/GramLab	persons	French

Table 3.2: Natural Language Processing Approaches Summary.

3.3.2 Works to Support Natural Language Processing Pipeline

NLP approaches have been studied by several authors for Portuguese language. However, our goal is not to development from scratch, but to adapt them to our objective. In the following paragraphs, we intend to describe the tasks already described in the literature that could be applied to our domain, starting with lexical and syntactic analysis and followed by semantic analysis.

Regarding to **sentence splitting**, different approaches have been proposed to solve sentence splitting. Beginning with Silla and Kaestner (2004) that published a study about different systems regarding text

[37] See www.stanfordnlp.github.io/CoreNLP/ [Accessed: 1 July 2020].
[38] See www.gendarmerie.interieur.gouv.fr/ [Accessed: 1 July 2020].
[39] See www.unitexgramlab.org/pt [Accessed: 1 July 2020]

segmentation applied to English or Portuguese texts. The systems under evaluation were, the *RE System*, *MxTerminator* and *Satz*.

- *RE System* uses an approach based on hand-crafted rules by using regular expressions to detect sentences boundaries by scanning it to the end of a sentence, adding exceptions to the system like decimal numbers, parenthesis at the end of a sentence and ellipsis. *RE system* has adapted to Brazilian-Portuguese by adding 240 regular expressions that indicate abbreviations linked to the used language;
- *MxTerminator* system proposed an approach that is language independent or text genre, supported by a machine learning algorithm, named as the MaxEnt.
- *Satz* proposed an approach that analyzed each punctuation marks and their context, flexible to use any machine learning algorithm, such as the decision tree classifier or neural networks algorithms.

Kiss and Strunk (2006) suggested a language-independent unsupervised approach for sentence boundary detection, named by *Punkt*. The system has based on the detection of ambiguities, such as abbreviations, that could create issues related to sentence boundaries identification. López and Pardo (2015) proposed a supervised algorithm that has extended to the sentence boundary detection method to the domain of the opinionated texts retrieved from the web. Rodrigues et al. (2018) proposed a task named SenPORT (Sentence Detector) for sentence splitting, integrated into the *NLPPort* [40] toolkit. Their work has based on the *OpenNLP* tool and the sentence detection model available for the Portuguese language, but with a list of abbreviations to avoid splitting sentences into periods that commonly occur in abbreviations, also adding an option were line breaks could result in a new sentence, by using regular expressions to help on this task, such as `(\n\r?)|(\r\n?)`.

The **tokenization** is a common lexical approach that integrates into a well-defined NLP pipeline. In the following paragraphs, we highlight the related work that try to resolve the tokenization issue. Branco and Silva (2003) that proposed a tokenization task applied to Portuguese texts that address different issues found in ambiguous strings by defining a set of rules to resolve the hard cases tokenization process encounter. Maršík and Bojar (2012) it aimed to describe a universal tool for segmentation and tokenization of textual data, applied to detect sentences in English text, and identify words in Chinese text, named by *TrTok*. Vijayarani and Janani (2016) proposed a comparative analysis against the following tokenization tools, such as Nlpdotnet Tokenizer, Mila Tokenizer, NLTK Word Tokenize, TextBlob Word Tokenize, MBSP Word Tokenize, Pattern Word Tokenize and Word Tokenization with Python NLTK. Based on the obtained results by Vijayaran et al., the best tool was Nlpdotnet Tokenizer that revels the best performance measures results. Rodrigues et al. (2018) proposed a tokenizer, named as *TokPORT*, which was integrated the *NLPPort* toolkit, based on the OpenNLP tool that uses a pre-trained model for the Portuguese language, with some improvements, such as contractions and clitics during the sentences pre-processing. For abbreviation detection, it has been considered a post-processing task.

The **POS Tagging** enables the identification in a sentence for each token (word), the part-of-speech tags, such as nouns, pronouns, verbs, adverbs, among others. Mendes et al. (2003) reuses an annotation tool developed to tag a spoken corpus with POS tags. Using Eric Brill's tagger, it has trained with a corpus with around 250000 words used to tag the Portuguese C-ORAL-ROM spoken corpus of 300000 words. Feldman (2006) proposed a system for automatic morphological analysis and tagging of Brazilian-Portuguese by avoiding extensive resources, such as a large annotated corpora and lexicons. Therefore, the authors used:

- (1) an annotated corpus of Peninsular Spanish, a language related to Portuguese, CLiC-TALP tagset;

[40] See `www.github.com/rikarudo/NLPPORT` [Accessed: 1 May 2020].

- (2) an unannotated corpus of Portuguese, NILC corpus;
- (3) a description of Portuguese morphology on the level of an essential grammar book.

Previous works have extended by adopting an alternative algorithm for cognitive transfer that affects Spanish emission probabilities into Portuguese. Gonçalves et al. (2006) used SVM, which are known to produce good results on text classification tasks. The proposals have applied to two different datasets written in the Portuguese language, like the Brazilian newspaper (Folha de São Paulo) news, and juridical documents from the Portuguese Attorney General's Office. de Holanda Maia and Xexéo (2011) proposed a probabilistic approach for POS that combined HMMs and character language models applied to Portuguese texts. Garcia et al. (2014) presented different dictionaries of the new orthography (Spelling Agreement) and a freely available testing corpus, containing different varieties and textual typologies. Fonseca et al. (2015b) proposed an architecture based on neural networks and word embedding, that achieved and that has achieved promising results in English. The classifier has been tested in different corpora: a new revision of the Mac-Morpho corpus, in which we merged some tags and performed corrections and in two other previous versions of it. The impact of using different types of word embedding and specific features as input has been evaluated.

The **Lemmatization** are common in different tools, but not available as isolated lemmatizers, although there are suites of tools, including morphological analyzers, that produce lemmas as a part of their outcome. There are available tools that address the lemmatization issue, such as: JSpell [41], Freeling [42], LX-Suite [43]. Each enumerated tool has detailed in the respective website or published papers (Carreras et al., 2004) (Simões and Almeida, 2001).

The *JSpell* and *Freeling* have a downloadable version, and the *LX-Suite* only have an online service. Moreover, *JSpell* is available as C and Perl libraries and an MS Windows binary; for the *Freeling* tool, a Debian package, and an MS Windows binary are available for download; additionally, an API in Java, Python, and a native C++ API. Regarding the author's lemmatization resolution approach, both *jSpell* and *Freeling* start with a collection of lemmas and the use of a set of rules to create inflections and derivations from those lemmas (Carreras et al., 2004) (Simões and Almeida, 2001). For words lemmatization, as the output of this process, words are matched against the produced inflections and derivations, retrieving the originating lemma.

The *LX-Suite* (Branco and Silva, 2007) is a set of NLP tools developed by the LX-Center, that uses a nominal lemmatizer (Branco and Silva, 2006) based on a lexicon to retrieve exceptions to the lemmatization rules: if the lexicon contains a word that would be processed by the rules, it is marked as an exception and added to the exception list. According to the authors, lemmatizer achieved an accuracy of 97.87%, when tested against a hand-annotated corpus with 260,000 tokens. For jSpell and Freeling, although evaluations for the morphological analyzers exist, no statement regarding the accuracy of the respective lemmatizers was found.

da Silva (2007) proposed on is the thesis, a nominal lemmatizer that is supported by morphological regularity found in word inflection by using a set of transformation rules that revert to the form of the lemma. Rodrigues et al. (2014) proposed a lemmatization method by sharing features with other approaches, such as the use of rules and, more recently, a lexicon.

Several approaches or tools address the **dependency parser** issue. Hence, the tools available for dependency parsers are:

[41] See www.natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell [Accessed: 1 August 2019].

[42] See www.nlp.lsi.upc.edu/freeling/index.php/node/1 [Accessed: 1 August 2019].

[43] See www.lxcenter.di.fc.ul.pt/services/en/LXServicesLemmatizer.html [Accessed: 1 August 2019].

- *DepPattern* (Otero and González, 2012) toolkit [44] is a linguistic package providing a grammar compiler, PoS taggers, and dependency-based parsers for several languages, provided with parsers for five languages: English, Spanish, Galician, French, and Portuguese. The parsers were implemented in PERL;

- *MaltParser* [45] (Nivre et al., 2007), a data-driven system for dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using an induced model;

- *LX-Parser* [46] is syntactic analyzer for Portuguese based on a statistical approach;

- *Freeling* [47] is presented as a open source suite of language analyzers, developed at TALP Research Center (Universitat Politécnica de Catalunya). Introduces as part of suite, the European Portuguese Freeling PoS-tagger - trained with the following linguistic resources, such as Bosque 8.0 from Linguateca.

Zilio et al. (2018) proposed the dependency parsing model for Portuguese, named *PassPort*, trained with the Stanford Parser. Rodrigues et al. (2018) proposed a dependency parser, named *DepPORT* based on MaltParser system parser, trained with Bosque 8.0 corpus from Linguateca (a 138,000 token corpus from the Floresta Sintá(c)tica, manually revised by linguists), after conversion to a processable format, the CoNLL-X format.

Regarding semantic analysis, previous studies have reported **Named-Entity Recognition** systems applied to the Portuguese language or variants that will be the ground of our work. In this section, we describe the studies related to NER systems, annotating the improvements and gaps.

Bick (2006a) proposes the *PALAVRAS-NER* (hereinafter mentioned as *PALAVRAS*) system that uses a Constraint Grammar composed by a NER task for grammatical tagging. The original version was released in 2003, at 6th International Workshop - Computational Processing of the Portuguese Language (Mamede et al., 2003), establishing a tag set of six NER categories (person, organization, place, event, semantic products, and objects) with about twenty subcategories.

Martins et al. (2007) proposed the *CaGE* system to deal with recognition and disambiguation of places. Two ontologies were developed focused on the Portuguese and the Global territory.

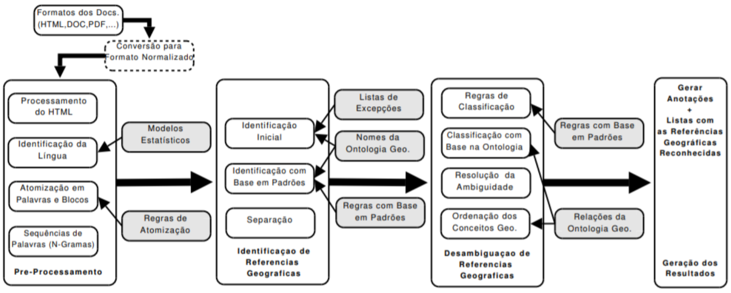
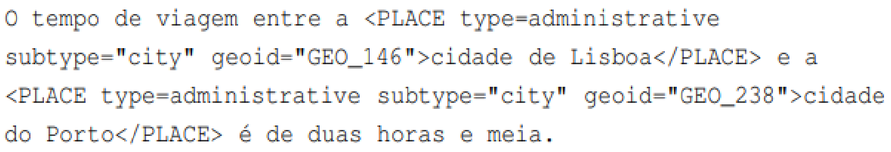

Formatos dos Docs. (HTML,DOC,PDF,...)
Conversão para Formato Normalizado
Processamento do HTML
Identificação da Língua
Atomização em Palavras e Blocos
Sequências de Palavras (N-Gramas)
Pre-Processamento
Modelos Estatísticos
Regras de Atomização
Identificação Inicial
Identificação com Base em Padrões
Separação
Identificaçao de Referencias Geograficas
Listas de Excepções
Nomes da Ontologia Geo.
Regras com Base em Padrões
Regras de Classificação
Classificação com Base na Ontologia
Resolução da Ambiguidade
Ordenação dos Conceitos Geo.
Desambiguaçao de Referencias Geograficas
Regras com Base em Padrões
Relações da Ontologia Geo.
Gerar Anotações + Listas com as Referências Geográficas Reconhecidas
Geração dos Resultados

Figure 3.3: *CaGE* Architecture (Martins et al., 2007).

[44] See www.gramatica.usc.es/pln/tools/deppattern.html [Accessed: 1 September 2019].
[45] See www.maltparser.org [Accessed: 1 May 2019].
[46] See www.lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXParser.html [Accessed: 1 September 2019].
[47] See www.gramatica.usc.es/pln/tools/freeling.html [Accessed: 1 September 2019].

Figure 3.3 describes the *CaGE* architecture divided into four main blocks: *Pre-Processsing* (in Portuguese “Pre-Processamento”); *Geographic References Identification* (in Portuguese “Identificação de Referências Geográficas”); *“Geographic References Disambiguation”* (in Portuguese “Desambiguação de Referências Geográficas”); and finally the *Results Generation* (in Portuguese “Geração de Resultados”). *Pre-Processsing* was split into sub-tasks, such as format conversion, HTML rendering, language classification, atomization, and pairing of n-grams, with the objective of processing data retrieved from WWW and uses a context pair approach for atomization and sentence recognition for posterior division on n-grams. *Geographic References Identification* uses the n-grams to recognize the geographic references by applying manual rules. *“Geographic References Disambiguation”* uses a set of rules based over the selected ontologies. Figure 3.4 shows the obtained results from the *Results Generation* task.



```
O tempo de viagem entre a <PLACE type=administrative
subtype="city" geoid="GEO_146">cidade de Lisboa</PLACE> e a
<PLACE type=administrative subtype="city" geoid="GEO_238">cidade
do Porto</PLACE> é de duas horas e meia.
```

Figure 3.4: *CaGE* SGML Annotated Tags (Martins et al., 2007).

Mota (2008) presents the *R3M* system for NE identification and classification, like persons, organizations, and places. The system identifies and classifies people, organizations, and locations (due to time limitations) using a flexible architecture that permits new entities and relations between entities by using a semi-supervised learning algorithm.

Figure 3.5 describes the *R3M* architecture, divided into an *Identification* (in Portuguese “Identificação”) and *Classification* (in Portuguese “Classificação”) that completes the five principal tasks and sub-tasks. The system is detached by training (in Portuguese “Treino”) and the testing (in Portuguese “Teste”) tasks. The *Identification* (in Portuguese “Identificação”) enables the NE identification and classification without discard the entity context. Inside *Identification* tasks have performed by rules to identify or eliminate a candidate. It also identifies the context in which the candidate appears by using a set of finite rules. The *Feature Extraction* (in Portuguese “Extracção de Caracteristicas”) task receives a list of pairs (entity, context) as input and generates a list of pairs with the entity and context-related features. The *Classification* (in Portuguese “Classificação”) receives a list from the last task and achieves a classification supported by a training algorithm (consulted in (Mota, 2008)). This task has helped by a *co-training* (in Portuguese “Co-treino”) that the classification rules using a semi-supervised approach. Finally, we have the *propagation* task added to improve the system recall by assigning the most frequent to an entity when the system did not identify it.

Chaves (2008) proposes *SEI-Geo* system to deal with the identification and classification of Location entity, based on manual patterns and geographic ontologies, also known as geo-ontologies. This system has integrated into a Geographic Knowledge Base used as a repository for integrating the extracted knowledge from heterogeneous sources.

Figure 3.6 shows the *SEI-Geo* system that starts with an *Identifier* (in Portuguese “identificador”) to identify sentences and patterns, supported by concepts and their occurrences extracted from a geo-ontologies. The *Classifier* (in Portuguese “Classificador”) that queries the geo-ontologies trying to disambiguate and identification of semantic relations. Linked to *Classifier* an *Tree Extractor* (in Portuguese “Extractor de Arbustos”) that builds a tree with NE and relations. Finally, an *Annotator* (in Portuguese “Anotador”) is capable of text annotation in the format used by the system. As already mentioned, the *SEI-Geo*

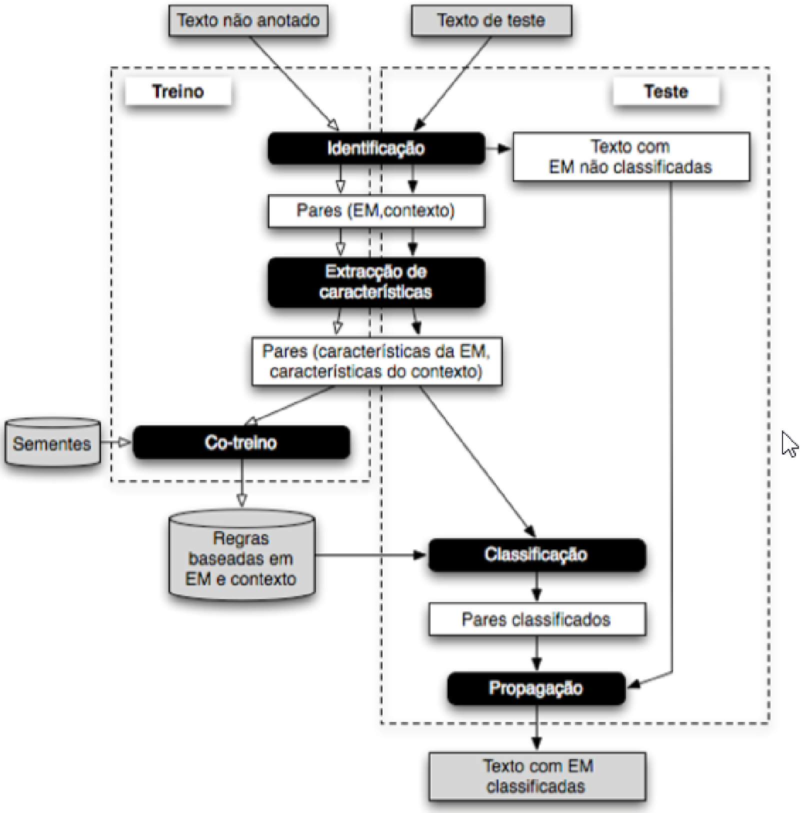


Figure 3.5: *R3M* Architecture (Mota, 2008).

system uses Geo-Net-PT [48] and WGO (World's Geographic Ontology), that contains concepts, names, and relations related to countries, cities and among others to support geographic knowledge bases.

Amaral et al. (2008) proposed a NER system, named *Priberam*, which is based on a lexicon, an ontology and a grammar. Each entry in the lexicon corresponds to multilingual ontology levels with a morpho-syntactic and semantic classification. Authors built the system supported by contextual rules, such as sequences of word combinations, grammatical categories, and lists of word combinations forming categories. These rules were applied to the NER task, considering the sequences of proper names (separated or not by some prepositions or the mentioned context).

Cardoso (2012) proposed a NER framework, called *Rembrandt*, with the purpose of document annotation by classifying the NE with unique identifiers, such as Wikipedia/DBpedia URLs. As we can see in figure 3.7, this system uses Wikipedia [49], Yahoo GeoPlanet, and DBpedia [50] as a knowledge base in order to classify the NE, associated with a set of grammatical rules, understood as standards, to extract their meaning through internal and external indications. *Rembrandt* arises from the necessity to create a system for marking texts indicating the NE related to geographical locations, such as countries, rivers, universities, monuments, or headquarters of organizations.

[48] See www.linguateca.pt/geonetpt [Accessed: 1 January 2020].

[49] See www.wikipedia.com [Accessed: 1 May 2020].

[50] See https://wiki.dbpedia.org/ [Accessed: 1 May 2020].

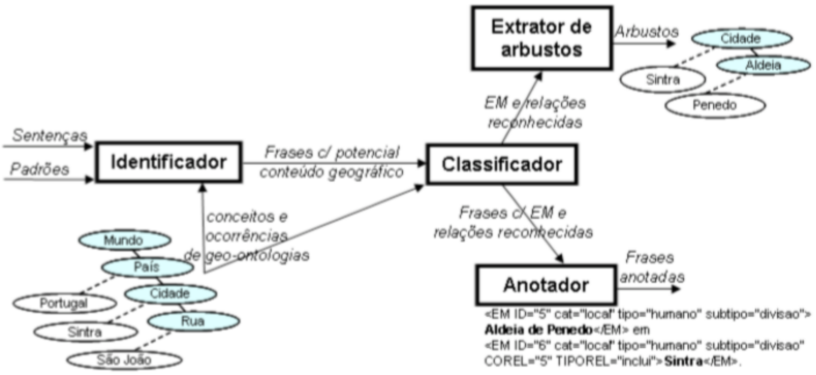

Extrator de arbustos
Arbustos
Cidade
Aldeia
Sintra
Penedo
EM e relações reconhecidas
Sentenças
Padrões
Identificador
Frases c/ potencial conteúdo geográfico
Classificador
conceitos e ocorrências de geo-ontologias
Frases c/ EM e relações reconhecidas
Mundo
País
Portugal
Cidade
Sintra
Rua
São João
Anotador
Frases anotadas
<EM ID="5" cat="local" tipo="humano" subtipo="divisao">
Aldeia de Penedo</EM> em
<EM ID="6" cat="local" tipo="humano" subtipo="divisao"
COREL="5" TIPOREL="inclui">Sintra</EM>.

Figure 3.6: *SEI-Geo* Architecture (Chaves, 2008).

Amaral et al. (2013) proposed the *NERP-CRF* system based on CRFs classifier, with two stages: train and test, using the HAREM corpus. The system achieved the best precision results compared to other systems (Priberam, R3M, Rembrandt, SEI-Geo, and CaGE) in the same corpus, proving a competitive and effective system.

Pirovani and de Oliveira (2015) an approach to identify entities, such as names, based on a local grammar built after a linguistic study of this texts. Pires (2017) for is Master Thesis evaluates existing NER tools in order to select the best tool for a NER in the Portuguese language, focus on the SIGARRA news (Porto University Portal about information news, related to the student community). The evaluation was performed based on two datasets (HAREM collection and a manually annotated subset of SIGARRA's [51] news) and calculated the information retrievals performance measures, such as precision, recall, and F-measure.

Pirovani and de Oliveira (2018) proposed the use of CRFs for NER in Portuguese texts, focused on terms classification obtained by a Local Grammar as an additional informed feature. Local grammars used hand-made rules to identify NEs. The Golden Collection of the First and Second HAREM to has used as training sets. Rodrigues et al. (2018) proposed a NER module integrated in NLPPort, named *EntPORT*, based on OpenNLP toolkit. Authors trained a supervised model based on Floresta Virgem corpus, being the NE classified as one of the following: abstract, artprod (article or product), event, numeric, organization, person, place, thing, or time.

The **Semantic Role Labelling** task in NLP that detects and classifies the semantic roles in a sentence and relates with the predicate (verb), making them dependent on it, thus, becoming an essential step towards understanding the meaning of natural language. There are SRL systems (well-studied) for languages like English, although there is a lack of these systems applied to the Portuguese language.

Bick (2007) proposed a semantic-role annotator that uses hand-written constraint grammar rules and a lexicon. Sequeira et al. (2012) presented a preliminary approach for obtaining a data-driven SRL by using Support Vector Machines and Conditional Random Fields to train the selected corpus and evaluating the three most frequent semantic roles (relation, subject, and object) with a subset of Bosque 8.0 corpus (also used for training). This approach brings some advantages regarding the hand-rule approach due to it using trained models that are not fixed to a rule or a lexicon. However, this approach evaluates three semantic

[51] See www.sigarra.up.pt [Accessed: 1 January 2020]

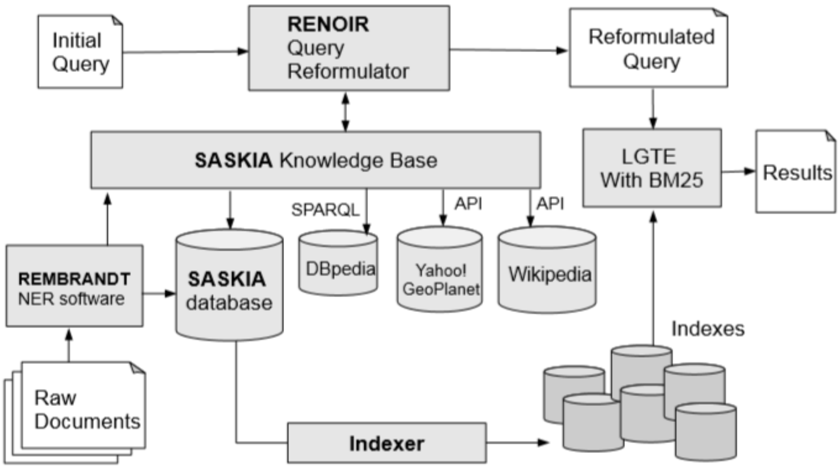


Figure 3.7: Rembrandt Architecture (Cardoso, 2012).

roles. Language has more than three semantic roles. Fonseca and Rosa (2012) proposed an adaptation of the SENNA [52] architecture for SRL, introducing the use of unlabeled data as an advantage, due to the lack of labeled resources in Portuguese. Fonseca and Rosa (2013) proposed an SRL based on two-step convolutional neural architecture (neural networks) to label semantic arguments in Brazilian Portuguese texts, called nlpnet [53]. Santos (2014) proposed a SRL module at the end of the parsing stage, as part of a NLP chain, named STRING (developed at INESC-ID Lisboa). The defined 37 semantic roles and SRL module are composed of 183 pattern-matching rules for labeling semantic roles. Like other approaches, the use of hand-rules limited system accuracy. Quaresma et al. (2019) trained a model for SRL module on top of the dependency parser (based on Freeling) using the modified data set from System-T.

3.4 Information Extraction

The documents produced during a criminal or journalistic investigation are in textual form for better understanding. Different authors have explored several tasks that aim to permit the IE from different sources, such as Relation Extraction (RE), Event, and *5W1H* extraction.

We focused on the IE approaches that deal with textual data already proposed by the academic community, mostly in the English language. However, some approaches to the Portuguese language have been proposed. The analysis of each view will support our decisions regarding the IE approaches.

3.4.1 Relation Extraction

RE is an approach to identify and classify relationships between entities identified in the text (Singh, 2018). In the following paragraphs, we described the existing approaches to solve RE issues applied to the Portuguese language and used to our domain.

[52] See www.ronan.collobert.com/senna [Accessed: 1 October 2019].
[53] See www.nilc.icmc.usp.br/nlpnet [Accessed: 1 September 2019].

Mota and Santos (2008) proposed the *SEI-Geo* system that is integrated in a Geographic Knowledge Base. *SEI-Geo* system aims to recognize *part-of* relation between geographic entities using hand-crafted patterns supported by linguistic features. To enable this, two ontologies were added, the *Geo-Net-PT* and *World Geographic Ontology*, by using entities-pairs identification mapped with ontologies. Mírian Bruckschen et al. (2008) developed a system, named by *SeRELeP*, to recognize three different relationships types: *occurred*, *part-of*, and *identity*. To accomplished this, heuristic rules using linguistic and syntactic features generated by PALAVRAS (Bick, 2006b) system were applied. Cardoso (2008) jointed with *Rembrandt* system, a feature that identify 24 different relations types by using hand-crafted rules based on linguistic features supported by two Knowledge Bases (KBs): DBpedia [54] and Wikipedia. Beamer et al. (2008) aims to develop a learning model that identifies and extracts noun features from WordNet *Is-a* backbone. Therefore, a semantic interpretation system of type-B relies on *WordNet* semantic features, employed by a supervised learning model.

Rodrigues et al. (2010) proposed architecture for information representation using an ontology, applied to government documents (Portuguese Municipal Boards meetings). The architecture has supported by an NLP pipeline with a NER module. A trained classifier has proposed using enhanced text to detect information patterns to notice relationships between entities. This architecture was the focus of the Portuguese language. However, for convenience, the authors translated them into English. They followed for a relation extraction done by other authors, such as *LEILA* (Suchanek et al., 2006), where the system can extract instances of arbitrary binary relations by using a grammar link parser for syntactic analysis. Oliveira et al. (2010) aims to extract five types of relations: *synonymy*, *hyperonymy*, *part-of*, *cause* and *propose*, between modified terms by adjectives or prepositions, based on hand-crafted patterns using lexical and syntactic features. This system was applied against Wikipedia texts, obtaining information that relies on the *Onto.PT* [55] which is a lexical network for Portuguese in the fashion of *WordNet* [56]. Lagos et al. (2010) defines an approach to semi-automatically extract information from a collection of lawyers documents, based on NLP techniques, by extracting the relations between key players (identified as NE) in a certain case by the shape of the events. An Event Recognition Module was proposed and supported by Xerox Incremental Parser [57] (XIP), with a NER system that detects and semantically classifies proper nouns associated with events. They based the Event Detection Module followed an approach that describes an event as *"a predicate (verb, adjective, and predicative noun) related to its arguments and modifiers."*.

Garcia and Gamallo (2011) proposed a system that approaches the impact of different features in extracting occupation relationships instances over Portuguese texts. Each training sentence has been evaluated by extracting each word, the lemma, and the POS tags. The system also computes the syntactic dependencies between words using a syntactic parser. The training sentences has used to train an SVM classifier.

Santos et al. (2012) proposed a system, named *News2Relations 1.0*, that extract relations in a form of triples - *(Subject, Predicate, Object)*, applied to news titles. The system extracts certain attributes, like adjectives, adverbs, and local negation, from the relationship's main elements and extracts the inter-relation between two adjacent relationships.

Souza and Claro (2014) developed an supervised Open Information Extraction (OIE) approach that extracts triples from Portuguese texts. To train and evaluate classifiers, authors used 500 annotated sentences extracted from corpus CETENFolha [58], where positive and negative examples of relationships are labelled. Taba and de Medeiros Caseli (2014) reviews the semantic relations, and how can be automatically extracted from Portuguese texts. The two main approaches reviewed are based on (i) textual patterns and

[54] See www.wiki.dbpedia.org/ [Accessed: 1 November 2018].

[55] See www.ontopt.dei.uc.pt/ [Accessed: 1 November 2018.]

[56] See www.wordnet.princeton.edu/ [Accessed: 1 November 2018.]

[57] See www.string.hlt.inesc-id.pt/w/index.php/XIP [Accessed: 1 May 2020].

[58] See www.linguateca.pt/cetenfolha/ [Accessed: 1 November 2018.]

(ii) supervised methods. Thus, this work investigates how to extent these two approaches to be applied for automatic extraction of seven binary semantic relations, such as *is-a*, *part-of*, *location-of*, *effect-of*, *property-of*, *made-of*, and *used-for*. For training, two trained corpus was used, such as CETENFolha and FAPESP [59], were used PALAVRAS system for annotation.

Sena et al. (2017) aims to extract facts in Portuguese without pre-determining the types of facts or modifying the approach proposed by Fader et al. (2011), with some changes in the inference process. Their approach has focused on the new facts that could be extracted from an inference process (identification of transitive and symmetric issues) to increase the quantity of extracted facts using open information extraction methods. Dividing the method in four-folds like the Syntactic Constraint, Inference Classifier, Transitivity Constraint and Symmetric Constraint. Gamallo and Garcia (2017) proposed a suite of tools, called *Linguakit*, for extracting, analyzing, and annotating that permits the integration into a big data infrastructure. Several tasks have introduced, such as POS, syntactic parsing, coreference resolution, and including relation extraction applications, sentiment analysis, summarization, extraction of multi-word expressions, or entity linking to DBpedia. *Linguakit* has applied to four languages, like the Portuguese, Spanish, English, and Galician languages. Related to RE, under an unsupervised approach, proposed to obtain an open set of relationships between two objects, and outputs triples *Obj1, Relation, Obj2*, selected by an OIE. *Linguakit* uses argOE (Gamallo and Garcia, 2015) as a rule-based tool that takes as input, a text analyzed in syntactic dependencies (in CoNLL-X format).

3.4.2 Event Extraction

Xu et al. (2006) propose an event-based approach to visualize the documents as a graph. Applying a graph-based ranking algorithm to illustrate the application of document graph to multi-document summarization. McCracken et al. (2006) introduces an approach that analyses the SRL used of a textual event description to understand events in summarized reports about individual people. This system has supported an approach using machine learning over the Propbank corpus with a rule-based approach using a sublanguage grammar of the summary reports. The system has been used to identify event/role usage patterns mapped to entity relations in the application's domain ontology.

Nothman et al. (2012) introduces an event linking approach using labels to enable an event reference with the article where it has first reported, which implicitly relaxes coreference to co-reporting and will practically enable augmenting news archives with semantic hyperlinks.

Siaw et al. (2013) presents an automated NLP pipeline for semantic event extraction and annotation (EveSem). The system outputs an XML annotated semantic event. This system integrates a temporal interpretation of events using the linguistic elements made available by using the given tools.

Zhou et al. (2016) proposed an approach of Chinese event extraction, namely, event argument identification. The approach introduces a SRL Based Event Argument Identification method, based on event extraction and event argument identification methods. It was then proposing an approach that extracts *5W1H* information mapped by heuristic rules. Yang and Mitchell (2016) describes a new approach that models the dependencies among variables of events, entities, and their relations and performs joint inference of these variables across a document. The main goal is to enable access to document-level contextual information and facilitate context-aware predictions.

Raiyani et al. (2019) proposed an automatic event extraction from Portuguese linked document, supported by an ontological structure that extracts events mapped as a knowledge graph that represents the NEs and the events associated with each document. Such graphs are accessible through SPARQL queries.

[59] See www.revistapesquisa.fapesp.br [Accessed: 1 December 2018.]

3.4.3 5W1H Extraction

Tanev et al. (2009) proposed event extraction systems and described its application for the Portuguese and Spanish languages. The main goals were to identify events in text, identifying *"Who did what to whom, when, with what, where, and why´´*. This proposal uses the *EMM* [60] as a multilingual news gathering and analysis system that works with 41 languages, that feeds the *NEXUS* (Tanev et al., 2008) event extraction system that integrates information and performing validation and merging. The *NEXUS* system was adapted to Portuguese and Spanish because it is original only for English, French, Italian, and Russian languages.

Das et al. (2010) proposed different methodologies to extract semantic role labels of Bengali nouns using 5W distilling. To answer the 5W questions (Who, What, When, Where, and Why), the authors used semantic information of nouns to build an annotated gold standard corpus and acquisition linguistics tools for feature extraction.

Wang et al. (2010) proposed an approach to extract semantic event elements, for *5W1H* (Who, What, Whom, When, Where and How) concepts, using a machine learning method to identify the critical events of Chinese news stories, followed by semantic role labeling enhanced by heuristic rules to extract event *5W1H* elements. Wang (2012a) proposed an event semantic elements extraction and an event ontology population supported by SRL, and the NER techniques and rule-based method, and NOEM ontology.

Chakma and Das (2018) developed an annotation of English tweets for identification of predicates and corresponding semantic roles by adopting a *5W1H* (Who, What, When, Where, Why, and How) concept which is widely used in journalism, to enable the identification and classification of questions, a SRL approach has used.

Hamborg et al. (2018) proposed an open-source tool, named by *Giveme5W*, that enables a syntax-based on five journalistic *W* questions (5Ws) extraction system for news articles. Figure 3.8 shows the *Giveme5W* pipeline.

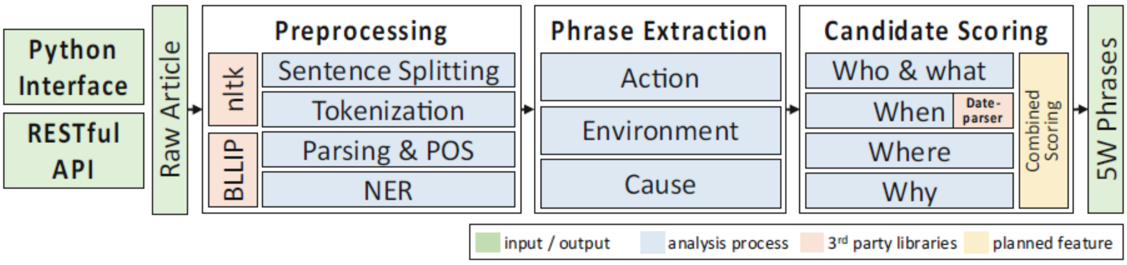


Figure 3.8: Giveme5W Architecture (Hamborg et al., 2018).

The pipeline has divided into three main phases: the *Preprocessing* uses sentence splitting, tokenization, POS, and NER as tasks; the *Phrase Extraction* performs three extraction chains to extract the events, the *action* chain extracts phrases to the journalistic "who" and "what" questions, the *environment* for "when" and "where", and *cause* for "why"; finally, the *candidate Scoring* determine the best candidate of each 5W question.

[60] See www.emm.newsbrief.eu/overview.html [Accessed: 1 July 2020].

3.5 Deep Learning

Deep learning is an emergent field applied to different applications, such as image processing or NLP tasks, and is used to extract knowledge from data with more meaning and accuracy.

Lydia and Seetha (2019) proposed to use deep learning techniques for timely analysis of crime, using a neural network (Deep Neural Network) prediction of crime rates and crime hotspots to improve proactive policing and prevention measures in India. M. Ramirez-Alcocer et al. (2019) proposed an approach to classify incidents of a crime of public safety through predictive analysis. Using a predictive model is based on a neural network Long Short-Term Memory (LSTM) trained with a small group of attributes. This approach was evaluated using a data set (real data) of OSINT source, containing information about the crimes.

Azeez and Aravindhar (2015) presents a visual analytics approach to provide decision-makers with tools, like proactivity and a predictive environment, to make effective resource allocation and deployment decisions. This is supported by historical crime records and various geospatial and demographic information. Also, it adds a semantic analysis and natural language processing of Twitter posts via latent Dirichlet allocation, Topic detection Sentiment analysis.

Regarding to works that used the Portuguese language, Souza et al. (2019) trained Portuguese BERT [61] models and employed a BERT-CRF architecture to the NER task on the Portuguese language, combining the capabilities of BERT, like transfer, with the CRF structured predictions. The BERT models were trained with the HAREM I dataset. Fernandes et al. (2018) describes the challenges, limitations, and benefits of applying the deep learning to NER in the Portuguese language. The corpus HAREM I (annotated data) were used to train/test the NER models. For word embeddings training, the authors used non-annotated data. To obtain pre-trained word embeddings, sets of raw textual data were required and retrieved from Portuguese Wikipedia articles. For training, two different embedding training architectures were used, such as the Word2Vec and Wang2Vec.

3.6 Graph-Based Approaches

The approaches using GDB have proposed for KR regarding the criminal domain. They could be used as a reference to help us approach and solve new problems of a similar nature.

Arora and Mishra (2014) support their work on a graph mining approach for graph structure to obtain frequent patterns of information. By applying this technique, the authors aim to measure a person's probability of committing a stock market crime. A graph database was developed using NetBeans [62] for algorithm implementation, *Neo4j* for graph database, and NeoEclipse [63] for graph analysis. Therefore, the method is the hotspot identification by searching, by using graph patterns in the graph database by processing the crime-related datasets into a transaction based dataset.

Alshammari and Alghathbar (2017) proposed the *CLogVis* system for a crime data analysis and visualization that helps police departments and security agencies to connect criminals and suspects by using their cell phones data. Figure 3.9 shows the ClogVis architecture that extracts data from a suspect or criminal devices and logs files as text files that form the following datasets: Phones, PhoneBook, CallLog, TSPtempCalleLog, CrimeContext, and CrimeType. The data has been cleaned, and weights are assigned based on a predefined list of communication patterns. A graph is constructed for a final representation in an interactive and

[61] See https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html [Accessed: 1 May 2021].
[62] See www.netbeans.org [Accessed: 1 July 2020].
[63] See www.github.com/neo4j-contrib/neoclipse [Accessed: 1 July 2020].

informative shape to enable investigators to obtain the possible evidence.

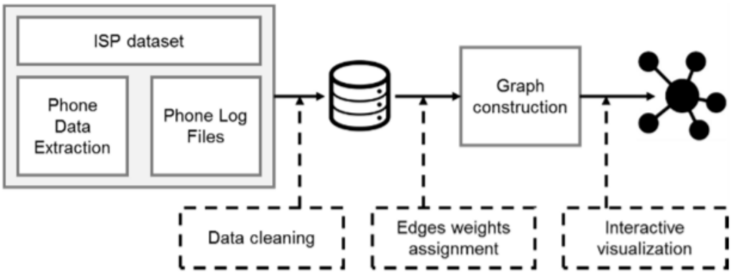


Figure 3.9: ClogVis Architecture (Alshammari and Alghathbar, 2017).

Dasgupta et al. (2018) proposed a system for the automatic extraction of crime-related information from crime-reporting news articles, using a deep convolution recurrent neural network model to analyze crime articles to extract crime-related entities and events. A crime ontology was setup to maintained the knowledge extracted. This system's outcome is a crime register that contains details of the crime committed across geographies and time.

Geepalla and Abuhamoud (2019) propose a model that uses graph technologies to analyze Call Detail Records to find potential criminals. Call Data Record is a data structure that stores relevant information about a given phone telecommunications activity involving a user of a telecommunications network. After this, it analyzes these databases, finds various links between various suspects (mobile numbers), and generates them as output, concluding this analysis. Figure 3.10 details a graph model to represent a phone call between potential criminals.

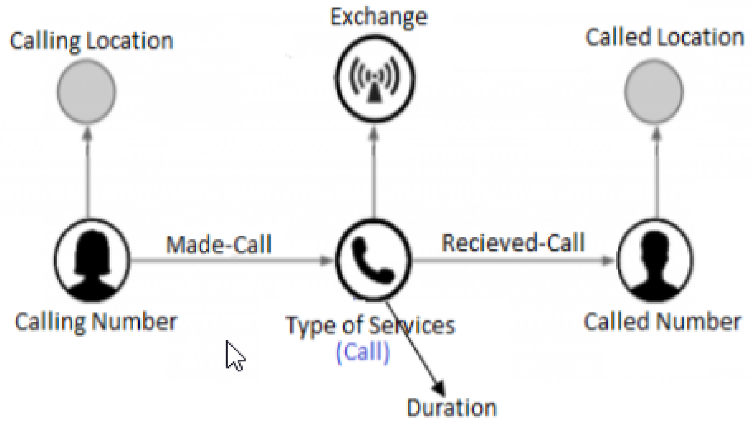


Figure 3.10: Phone Calls Graph Model (Geepalla and Abuhamoud, 2019).

Das et al. (2019) proposed an unsupervised approach for extracting relations from newspapers based on crime data. This approach established relations between persons, organization, and locations, where a similarity score are measured based on the intermediate words. A weighted graph has been built, and a similarity threshold is set to partition the graph based on the edge weights. Figure 3.11 shows the flowchart that explains the approach methodology.

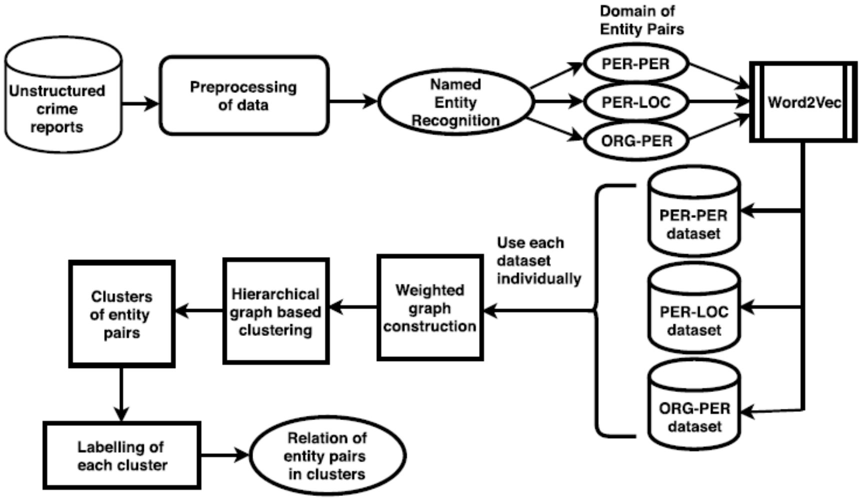

Domain of Entity Pairs
Unstructured crime reports
Preprocessing of data
Named Entity Recognition
PER-PER
PER-LOC
ORG-PER
Word2Vec
PER-PER dataset
PER-LOC dataset
ORG-PER dataset
Use each dataset individually
Weighted graph construction
Hierarchical graph based clustering
Clusters of entity pairs
Labelling of each cluster
Relation of entity pairs in clusters

Figure 3.11: Graph-Based Approach Flowchart (Das et al., 2019).

3.7 Concluding Remarks

This section summarizes the analysis of the several approaches described by pointing out their strengths and weaknesses. The main conclusions are presented below, organized by the sequence of related work sections. In section 3.1 we have described the end-to-end frameworks applied to the criminal domain:

- partnerships have been presented between universities and other public institutions, such as police departments and faculty departments (Chen et al., 2003)(Albertetti and Stoffel, 2012), which allow us to use classified data. These partnerships permit access to classified data. (Mata et al., 2016) uses mixed data sources from official and public sources;

- authors proposed pipelines developed from scratch, or configurable (Wiedemann et al., 2018) already proposed, to deal with data extracted from different sources, such as structured (relational databases) or unstructured data (textual data) in several file formats. However, the authors didn't give a good description of theses tasks, giving an idea of a minor activity;

- the use of relational databases (Albertetti and Stoffel, 2012) or ontologies, such as (Adderley et al., 2014) or (Onnoom et al., 2015) for knowledge representation. These approaches introduce two different issues: the databases for the criminal domain fail to represent its unstructured data; the ontologies are fitted to the domain but are time-consuming and difficult to build from scratch;

- the use of external knowledge bases for data enrichment, like the *GeoNames* [64] geographical database (Mata et al., 2016);

- the frameworks analyzed have focused on unstructured data (textual data), such as police reports. The use of NLP tasks have been considered for different approaches, such as Gianola (2020). These approaches are conditioned by the applied language, which motivates the proposal of new NLP methods applied to the Portuguese language in the criminal domain. The analyzed frameworks used, mostly, the English language;

[64] See www.geonames.org [Accessed: 1 January 2020].

In section 3.3.1 we have described the NLP applied to the criminal domain:

- the identification and classification of *NEs* related to the criminal domain, such as narcotics, crime types, weapons, vehicles, scenes, clothes, shoes, physical features. To identify and classify NEs use different learning approaches, such as supervised or unsupervised learning. However, the NEs are linked to the language, which needs different approaches such as annotated corpus for training;
- The use of NLP methods to deal with expressions used by criminals, named as slang. This is an issue because it makes it difficult to identify relevant information, leading to misnote terms out of context.

In section 3.6 we have described the GDB applied to the criminal domain:

- the works are focused on the English Language;
- the studied representations are focused on the representation of entities and defined relations. This approaches have been proposed to resolve specific issues, such as the representation of phone calls between persons.

We have described the related work that can be used, by adaption or inspiration, to our approach (sections 3.2, 3.3.2, and 3.4):

- the ETL approaches can be applied to different domains by adaption to resolve a specific problem to the study environment — by allowing the adaptation of the same or the use of available tools for configuring environments of data extraction and associated tasks;
- the lexical and syntactic analysis of documents written in the Portuguese language is not dependent on the analyzed domain. Therefore, this analysis is transverse to all domains, with few adaptations;
- the NER approaches applied to different domains reveals a set of NEs that could be applied to our domain, such as persons, organizations, locations, time/date, events, objects, products, artprod, or thing;
- the use of annotated corpus for training and testing of the NER or RE approaches are available for download, despite not having one adapted to the studied environment;
- the relation extraction has several approaches based on hand-crafted rules through supervised techniques. Some of them are supported by KBs to extract relations focused in several domains;
- the *5W1H* approach allows us to answer the *5Ws* and *1H* questions. This approach is a technique used in the journalistic domain and despised by the criminal domain. Therefore, it can be adapted to help in the interpretation of events in criminal-related documents.

We have analyzed previous works and retained the improvements proposed. However, from the analysis, some gaps were identified in the analysis of criminal-related documents:

- the lack of works that study the interaction between computer science and criminal domain in the Portuguese language, such as criminal investigation reports;
- the use of an ETL approach with adaptations to the criminal domain, must be included in our framework. The amount of documents that need to be processed needs an approach that automatically extracts, transforms, and loads data to a computer-readable format;

- in the analyzed NER approaches, we detected a set of domain-independent NEs that are useful for our domain, such as persons or organizations. The use (with adaptions) of such approaches to the criminal domain in combination with NEs that are related to the criminal domain need a closer study;

- the lack of trained models for NER applied to NEs related to the domain, such as narcotics;

- the *5W1H* approach is a viable path to be used in our problem and the criminal domain. However, no relevant literature applied to the Portuguese language and domain have been found, which highlights the lack of proposals in the Portuguese language, and for the criminal domain;

- Gianola (2020) embraced a similar problem to ours when using the IBM™ I2 Analyst's Notebook tool by the Police departments, where some gaps related to the processing of textual documents have identified, such as witnesses interviews. The approach used the NE identification and classification supported by a set of rules using Unitex/GramLab [65] tool. Even having common points, it uses the French language and has some gaps regarding NEs extraction. However, no reference to relations or events extractions gives room for improvement.

[65] See www.unitexgramlab.org [Accessed: 1 November 2020]

4

SEMCrime Framework Overview and Preprocessing

"The more original a discovery, the more obvious it seems afterwards."

Arthur Koestler

This chapter presents the SEMCrime framework from the initial approach to the final architecture. The framework uses as input the criminal-related documents, namely their origin, size, and structure. The Preprocessing Criminal-Related Documents is the first module of our proposal, and it is divided into the Document Processing and the NLP Pipepline. Document Processing enables the transformation of unstructured into semi-structured data using a XML file and performs a group of tasks that allow us to extract, clean, and load the plain text into a XML file. NLP Pipepline module is a set of NLP tasks introduced to perform syntax analysis over data, such as tokenization. Abbreviations and acronyms detection methods were also proposed to enable normalization of each detection's extended form.

We start by discussing the challenge presented in Chapter 1 with an excerpt from a criminal investigation report (person and location were changed to maintain anonymity) to achieve an end-to-end framework that allows the representation of unstructured data retrieved from the criminal-related documents.

> *"Em 18 de Abril de 2008 , durante a busca ao domicilio do Pedro Silva , sita na Rua Rui Leite , no Bairro de Santa Apolónia , em Coimbra , foi encontrado e apreendido no interior da carteira do arguido: uma pequena lingua de haxixe , com o peso de 1,8 gramas ."*

The sentence above outlines some challenges that we need to understand and fulfill on the framework:

- the text must be retrieved from the criminal-related documents in their original file formats, like the Microsoft™ Word, Portable Document Format (PDF), and Hypertext Markup Language (HTML);
- the text retrieved could have errors, which can cause problems with interpretation by later tasks, such as tokenization;
- the text must be understood from a lexical and syntactical perspective, like tokenization or sentence splitting;
- identifies entities such as peoples and locations names, references to dates and times, and entities related to the criminal domain, such as narcotics. The extraction of these entities are useful to understand the documents semantically;
- the relations between entities need to be identified and extracted because they enable the understanding of the meaning of each sentence;
- the data retrieved must be represented in a structured form, such as a graph database, to permit end-user queries.

Figure 4.2 shows the defined five tasks to deal with our challenge:

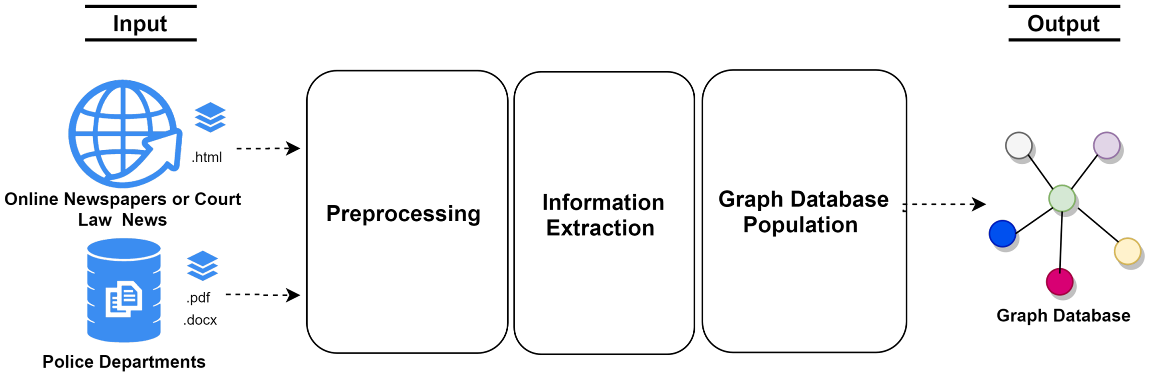


Figure 4.1: Our Initial Approach.

The meaning of each phase are described below:

- the documents feed the framework as **Input**;

- the **Preprocessing** module to extract data from documents, and perform syntactic tasks, like sentence splitting;
- an **Information Extraction** module to perform semantic tasks over the extracted data;
- an **Graph Database Population** module populates the graph database with the extracted data;
- the **Output** where data is represented into a graph database.

However, the phases proposed are too high-level, so we need to detail them. We propose a framework that automatically understands the criminal-related documents without losing their semantics during several computational methods and then populates a graph database. We established the following key steps:

- the path followed to observe the documents (similar to a criminal investigator task), by trying to identify the answers to the five ***W*** and one ***H*** questions (*5W1H*) described in each sentence of criminal-related documents, such as, *Who did What, When, Where, Why, and How* (nevertheless, the last one was discarded from this proposal). To enable this:
  - we need to identify and classify the semantic roles, the NEs, and related domain-specific terms to the criminal domain for each sentence.
- the results must be represented in a *Neo4j* graph database. For this, we have established a population and enrichment method.

The final proposal for the *SEMCrime* framework is built by the following modules, as shown in figure 4.2:

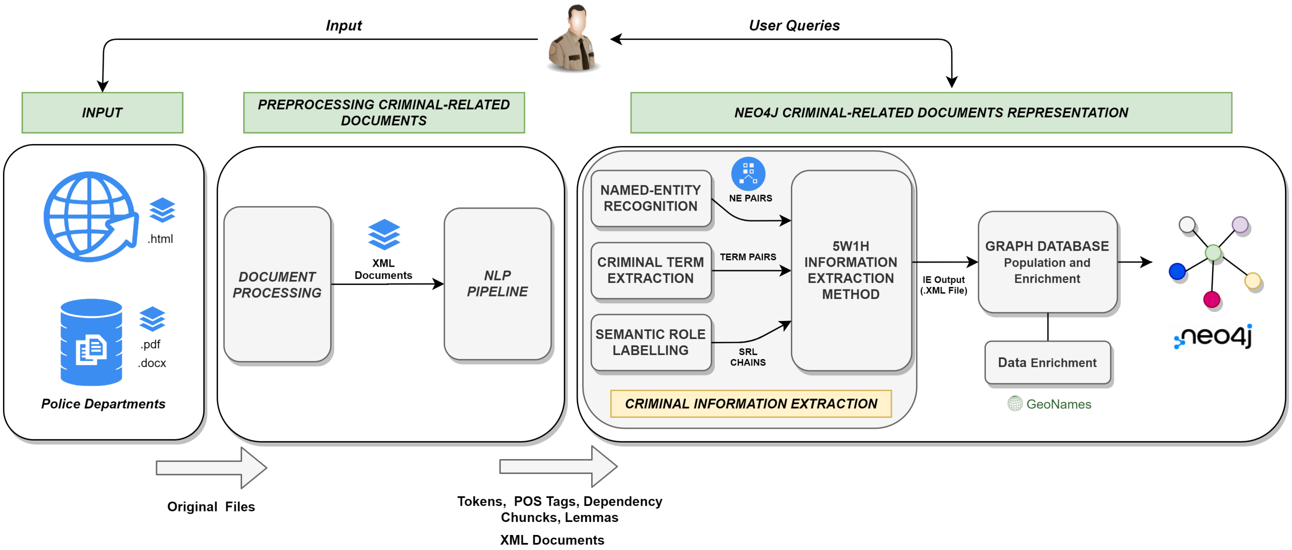


Figure 4.2: *SEMCrime* Framework Architecture.

- *Input*: takes as input a set of documents related to crime, obtained from online newspapers and police departments, in its original formats (Microsoft™ Word, PDF or HTML file formats);
- *Preprocessing Criminal-Related Documents*: formed by a pipeline with:

  - *Document Processing*: this module permits the extraction, transformation and loading of documents. For example, using a cleaning task (extracting words or symbols that may cause "noise" or not relevant). The output of this module is a XML format (a machine-readable format);
  - *NLP Pipeline*: permits the syntactic analysis of documents. This module outputs tokens, POS Tags, dependency chunks and lemmas.

- *Neo4j Criminal-Related Documents Representation*: this module is divided into two modules:
  - *Criminal Information Extraction*: uses a *Named-Entity Recognition* module to identify the NEs relevant to the domain, an *Criminal Term Extraction* was introduced to extract domain-specific terms that are relevant to the criminal domain, and a *Semantic Role Labelling* module to identify the predicate and its semantic role that will be used in *5W1H Information Extraction Method* that aggregates the other two modules to deliver the identification of the *5W1H* information and crime type detection in documents. This module outputs an *Information Extraction XML File*;
  - *Graph Database Population and Enrichment*: this enables the population of the Neo4j graph database, and the data enrichment using the *GeoNames* [1] geographical database.

For the implementation of each framework component, an initial system version of our framework was proposed, which will enable us to explore the design issues, functionality, and evaluation measures. During this study, we have felt that when we have good tools and approaches that fit our objectives in the study domain, therefore, we need to follow *Occam's razor* principle, where that *"entities should not be multiplied without necessity"* (Schaffer, 2015) or *"the simplest explanation is most likely the right one"* (Schaffer, 2015). Thus, we have selected the tools and approaches that satisfy our requirements. Our prototype was developed using Java [2] and Python [3] programming language, *Apache Tika* toolkit for Java, *Newspaper3k* for article scraping and curation, *NLPNET* a python library for NLP tasks based on neural networks, *Apache OpenNLP* toolkit, and the *NLPPort* toolkit. Along with the following chapters, we will present the adaptations performed to fit our requirements and each tool and the approaches used in more detail.

The rest of the chapter is organized as follows. In section 4.1 described the criminal-related documents by defining their origins, formats, and structures; in section 4.2 describe the preprocessing tasks performed over the documents, such as the document processing (an adapted ETL pipeline) and the NLP pipeline.

4.1 Criminal-Related Documents

The **Criminal-Related Documents** is a set of textual documents constituted by Criminal Investigation Reports, Criminal News, and PGdLisboa News, retrieved from different origins. For example, in a criminal investigation ending, a report is developed with all relevant data to the investigation; this is all written in a **Criminal Investigation Reports**. Also, investigative journalists report specific police investigations in online newspapers about particular crimes, named by **Criminal News**. Another source of documents is the Procuradoria-Geral Distrital de Lisboa [4] website [5] (news section) that reports the ongoing or finished investigations, named by **PGdLisboa News**. The previously mentioned share some features, such as describing a specific crime and the involved actors. The choice falls on these documents because they intended to answer the following five questions: the *"Why?", "Where?", "When?", "What?" and "Who"*

[1] See www.geonames.org [Accessed: 1 September 2020.]

[2] See www.java.com [Accessed: 1 July 2020].

[3] See www.python.org [Accessed: 1 July 2020].

[4] In English: District Attorney General of Lisbon.

[5] See www.pgdlisboa.pt [Accessed: 1 May 2020].

and *"How?"* (Braz, 2019), which are intrinsically related to the investigation process, whether it be the criminal investigation or investigative journalism.

The research community does not have, to our knowledge, developed a gold standard (training and test corpus) of annotated criminal-related documents written in Portuguese, which can be used as a standard for evaluating an automatic knowledge extraction system. Therefore, developing a manually annotated corpus for training, validating, and evaluating is of extreme importance. Table 4.1 enumerates the criminal-related documents, as our corpora, to be analyzed in the different stages of the framework, thus allowing a suitable corpus as validation data, minimize the problem of not having an annotated corpus in the Portuguese language related to the criminal domain.

Corpus	Words	Characters	Sentences	Texts
Criminal News	18192	111637	667	80
PGdLisboa News	16020	100720	533	80
Criminal Investigation Reports	4781	24567	380	3
Total	**38993**	**236924**	**1580**	**163**

Table 4.1: Criminal-Related Document Corpora.

The documents analyzed are in Microsoft™ Word, PDF, and HTML file formats. Due to the reason that we are extracting data from different sources, different encodings are present. Thus, we need to apply a unique encoding, and other encodings must be changed into it. Our choice falls on to the UTF-8.

The remainder of this section is organized as follows: Section 4.1.1 we describe the crime investigations reports; Section 4.1.2 describes the crime news reported in online newspapers, and the news retrieved from Procuradoria-Geral Distrital de Lisboa website (news section).

4.1.1 Criminal Investigation Reports

The criminal investigation reports synthesize in one or multiple documents the information collected during a criminal investigation by grouping the contents of an investigation, such as the witnesses, the suspects, the police investigators, or the fact descriptions. Analyzing these documents is a complex task, mainly because they are closed or classified documents. In addition, the documents information that identifies persons, phone numbers, and other sensitive data were anonymized due to confidentiality constraints.

Figure 4.3 describes the data layout structure regarding criminal investigation reports. A layout structure was defined by selecting a set of tags for data identification. Thus, the document's name, author(s), publication date, crime process number (internal police number to identify each report), title, and document body (facts description, suspects identification, and conclusions) were defined for document annotation. Images were discarded from our evaluation.

From the manual analysis of the documents, we detect several challenges that computational methods could face for a machine-understandable analysis of criminal investigation reports, as shown below:

- the existence of a person's name in the following formats:
  - using its full name, where all letters are in upper or lower case;
  - using partial names;
  - using only initials, like an acronym;

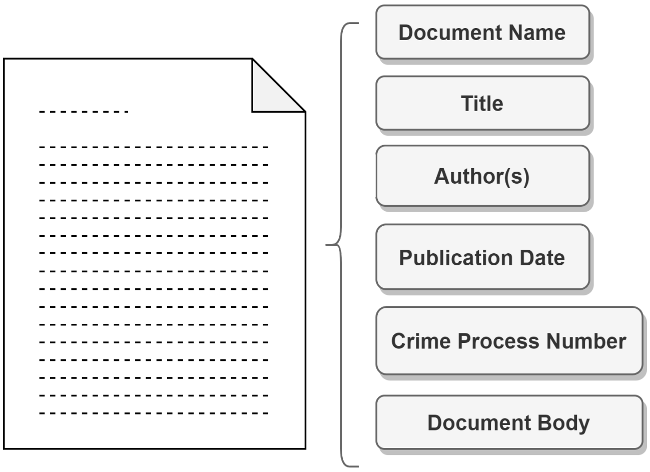


Figure 4.3: Criminal Investigation Report Layout.

  - using its street name or nickname. For instance, Alphonse Gabriel Capone was known by the nickname "Scarface".

- the date format is not standardized, so different formats are displayed, such as 12.03.2003 or "March 12, 2003";
- the use of abbreviations, where some of then are related to the domain;
- the use of acronyms, where some of them are related to the domain;
- the address names are not normalized and contain abbreviations and numbering, such as *"Rua Miguel Braga, nº 644, R/C d.to - Coimbra"*;
- the measures are not standardized;
- the use of slang that is annotated between quotation marks.

4.1.2 Criminal and PGdLisboa News

The report of a crime is spread all over the media, whether in the form of text, images, or videos, performed by investigative journalists and published in online newspapers (Wiedemann et al., 2018) as during criminal investigations performed by police departments. Criminal domain experts advised using these documents by arguing that they followed the same narrative form and the same requirements of the criminal investigation reports. Therefore, the texts are included in the criminal-related documents because they contain relevant or significant information to the study domain.

The online newspapers have news regarding crimes, like the description of facts, their actors, and pieces of evidence. For our case study, we manually crawled and extracted, using scrapping tools, news from online newspapers Diario de Coimbra [6], Jornal As Beiras [7], Diario de Noticias [8], and Publico [9].

[6] See www.diariocoimbra.pt [Accessed: 1 May 2020].
[7] See www.asbeiras.pt [Accessed: 1 May 2020].
[8] See www.dn.pt [Accessed: 1 May 2020].
[9] See www.publico.pt [Accessed: 1 May 2020].

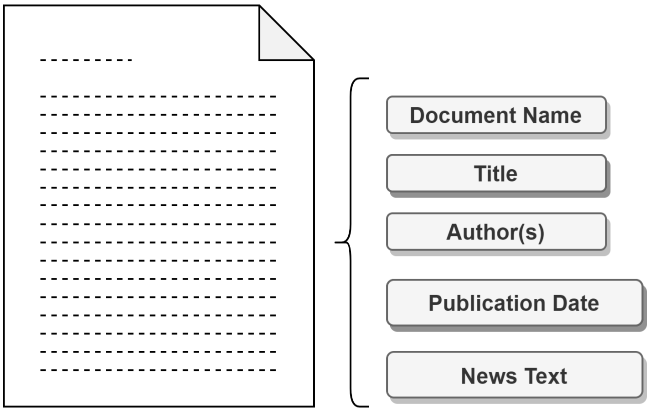


Figure 4.4: Criminal and PGdLisboa News Layout.

Figure 4.4 shows the layout structure extracted from crime news (PGdLisboa news followed the same schema), such as Document Name, Author(s), Publication Date, Title, News Text. Tables and images were discarded.

One of the Portuguese sources for criminal reports is the Procuradoria-Geral Distrital de Lisboa website. The existing news is about crimes already finished. Since PGdLisboa news follows the same format as Criminal News, we use the same layout for data representation (see figure 4.4).

4.2 Preprocessing Criminal-Related Documents

The introduction of this module is deeply connected to the necessity of criminal investigations that involve the dealing of large amounts of data from different sources. This continuous growth of the data makes data more appealing but difficult to deal with by police departments (Wall, 2018). The data are present across web pages, such as the Court's Law or the online newspapers, in the criminal investigation reports produced by criminal investigations upon police departments. Therefore, these documents are written in a free text form, as unstructured data, and in multiple formats, namely in Microsoft™ Word, PDF, and HTML file formats. These kinds of documents were named as criminal-related documents (detailed in Section 4.1). Thus, the criminal-related documents needed to the extracted, transformed and loaded into a common area, using a configurable and well-known process called ETL. Another important step in understanding the written text is the introduction of NLP task that performs a lexical and syntax analysis of each text, such as sentence splitting, tokenization, or lemmatization.

This module is divided into two modules: **Document Processing** module focused on the adaptation of an ETL approach to identify the required mappings and transformations that need to be done automatically and perform operations that lead to a transformation of unstructured into semi-structured data (represented in a XML file). Followed by the **NLP Pipeline** module enables the NLP tasks regarding lexical and syntax analysis of each document in the Portuguese language.

The following items resume the chapter contribution:

- the metadata annotation of the criminal-related documents that enable the content structuring in a XML file format;
- an approach to the abbreviations detection based on a dictionary with abbreviations (that include

common abbreviations on the criminal domain);

- an approach to the acronyms detection based on a dictionary with acronyms (that include common acronyms on the criminal domain);
- introduce an ETL adapted to module needs to be normalized into a unique file format and document content (by transformation and cleaning tasks);
- implements a methodology to clean up “noisy” data based on regular expressions;
- propose a functional NLP pipeline based on tasks to understand the syntax of each analyzed text document;
- adapted state of the art approaches in the proposed NLP pipeline, and made improvements regarding the Portuguese language context:
  - add an elision terms list to increase tokens detection.

The remainder of this chapter is organized as follows: the Section 4.1 presents criminal-related documents by describing each group of analyzed documents; in section 4.2.1 explains the document processing module that is the focus of this chapter; in Section 4.2.2 describes the NLP tasks introduced for lexical and syntax analysis; Finally, the concluding Remarks are performed in Section 4.3.

4.2.1 Document Processing Module

The criminal-related documents needed to be extracted and transformed into a unique format after a set of tasks that are common in ETL frameworks. This is a ”tailor-made” (customized to our needs) module, where the specificity of application to a criminal investigation is considered and has a considerable impact on the proposed artifact. This processing activity was performed offline. The pipeline was built from scratch, using the Java programming language and some available frameworks, such as Apache Tika [10] toolkit. Figure 4.5 shows the *Document Processing* module divided into three phases:

- *Input*: the criminal-related documents in original file formats;
- *Extraction*: set up the access task (Connectors) and reading task (Reader);
- *Transformation*: performs the cleaning task supported by a set of rules, and a normalizer task based on a set of rules, and an abbreviation list;
- *Loading*: outputs the criminal-related documents in XML format to the next module;

Document Processing Module data flow is enumerated in figure 4.5:

- (1) original criminal-related documents in several formats, like Microsoft Word, PDF, and HTML;
- (2) criminal-related documents in plain text;
- (3) plain text cleaned and normalized;
- (4) built XML files following the XML Schema and loaded into the next module.

[10] See www.tika.apache.org [Accessed: 1 March 2020].

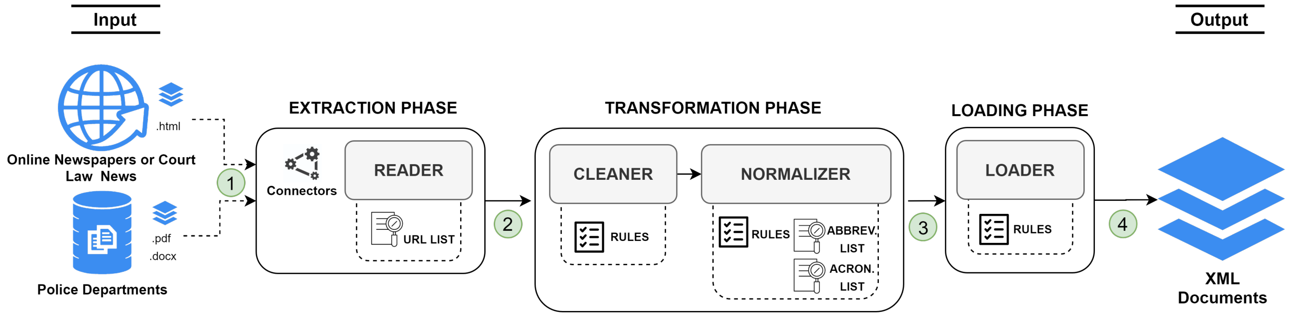


Figure 4.5: Document Processing Module.

We used a XML format for the exchange of semi-structured data. An XML document consists of nested element structures with a root element, where each one may contain other elements and attributes. An XML Schema was associated (see section 4.2.1) to the XML document, describing each element declarations and type definitions, and also relationships between entities that are represented by nesting elements or by references.

Extraction Phase

Extraction phase has two sub-tasks: *Connectors* and *Reader*. The *Connectors* make the extraction of data from the data sources. For the task execution, a group of steps was added:

- File formats identification;
- Data extraction: fetching data with an appropriated connector to the external sources;
- Data reading: the reading data stream and building an in-memory model to facilitate the *Reader* task. We defined an Uniform Resource Locator (URL) List that enables us to set up a list of online newspapers URL to be extracted from the WWW.

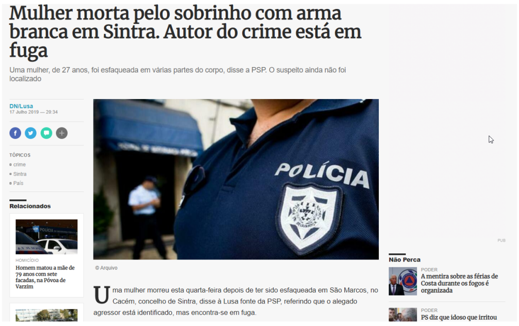
The prototype was developed in Java and Python programming languages and followed the defined methodology (such as cleaning rules). We use the Apache Tika toolkit and Newspaper3k [11] for article scraping and curation. Figure 4.6 shows an example from online *Jornal de Noticias* [12] newspaper. This shows the result of crime news after being processed by the extraction phase, in plain text. During task execution, uncleaned and malformed text could also be extracted, bringing issues to the next stage.

Transformation Phase

The *Transformation Phase* is designed with two main tasks: the *Cleaner* and the *Normalizer*. The *Cleaner* was proposed to correct, remove erroneous and misleading data, such as contradictions or disparities. This task deals with textual data to represent information and can also be deceptive in terms of "dirty data" (Chu et al., 2016) (Li et al., 2019) misleading the output results. To reduce the "garbage in, garbage out" (Geiger et al., 2019). Thus, table 4.2 describes a rules list to ensure the cleaning task. The enumerated list arises

[11] See www.newspaper.readthedocs.io [Accessed: 1 September 2019].

[12] See www.jn.pt [Accessed: 1 October 2019].



Mulher morta pelo sobrinho com arma branca em Sintra. Autor do crime está em fuga

Uma mulher, de 27 anos, foi esfaqueada em várias partes do corpo, disse a PSP. O suspeito ainda não foi localizado

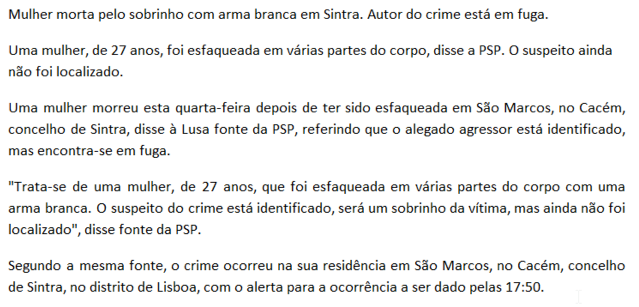
Uma mulher morreu esta quarta-feira depois de ter sido esfaqueada em São Marcos, no Cacém, concelho de Sintra, disse à Lusa fonte da PSP, referindo que o alegado agressor está identificado, mas encontra-se em fuga.

Homem matou a mãe de 79 anos com sete facadas, na Póvoa de Varzim

A mentira sobre as férias de Costa durante os fogos é organizada

PS diz que idoso que irritou

Figure 4.6: Criminal News Screenshot.



Mulher morta pelo sobrinho com arma branca em Sintra. Autor do crime está em fuga.

Uma mulher, de 27 anos, foi esfaqueada em várias partes do corpo, disse a PSP. O suspeito ainda não foi localizado.

Uma mulher morreu esta quarta-feira depois de ter sido esfaqueada em São Marcos, no Cacém, concelho de Sintra, disse à Lusa fonte da PSP, referindo que o alegado agressor está identificado, mas encontra-se em fuga.

"Trata-se de uma mulher, de 27 anos, que foi esfaqueada em várias partes do corpo com uma arma branca. O suspeito do crime está identificado, será um sobrinho da vítima, mas ainda não foi localizado", disse fonte da PSP.

Segundo a mesma fonte, o crime ocorreu na sua residência em São Marcos, no Cacém, concelho de Sintra, no distrito de Lisboa, com o alerta para a ocorrência a ser dado pelas 17:50.

Figure 4.7: Criminal News After Extraction Phase.

from the hand-analyzed documents and the incorrect or misleading data we spotted, such as extra white spaces.

Rules	Description
R1	remove spaces, line breaks, duplicate white-spaces and tabs
R2	commas are followed by a space
R3	symbols are ignored, such as cardinals
R4	each sentence contains a single end-mark
R5	remove all characters that are different from ASCII charset
R6	split attached words

Table 4.2: Cleaning Task Rules.

Another issue that arose from a manual analysis of documents is the use of different formats to represent the same information, such as date formats or currency. Table 4.3 describes the *Normalizer* rules to solve those problems.

Rules	Description
R1	convert years
R2	convert money to standard form
R3	convert street address names
R4	normalize ZIP codes
R5	normalize identifiers unique (such as sex categories Male/Female/Unknown)
R6	abbreviation detection
R7	acronym detection

Table 4.3: Normalization Task Rules.

We proposed a rule-based approach that uses a dictionary to deal with the abbreviations, such as the abbreviation “Insp. Silva” can be expanded to “Inspector Silva” (when these abbreviations are not expanded, and the sentence detection is erroneous because the dot punctuation will introduce a new line in the text). This task is performed by looking up a token in a list and comparing it against the candidate abbreviation, returning the correct abbreviation. Finally, an extract is showed in listing 4.1. However, this language-dependent list of abbreviations lacks abbreviations regarding terms founded in criminal-related documents. That is solved by extending this list with new terms because resolving abbreviations will reduce erroneous results in the following modules, such as in the tokenization task.

```
<target="abrev.">abreviação</replacement>
<target="Abr.">Abril</replacement>
<target="acad.">academia</replacement>
<target="adapt.">adaptação</replacement>
<target="adit.">aditamento</replacement>
<target="afr.">africano</replacement>
<target="Ag.">Agosto</replacement>
<target="ág.f.">água-forte</replacement>
<target="alarg.">alargada</replacement>
<target="álb.">álbum</replacement>
<target="alf.">alfabético</replacement>
```

Listing 4.1: Abbreviations List in a XML Format (extract).

Another issue that we highlight during manual document review is acronyms, such as "PJ" , which means "Polícia Judiciária". To solve this, we have applied the same method used for abbreviations: a list of general acronyms was extended with a list of acronyms [13] with general terms (extracted from a collection of Portuguese news gathered from the on-line newspaper Publico) and acronyms related to the criminal domain. The listing 4.2 shows an extract of the XML file that enables the abbreviations detection and span.

```
<target="PJ">Polícia Judiciária</replacement>
<target="PSP">Policia de Segurança Publica</replacement>
<target="GNR">Guarna Nacional Republicana</replacement>
<target="SEF">Serviço de Estrangeiros e Fronteiras</replacement>
<target="SIS">Sistemas de Informação e Segurança</replacement>
<target="OPC">Orgões de Polícia Criminal</replacement>
<target="AST">Autoridade de Segurança no Trabalho</replacement>
```

Listing 4.2: Acronyms List in a XML Format (extract).

Loading Phase

In the *Loading Phase*, we defined two tasks: to build XML documents from a defined XML Schema, and to load the XML documents into the *NLP Pipeline*. The XML format provides a way to store criminal-related documents with metadata and structural information. We defined an XML Schema for each criminal-related documents. For the Criminal Investigation Reports, see Listing 4.3 and for the Criminal and PGdLisboa News, see Listing 4.4.

```
<?xml version = "1.0" encoding = "UTF-8"?>
<xs:schema xmlns:xs = "http://www.w3.org/2001/XMLSchema">
  <xs:element name = "ReportNameID">
     <xs:complexType>
        <xs:sequence>
          <xs:element name = "documentname" type = "xs:string" />
          <xs:element name = "authors" type = "xs:string" />
          <xs:element name = "publicationdate" type = "xs:date"/>
```

[13] See www.davidsbatista.net/nlp_datasets [Accessed: 1 April 2020].

```
          <xs:element name = "cpn" type = "xs:string" />
          <xs:element name = "title" type = "xs:string" />
          <xs:element name = "documentbody" type = "xs:string" />
        </xs:sequence>
     </xs:complexType>
   </xs:element>
</xs:schema>
```

Listing 4.3: Criminal investigation reports - XML Schema.

```
<?xml version = "1.0" encoding = "UTF-8"?>
<xs:schema xmlns:xs = "http://www.w3.org/2001/XMLSchema">
   <xs:element name = "NewsID">
     <xs:complexType>
        <xs:sequence>
          <xs:element name = "documentname" type = "xs:string" />
          <xs:element name = "authors" type = "xs:string" />
          <xs:element name = "publicationdate" type = "xs:date"/>
          <xs:element name = "title" type = "xs:string" />
          <xs:element name = "newstext" type = "xs:string" />
        </xs:sequence>
     </xs:complexType>
   </xs:element>
</xs:schema>
```

Listing 4.4: Crime News and PGdLisboa news - XML Schema.

Listing 4.5 shows an example of the XML output file:

```
<?xml version="1.0" encoding="UTF-8"?>
<News1>
    <docname>News1</docname>
    <authors>['Global Media Group']</authors>
    <publicationdate>2019-07-17 20:34:00+00:00</publicationdate>
    <title>Mulher morta pelo sobrinho com arma branca em Sintra. </title>
    <newstext>O amigo levou facada e outro um tiro (...) </newstext>
</News1>
```

Listing 4.5: Crime News XML Source Format (extract).

As we can see, the XML file followed the XML Schema defined for each criminal-related document, with the defined tag annotations added to enable posterior text analysis.

4.2.2 NLP Pipeline

Most of the data related to a criminal investigation are written in natural language, which is the Portuguese language in this thesis's scope. Therefore, this section describes the NLP tasks related to syntax analysis after extracting and transforming the original documents into a XML document template. Figure 4.8 illustrates the NLP tasks that are integrated into a module design that analyses (syntactically) the textual data.

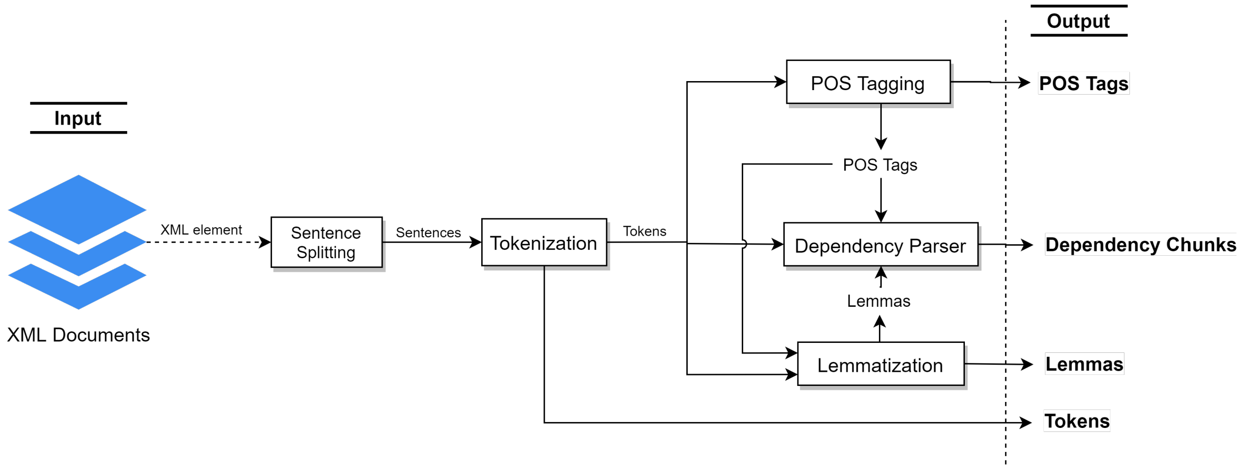


Figure 4.8: *SEMCrime* NLP Module.

Each item describes (briefly) the NLP tasks:

1. *Input*: documents from the previous module;
2. *Sentence Splitting*: splits into sentences each text document into sentences;
3. *Tokenization*: each sentence is transformed into tokens to separate all words and punctuation marks;
4. *Part-of-Speech Tagging*: each token is assigned the corresponding Part-Of-Speech tag by identifying its syntactic function in a sentence for interpretation of each token;
5. *Lemmatization*: With both the tokens and respective Part-Of-Speech tags, obtains the corresponding lemmas (base form of a word), allowing the system to use just the base form of a word;
6. *Dependency Parser*: using token together with POS tags and lemmas, dependency parsing is done by yielding dependency chunks — chunks that are built by using the dependencies between tokens, by grouping them according to significant dependencies, such as subjects, verbs, and objects.

Table 4.4 describes the input and output data for each module:

		Sentence Splitting	**Tokenization**	**POS Tagging**	**Lemmatization**	**Dependency Parser**
Input	XML documents	XML documents	sentences	tokens	tokens, POS tags	tokens, POS Tags lemmas
Output	XML documents	sentences	tokens	POS Tags	lemmas	dependency chunks

Table 4.4: Data Input/Output (Preprocessing).

Sentence Splitting

Sentence splitting is used to detect sentences that compose the documents under analysis. This task reduces the dimension of the elements for better processing by other NLP tasks, such as POS or lemmatization.

In our domain, documents are written in the Portuguese language (European-Portuguese, some words or sentences, like slang, could be in other natural languages, such as the English language). Sentence splitting aims to break down the text documents into sentences, deciding where each sentence begins and ends, despite the challenges aroused due to the potential ambiguity of punctuation marks inherent to the used language.

Figure 4.9 shows the metrics of the sentences and punctuation marks in criminal-related documents. The results denote an increase in punctuation marks on documents produced by Official channels (Courts and Police Departments), such as **Criminal Investigation Reports** and **PGdLisboa News**. This is mainly due to the author's writing style and the fact that these documents are official.

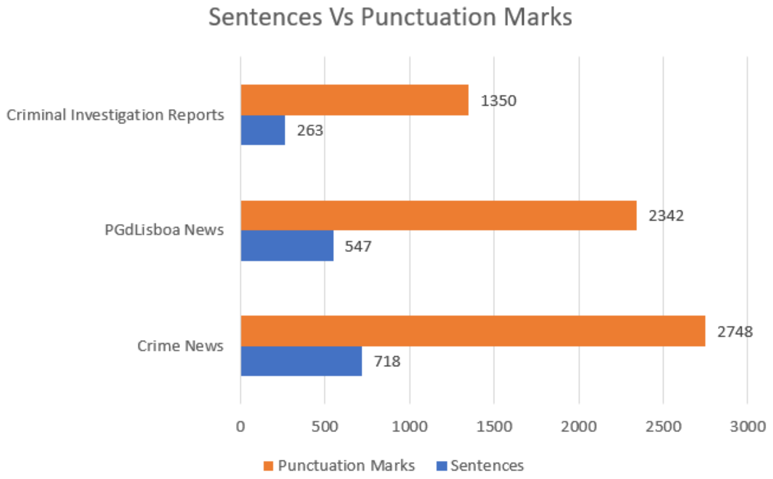


Figure 4.9: Sentences versus Punctuation Marks.

Regarding the Portuguese language, the sentences can be formed by a single sentence (simple sentence) or by more than one sentence (compound sentence). Furthermore, according to the meaning, there are several types of sentences whose purpose is more or less predictable, such as interrogative, exclamatory, and declarative (Matos, 2010). Finally, directly linked with sentence splitting, we have the punctuation marks: to make communication more effective and expressive by clarifying the understanding of the reading, rhythm, and intonation of the text. The proper use of punctuation is thus critical for the correct interpretation of the text.

Figure 4.10 depicts the sentence splitting task, with raw input and the output of the text, on individual sentences.

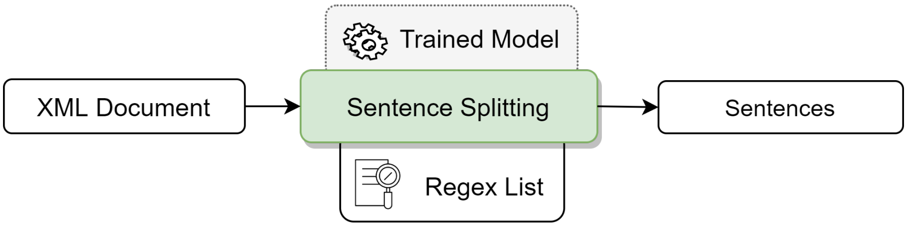


Figure 4.10: Sentence Splitting Schema.

We have selected as baseline the SenPORT (that is part of NLPPort (Rodrigues et al., 2018)), an work based on a supervised algorithm (OpenNLP train model - pt-sent.bin [14], trained with CoNLL-X Bosque [15] data), with an abbreviation and regular expressions lists. This abbreviation list has a reduced number of abbreviations with terms that are present in the criminal-related documents, already dealt in section 4.2.1.

Figure 4.5 shows the obtained results after executing sentence splitting over criminal-related documents.

	Sentences (#Count)
Criminal News	718
PGdLisboa News	547
Criminal Investigation Reports	263
Total	**1528**

Table 4.5: Sentence Splitting over Criminal-Related Documents.

Tokenization

After being split, sentences are subjected to a tokenization process. It consists of identifying the tokens (words, numbers, or punctuation marks) in a sentence. This task is done by splitting the sentence using delimiters like spaces, commas, and dots.

Figure 4.11 shows the tokenization task proposal that uses sentences as input and produces a list of tokens.

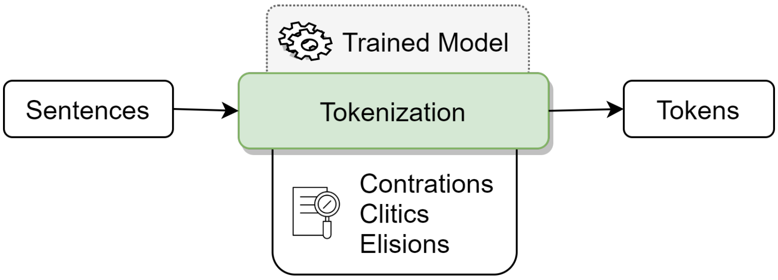


Figure 4.11: Tokenization Schema.

Our tokenization task approach is supported on the *TokPORT* (Rodrigues et al., 2018), which uses the OpenNLP pre-trained model (pt-token.bin). With the following features:

- to check the presence of contractions and clitics (pre-existing features);
- to check the presence of elision (feature added by us).

By expanding clitics in tokens, separating the verb, and the personal pronouns solves some of the issues that are related to the Portuguese language, such as contractions (an example in Listing 4.7) and clitics (an example in Listing 4.6).

[14] See www.opennlp.sourceforge.net/models-1.5 [Accessed: 1 September 2019]
[15] See www.linguateca.pt/floresta/CoNLL-X [Accessed: 1 September 2019]

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<clitics>
  <!-- Rules for breaking apart verbs with clitics -->
  <replacement target="-a"> ela</replacement>
  <replacement target="-as"> elas</replacement>
  <replacement target="-lhe-á">á a ele</replacement>
  <replacement target="-lhe-ão">ão a ele</replacement>
  <replacement target="-lhe-ás">ás a ele</replacement>
  <replacement target="-lhe-ei">ei a ele</replacement>
  <replacement target="-lhe-emos">emos a ele</replacement>
  <replacement target="-lhe-ia">ia a ele</replacement>
  <replacement target="-lhe"> a ele</replacement>
  <replacement target="-lhes-á">á a eles</replacement>
 </clitics>
```

Listing 4.6: Examples of clitics examples.

The reason to process contractions is similar to that of clitics, that is, to break prepositions and pronouns, as shown in listing 4.7.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<contractions>
  <!-- Rules for breaking apart verbs with clitics -->
  <replacement target="-a"> ela</replacement>
  <replacement target="-as"> elas</replacement>
  <replacement target="-lhe-á">á a ele</replacement>
  <replacement target="-lhe-ão">ão a ele</replacement>
  <replacement target="-lhe-ás">ás a ele</replacement>
  <replacement target="-lhe-ei">ei a ele</replacement>
  <replacement target="-lhe-emos">emos a ele</replacement>
  <replacement target="-lhe-ia">ia a ele</replacement>
  <replacement target="-lhe"> a ele</replacement>
  <replacement target="-lhes-á">á a eles</replacement>
 </contractions>
```

Listing 4.7: Examples of contractions.

However, the *TokPORT* does not work properly with elisions. For example, in the following sentence:

> *”O João viu uma cobra d’água.”*

The *TokPORT* outputs the following tokens: *”O”* , *”João”* , *”viu”* , *”uma”* , *”cobra”* , *”d”* , *”’”* , *”água”* , *”.”*.

We have to expand the word *”cobra d’água”* into *”cobra de água”* and other compound words that use apostrophes to avoid errors in tokens detection. To solve this issue, we added a lexicon with elisions (Figueira et al., 2011). Listing 4.8 lists a segment of the rules for breaking apart elision.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<elision>
```

```
  <!-- Rules for breaking apart elision -->
  <replacement target="d'água"> de água</replacement>
  <replacement target="d'Oeste"> de Oeste</replacement>
  <replacement target="d'Ele"> de Ele</replacement>
  <replacement target="d'Aquele"> de Aquele</replacement>
  <replacement target="d'Aquela"> de Aquela</replacement>
</elision>
```

Listing 4.8: Apostrophes (elision) examples added as rules to tokenizer.

Table 4.6 shows an example of the tokenization task.

Sentence	A GNR interceptou uma embarcação com duas toneladas de haxixe no Rio Guadiana.
Tokens	[A] [GNR] [interceptou] [uma] [embarcação] [com] [duas] [toneladas] [de] [haxixe] [em] [o] [Rio] [Guadiana] [.]

Table 4.6: Tokenization task example.

Figure 4.7 shows the results obtained after executing tokenization over criminal-related documents.

	Tokens
	(#Count)
Criminal News	22745
PGdLisboa News	20479
Criminal Investigation Reports	10359
Total	**53583**

Table 4.7: Tokenization over Criminal-Related Documents.

Certain groups of toked need special attention during this process, such as email addresses, website links, or license plates, that should be recognized as a whole.

POS Tagging

Each token has a piece of linguistic information that is extracted by using a POS task. The syntactic category assigned to POS Tags are adjectives, adverbs, articles, nouns, numbers, proper nouns, prepositions, verbs, and punctuation marks.

The criminal domain does not differ from the application of POS task to other domains, as long as the language is in Portuguese. Figure 4.12 shows our POS task, using the tokens as input. This task aims to analyze the text at an upper level. This task aims to be seen not just as a sequence of words (tokens) but with morph-syntactic features with meaning. In our study, we opt to use a supervised approach that uses a trained model. For our domain, our POS task is essential to identify verbs, which are valuable elements to determine events or relationships in criminal-related documents.

We adopted the *TagPORT* (Rodrigues et al., 2018), that uses the *POS tagger* available in the *OpenNLP toolkit*, with a model for Portuguese (pt-pos-maxent.bin).

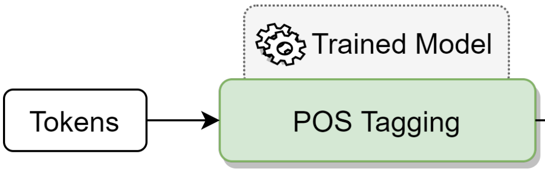


Figure 4.12: POS Tagging Schema.

As example, we used the same sentence that was processed in the tokenization task. Table 4.8 shows the POS task output:

Tokens	[A] [GNR] [interceptou] [uma] [embarcação] [com] [duas] [toneladas] [de] [haxixe] [em] [o] [Rio] [Guadiana] [.]
POS Tags	[art] [prop] [v-fin] [art] [n] [prp] [num] [n] [prp][n] [art] [art][prop][prop] [punc]

Table 4.8: POS Tagging Results.

Lemmatization

Lemmatization task enables the representation of words in their dictionary form, called *lemmas*. For Portuguese language, lemmatization may include the following types of normalization:

- noun (gender, number, augmentative and diminutive);
- adjective (gender, number, augmentative, diminutive, and superlative);
- article (gender and number);
- pronoun (gender and number);
- preposition (gender and number);
- adverb (manner) and verb (regular and irregular);
- proper nouns, numbers, interjections, and conjunctions are ignored, as they are not generally inflected in Portuguese.

The figure 4.13 shows the lemmatization schema that makes part of our proposal.

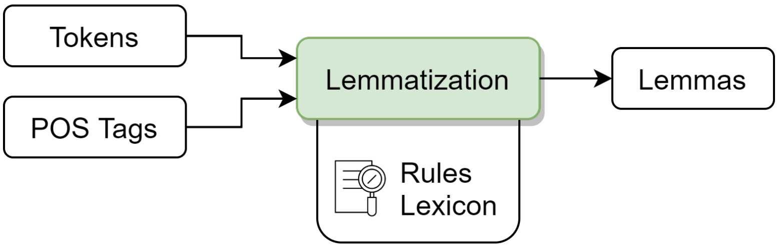


Figure 4.13: Lemmatization Schema.

We used the *LemPORT* Rodrigues et al. (2014) that achieved an accuracy $> 98\%$. This sharing method is featured with other approaches, such as the use of rules and a lexicon.

Table 4.9 shows the obtained results after lemmatization. The verb "intercetou" in finite verb form, after lemmatization turns into "interceptar", in infinite verb form.

Tokens	A	GNR	intercetou	uma	embarcação	com	duas	toneladas	de	haxixe	em	o	Rio	Guadiana	.
POS Tags	art	prop	v-fin	art	n	prp	num	n	prp	n	prp	art	prop	prop	punc
Lemmas	O	gnr	intercetar	um	embarcação	com	dois	toneleda	de	haxixe	em	o	rio	guadiana	.

Table 4.9: Lemmatization Example.

The purpose of the lemmatization task is to normalize each token (mainly verbs) from the criminal-related documents into its dictionary form, or infinitive form. This way, texts became uniform and facilitates words association.

Dependency Parser

For dependency parsing, the parsing algorithms outline from sentences its grammatical structure and determine the relationships between the *HEAD* (root of the sentence) words and the others by aggregating tokens from a sentence into chunks. Chunks are groups of all tokens related to a feature (such as a noun). In a more straightforward approach, tokens have a neighborhood and their location in a sentence related to other tokens, such as subject or determinant, or their meaning similarity.

Figure 4.14 shows the dependency parser task. This task is based on Maltparser [16] system supported by a trained model with Linguateca's Bosque 8.0 corpus where each token is assigned to one of many grammatical various functions (see table 4.10).

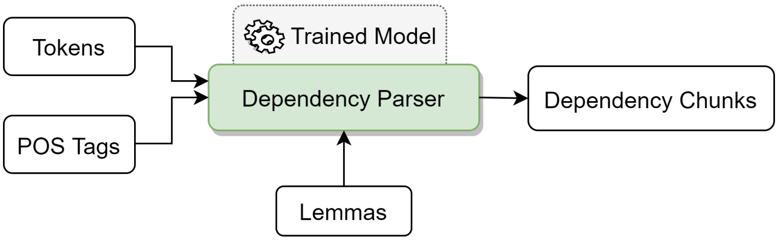


Figure 4.14: Dependency Parser Schema.

The dependency parser outputs dependency chunks. Each chunk is formed by tokens, constituted by head, as root, and corresponding dependent tokens. Take as an example, the following sentence:

"A GNR interceptou uma embarcação com duas toneladas de haxixe no Rio Guadiana."

Table 4.11 shows the inputs of dependency parser:

[16] See www.maltparser.org [Accessed: 1 March 2020].

Grammatical Functions	Abbrev.	Grammatical Functions	Abbrev.
(Predicate) Auxiliary Verb	PAUX	Vocative Adjunct	VOC
(Predicate) Main Verb	PMV	Topic Constituent	TOP
Adjunct Adverbial	ADVL	Top Node Noun Phrase	NPHR
Adjunct Predicative	PRED	Subject Related Argument Adverbial	ADVS
Auxiliary Verb	AUX	Subject Complement	SC
Complementizer Dependent	>S	Subject	SUBJ
Dative Object	DAT	Statement Predicative	S<
Direct Object	ACC	Root	ROOT
Focus Marker	FOC	Punctuation	PUNC
Main Verb	MV	Prepositional Object	PIV
Object Complement	OC	Predicator	P
Object Related Argument Adverbial	ADVO	Passive Adjunct	PASS

Table 4.10: Grammatical Functions.

Tokens	A	GNR	intercetou	uma	embarcação	com	duas	toneladas	de	haxixe	em	o	Rio	Guadiana	.
POS Tags	art	prop	v-fin	art	n	prp	num	n	prp	n	prp	art	prop	prop	punc
Lemmas	O	gnr	intercetar	um	embarcação	com	dois	toneleda	de	haxixe	em	o	rio	guadiana	.

Table 4.11: Dependency parser input.

Figure 4.15 shows the dependencies between tokens and the corresponding POS tags. To better visualize the dependency parser output, we have used an online tool named CoNLL-U Viewer [17]. The dependency tree is defined as a directed graph, with nodes (vertices) that correspond to a token (word) and arcs corresponding to its relations syntactic dependencies. These dependencies are labelled with dependency type (see table 4.10).

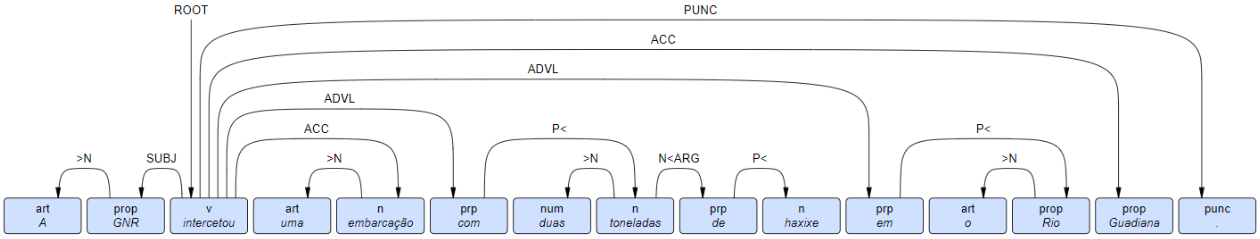


Figure 4.15: CoNLL-U Viewer Output.

The figure 4.15 shows the execution of the dependency parser over a sentence with root, the word "intercetou", and the words related to him. Regarding the study domain, this task is not different from other domains, which is an obvious conclusion because it does not deal with the meaning of words but with the syntax and dependencies. Indeed, our objective is to provide our framework with a parsing algorithm that could be used to check dependencies between tokens in each analyzed sentence.

[17] See www.let.rug.nl/kleiweg/conllu [Accessed: 1 October 2019].

4.3 Concluding Remarks

Along with this chapter, we start to answer the research question *"Is it possible to understand the unstructured data concerning the criminal domain by applying computational methods?"* since we propose a well-defined framework, named *SEMCrime*, with computational methods that allow us to understand and represent the unstructured data. Of course, the other chapters of this thesis will contribute to pursuing a complete answer to this question.

This chapter would have a significant impact on subsequent processing steps and in the criminal domain because:

- the introduction of the *SEMCrime* framework with the description of its modules;
- introduces the description of origin, structure, and content of criminal-related documents. This gave us an evaluation dataset to submit to our framework;
- the *Preprocessing Criminal-Related Documents* enables us:
  - to have the criminal-related documents into a semi-structured format, with the XML format, tagged with meta-information related to documents structure;
  - to adopt the abbreviation and acronyms lists with new terms obtained from criminal-related documents;
  - to understand lexical and syntactically the obtain text from the documents;
  - to have a new feature added to the tokenization task enables the elisions found in some compound words.

The *Preprocessing Criminal-Related Documents* module outputs tokens, POS Tags, dependency chunks, lemmas, and the criminal-related documents in XML format (this way, allowing to be used for different purposes). As long as the document are written in Portuguese, the results of this module are similar to the ones that we can get in other fields, such as the medical one. Despite, our documents having terms intrinsically linked to the criminal domain, that is not relevant in this phase.

5

Recognizing Named-Entities in Criminal-Related Documents

> *"Know how to solve every problem that has ever been solved".*
>
> *Richard Feynman*

This chapter proposes a unified approach for a NER module, named as NER-SEMCrime, applied to the criminal domain using the Portuguese language, that is capable of handling both dependent and independent domain named-entities, like persons, locations, organizations, narcotics, and crime types. A collection of combined classifiers was submitted to provide a unified annotated output.

The criminal-related documents besides named-entities like a person, location, and organization also have domain-specific entities (identified by their proper nouns) derived from unstructured data. These entities could be domain-dependent, such as narcotics or crime type names, or domain-independent, such as a person's name. For our framework, we need to identify and classify NEs relevant to the domain extracted from documents; therefore, we propose an automatic annotation of relevant NEs for the specific domain, such as narcotics, crime type names, mobile phone numbers, license plates, emails, person's names and among others.

To propose a NER module, we have established a two-fold approach that identifies the NEs in two main groups. First, table 5.1 lists the NEs that were identified and classified in the documents, which form a list of annotated entities.

	Domain-Independent	**Domain-Dependent**
	Persons, Organizations, Locations, Time/Date	Narcotics
Named-Entities	Numeric, License Plates, Emails Accounts	Crime Type
	ZIP Codes	Criminal Role

Table 5.1: Named-Entities Two-Fold Approach.

In the following items, we resume this chapter contribution:

- to make available NER classifiers using the Portuguese language annotated corpus (individually and combined) using different learning algorithms;
- to provide three domain-specific NER classifiers:
  - a narcotics NE classifier, by using their generic and street name using a supervised approach;
  - a crime type NE classifier, by using a crime type gazetteer;
  - a criminal role NE classifier, by using a criminal role gazetteer;
- to make available a NER classifier based on patterns for specific NE, such as mobile phones, license plates, and email accounts;
- to propose a NER module to be integrated into our framework.

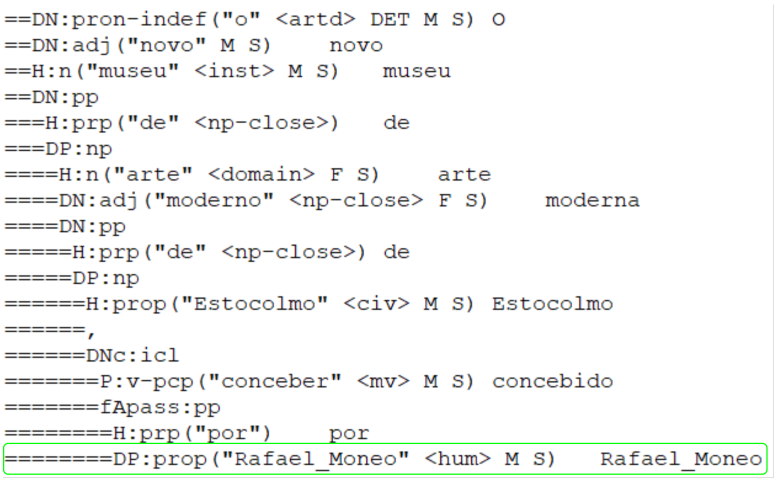
The remainder of this chapter is organized as follows: Section 5.1 aims to propose a methodology that combined trained *Floresta Sintática* corpora to improve the NER classifiers in order to enhance evaluation measures using as learning algorithms, the Perceptron, MaxEnt and Naïve Bayes, achieving an F1 measure of 0.815. Section 5.2 is focused on NER module proposal divided into sections. Firstly, in section 5.2.1, we proposed the domain-independent NER classifiers that extract relevant NE, identified by domain experts, like persons, organizations, locations, license plates numbers, emails, or postal codes, among others; Secondly, section 5.2.2 provides a domain-dependent NE that exploits new entities classification, like narcotics, crime type, and criminal role. Finally, section 5.3 describes an experimental setup performed over an evaluation dataset and the obtained results. The chapter ends with conclusions and results.

5.1 A Combined Corpus to Improve NER Classifier Performance

Our NER classifier combines unsupervised and supervised learning approaches. Regarding the supervised learning approaches, we use an annotated corpus, validated by the research community, and available for download. For training and testing purposes, we have identified two corpus: the *Floresta Sintática* [1] (described in Section 2.3.3) corpora and the *HAREM* [2] corpus (described in Section 2.3.5).

However, for our proposal, we have selected the *Floresta Sintática* corpora, particularly, the *Amazonia*, *Selva* and *Floresta Virgem* corpus. Our selection falls on this corpora because it allows the analysis of different corpus from different origins (in Portuguese and Brazilian), its volume of data, and the possibility of bringing them together in one corpus (as they follow the same annotation format).

The *Floresta Sintatica* corpora has been annotated by *PALAVRAS* (Bick, 2000) system by using an semantic tag, as can see in excerpt showed in figure 5.1. The annotated semantic tags [3] permits to label the NEs. In the boxed text of figure 5.1 we can see the tag *<hum>* (shortand for human) applied to the proper name "Rafael Moneon". This is the method used by *PALAVRAS* to annotated the semantic tags that identifies and classifies the NEs (the semantic tags have a corresponding HAREM [4] tags, available online in www.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags_nouns).



```
==DN:pron-indef("o" <artd> DET M S) O
==DN:adj("novo" M S)     novo
==H:n("museu" <inst> M S)    museu
==DN:pp
===H:prp("de" <np-close>)    de
===DP:np
====H:n("arte" <domain> F S)     arte
====DN:adj("moderno" <np-close> F S)     moderna
====DN:pp
=====H:prp("de" <np-close>) de
=====DP:np
======H:prop("Estocolmo" <civ> M S) Estocolmo
======,
======DNc:icl
=======P:v-pcp("conceber" <mv> M S) concebido
=======fApass:pp
========H:prp("por")     por
========DP:prop("Rafael_Moneo" <hum> M S)    Rafael_Moneo
```

Figure 5.1: Lying Trees (Excerpt).

Hence, we have joined the *Amazonia*, the *Selva* and the *Floresta Virgem* into one corpus, named *JointCorpusFS* corpus, for the following reasons:

- increasing of data volume for training purposes;
- variety of vocabulary (Portuguese and Brazilian-Portuguese variations) that enrich the content.

Such corpus merge was possible due to the fact that they are all:

- in Portuguese (or variant, Brazilian-Portuguese) language;
- in the same format, namely the AD (Afonso, 2004) (see figure 5.1);
- freely available for download.

[1] See www.linguateca.pt/Floresta/corpus.html [Accessed: 1 May 2019].
[2] See www.linguateca.pt/HAREM [Accessed: 1 May 2019]
[3] See www.visl.sdu.dk/visl/pt/info [Accessed: 1 December 2019].
[4] See www.linguateca.pt/acesso/corpus.php?corpus=CDHAREM [Accessed: 1 December 2019].

For NER training and testing purposes, three different learning algorithms were used, named Naïve Bayes, Perceptron and MaxEnt (described in Section 2.5), along with a file for their configuration.

For training and evaluation, we have chosen the *Apache OpenNLP* [5], because:

- it is a versatile resource with several NLP tasks, although it does not offer a NER classifier trained in the Portuguese language. It offers a *NameFinder* library that is capable of training and testing models in any language. The tool includes a file for features configuration in XML format, thus allowing to choose the features to consider to generate the classifiers (Fonseca et al., 2015a), such as algorithms types;
- has tools to convert corpus formats between them, which was useful for our proposal because we convert the AD into *Apache OpenNLP* format;
- offers a set of configurable learning algorithms allowing comparison between these.

Figure 5.2 shows a diagram with our methodology supported by the selected corpus, learning algorithms, and training and testing tasks. In (A), we listed the corpus segmentation for classifier training and evaluation.

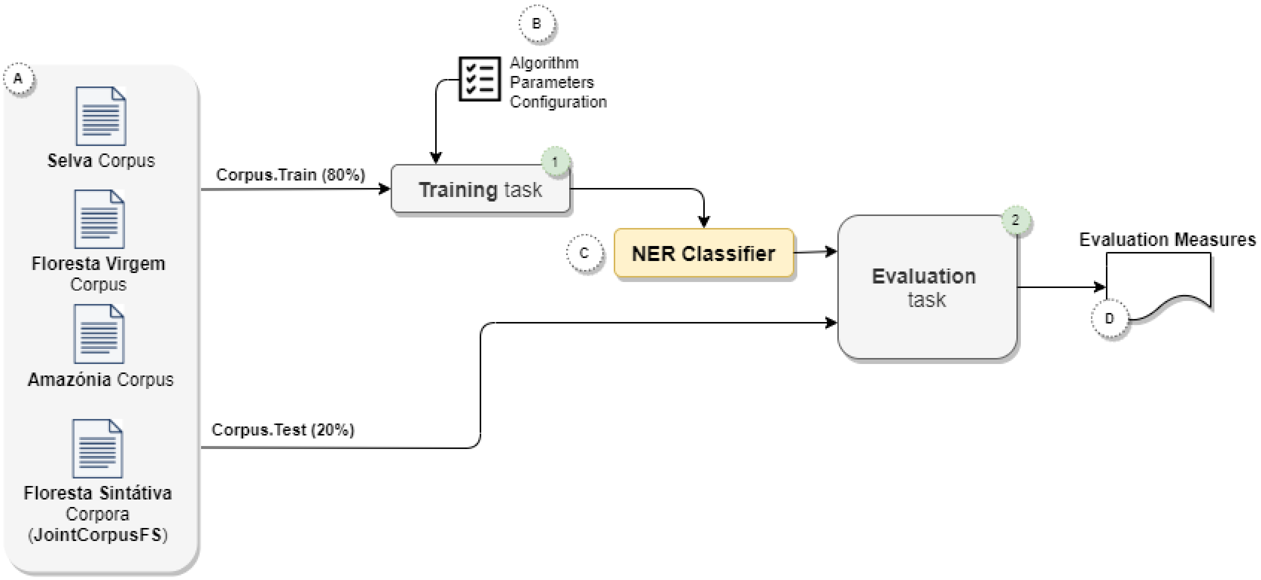


Figure 5.2: NER Training and Evaluation Methodology.

All corpus were divided into two segments: the *Corpus.Train* (80%) and the *Corpus.Test* (20%). The training and evaluation process used the following steps:

- In (1), a task for classifier training that is configured by a (B) Algorithm Parameters Configuration file with setting parameters: Algorithm name (Naïve Bayes, Perceptron, and MaxEnt), Interactions, and Cutoff. Training task result is a (C) NER classifier;
- In (2), we perform the evaluation task by using as input the *Corpus.Test* (20%) corpus.

[5] See www.opennlp.apache.org [Accessed: 1 July 2020].

5.1.1 Obtained Results

Table 5.2 shows the results obtained for *Precision (P)*, *Recall (R)*, and *F-Measure (F1)* for the best-trained model, with the selected corpus (individually or combined) and a learning algorithm with setup parameters.

	Perceptron			**MaxEnt**			**Naïve Bayes**		
NER Classifiers	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Amazonia	0.821	0.794	0.807	0.814	0.732	0.771	0.568	0.706	0.629
Selva	0.803	0.713	0.755	0.832	0.627	0.715	0.623	0.677	0.649
Floresta Virgem	0.836	0.773	0.804	0.810	0.721	0.763	0.602	0.723	0.657
JointCorpusFS	**0.841**	**0.789**	**0.815**	0.832	0.761	0.796	0.589	0.719	0.648

Table 5.2: NER Classifiers Evaluation.

The obtained results show that Naïve Bayes has the lowest score in all the measurements. In contrast, the Perceptron algorithm reaches the best result between the generated classifiers. The *Precision* measure reaches a 0.841 score, which tells us our classifier's ability to reject any non-relevant documents in the retrieved set. The *Recall* measure achieves a 0.789 score, the classifier ability to find all the relevant documents, where the proportion of positive cases is correct. Finally, the *F-Measure*, achieves 0.815 score, which is trade-off between *Precision* and *Recall*.

Table 5.3 summarizes the evaluation results for our best score classifier in terms of each target entity type.

	Perceptron			**MaxEnt**			**Naïve Bayes**		
Entity Type	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
numeric	0.942	0.899	**0.919**	0.924	0.874	0.899	0.499	0.687	0.578
time	0.956	0.862	**0.917**	0.903	0.913	0.908	0.721	0.899	0.790
event	0.924	0.846	**0.883**	0.919	0.802	0.852	0.627	0.879	0.732
location	0.851	0.813	0.832	0.824	0.834	**0.839**	0.635	0.772	0.697
person	0.828	0.767	**0.796**	0.812	0.767	0.789	0.638	0.714	0.674
organization	0.811	0.762	**0.786**	0.788	0.722	0.753	0.599	0.664	0.620
artprod	0.709	0.679	**0.694**	0.739	0.502	0.598	0.283	0.561	0.376
thing	0.753	0.609	**0.674**	0.784	0.509	0.618	0.652	0.516	0.576
abstract	0.605	0.600	**0.603**	0.745	0.494	0.594	0.539	0.433	0.480

Table 5.3: Best-Trained Classifiers Evaluation (by each target entity type).

Figure 5.3 shows the comparison between F1 results for each target entity type, for better visualization.

Evaluation measures were identified by highlighting where the *F-Measure* has had the best result, with the best trade-off for each entity type. The best evaluation results occurred with the Perceptron algorithm; however, MaxEnt achieved a better result regarding the Location entity type.

In chapter 4, we have used the *NLPPort* system (Rodrigues et al., 2018) to support our framework with

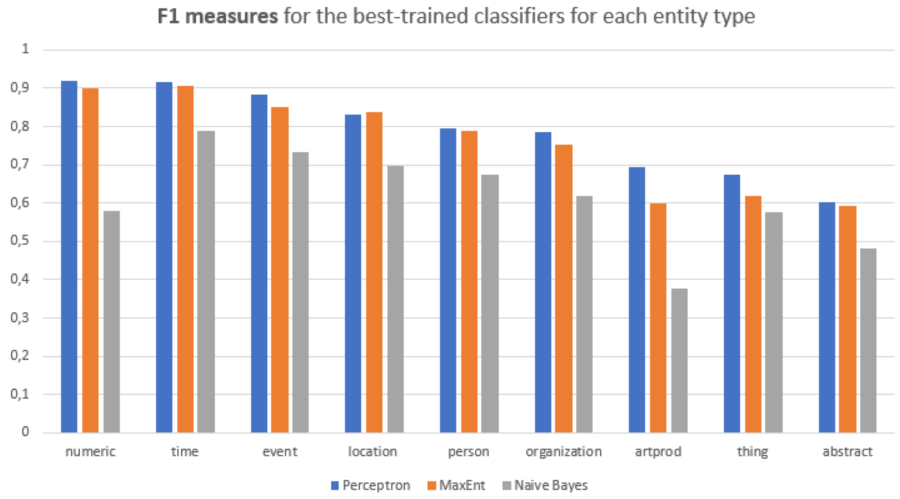


Figure 5.3: F-Measure Results for Best-Trained Classifiers (by NE Type).

some teaks. However, concerning to the NER module, named as *EntPORT*, the authors delivered a NER classifier that was trained with the *Bosque 8.0* treebank (integrated in *Floresta Virgem)*, and also, by using the *Apache OpenNLP* toolkit.

	Precision	**Recall**	**F-Measure (F1)**
EntPORT NER Classifier	0.778	0.699	0.736
Our NER Classifier	**0.841**	**0.789**	**0.815**

Table 5.4: Comparing NER Classifier Performance Results

The table 5.4 shows the evaluation results regarding our best-trained classifier and *EntPORT* NER Classifier. As we observe, the results obtained by our classifier are quite better in all performance metrics.

5.2 NER-SEMCrime Module Proposal

In this section, we propose a NER module, named **NER-SEMCrime**, for the Portuguese language applied to the criminal domain. We add new classifiers with specific training. Some difficulties were found during analysis/training, namely:

- there is a vague definition of the criminal domain in terms of NEs, due to different contexts, such as the expression: "She is so heroine";
- we can spot terms that are domain-dependent; however, terms used in other domains are used in this context, such as vehicle brands;
- the lack of NER tools and trained classifiers applied to the Portuguese language;
- the lack of freely available corpus for the criminal domain, with annotated criminal-related entities.

For those reasons, we have developed efforts to adapt capabilities from other fields on the task at hand. Therefore, to enable entities to identify and classify (and consider the restrictions already mentioned), we have studied the following learning methods: pattern, gazetteer-based, and supervised learning methods. Thus, having an approach that integrates NEs mentions from domain-dependent/independent to resolve a NER problem and adds value to the domain. We move a step forward to enable the criminal-related entities identification and classification and other entities relevant to the domain; we have followed three different learning methods to implement our proposal:

- gazetteer-based: using dictionaries with terms related to the criminal domain that needs to be detected in the documents;
- patterns rule: for example, regular expressions that enable the identification of patterns in text portions;
- supervised learning: using manually annotated corpus and learning algorithms to train classifiers to identify and classify specific NE.

The figure 5.4 shows the **NER-SEMCrime** module that combines the following processing chain: sentence detection and tokenization tasks to perform a preprocessing task; followed by a parallel-group of NER classifiers; and finally, the *ENSEMLE Method Named-Entities* that joins the *NE Pairs* from each classifier.

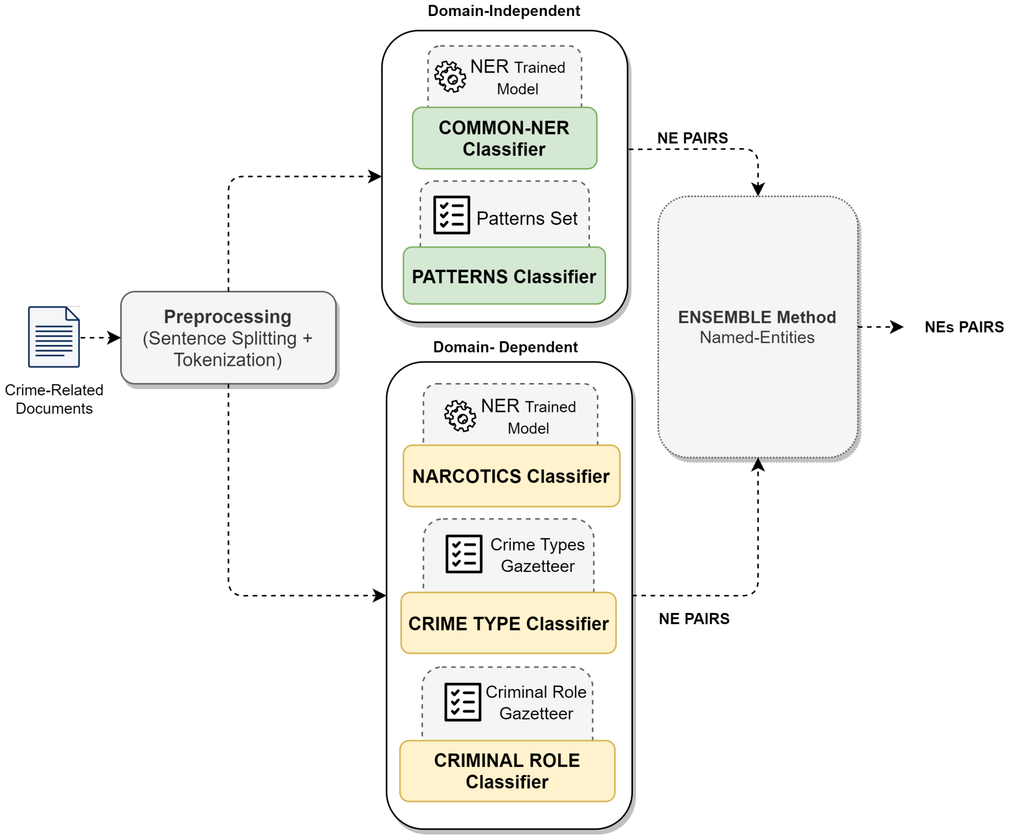


Figure 5.4: *NER-SEMCrime* Module Architecture.

5.2.1 NER Classifiers for Domain-Independent Named-Entities

We propose two NER classifiers for domain-independent NEs divided into:

- the mentions to person names (i.e., criminals, victims, or police officers), organizations (i.e., police agencies, financial institutions, etc.), locations names, along with other relevant entities. For this, we have established a NER classifier, named as *COMMON-NER*;
- a pattern set that uses regular expressions to extract entities, such as mobile phone numbers, license plates, emails, or others. Therefore, we establish a NER classifier, named as *PATTERNS*.

The following two sections describe each one of those NER classifiers.

COMMON-NER Classifier

The *COMMON-NER Classifier* proposed to classify the following NER types: numeric, organization, person, location, and time NEs. To enable this, we have used a trained model from section 5.1 that obtained the best performance results, which is the *JointCorpusFS*.

The sentence below illustrates annotation with the *COMMON-NER* classifier:

> *"O <person> Luis Silva </person> foi identificado como autor do assalto ao <organization> Banco de Portugal </organization> pelas <time> 14 horas </time>, na dependência de <location> Coimbra </location>."*

PATTERNS Classifier

The *PATTERNS Classifier* is proposed to detect the following entities - mobile phone numbers, email addresses, license plates, and zip codes. This approach relies on a set of patterns in a configurable file (which enable us to add more entities and patterns), defined by the following steps:

- identifying the entity to extract that follows a specific pattern, such as license plates;
- defining the regular expression (pattern) for each entity;
- adding to the file the defined pattern rule with the following format: *<entityname>* ***pattern rule*** *</entityname>*.

After defining the list, we implement a method that enables the lookup of entities mentioned in criminal-related documents and enables the extraction of *NE Pairs*. Listing 5.1 shows the configuration file in XML format.

```
?xml version="1.0" encoding="UTF-8"?>
<?process <"'&> ?>
<patternrulesentityfinder>
    ---- MORE
    <licenseplates>((?:[A-Z]{2}-\d{2}-\d{2})|(?:\d{2}-[A-Z]{2}-\d{2})|(?:\d{2}-\d
       {2}-[A-Z]{2}))</licenseplates>
```

```
   <mobilephone>(9[1236]\d) ?(\d{3}) ?(\d{3})</mobilephone>
   <zipcode>[0-9]{4}-[0-9]{3}</zipcode>
 ---- MORE
</patternrulesentityfinder>
```

Listing 5.1: PATTERNS Classifier configuration file.

The sentence below illustrates the annotations with the *PATTERNS Classifier*:

> *"O veiculo com a matricula <licenseplates> XX-XX-AA </licenseplates>, pertence ao suspeito com o telefone <mobilephone> 998765234 </mobilephone>, email: <email> sapo@sad.pt </e-mail>, e com o código postal <zipcode> 3210-554 </zipcode>."*.

This approach allow us to add a new pattern rule into the configuration file and thus a new entity classification. Therefore, the classifier is extendable.

5.2.2 NER Classifiers for Domain-Dependent Named-Entities

In the criminal domain, we have entities that populate the criminal-related documents, and their identification and classification as NEs, will benefit the information representation regarding a particular crime, such as cocaine or heroin, classified as narcotics.

NER Classifier for Narcotics Names

This section presents a new entity classification applied to the criminal domain using the Portuguese language, named by narcotics. From the analysis of criminal-related documents, we noted:

- the narcotics names are mentioned in their current and street name across criminal-related documents;
- the drug trafficking is one of the most reported [6] and typified crimes, investigated by the criminal police.

In the following sentence, narcotics term candidate is highlighted by its current name:

> *"O arguido foi detido por tráfico de cocaína e heroína ."*
>
> In English: "the defendant was arrested for cocaine and heroin trafficking.".

Also, the narcotics term candidate is highlighted by its street name, such as:

> *"O arguido foi detido por tráfico de neve e cavalo ."*
>
> In English: "the defendant was arrested for snow and horse trafficking.".

[6] See www.pordata.pt/Europa/Crimes+por+categoria-3285 [Accessed: April 2020].

We have used a supervised learning approach and a manually annotated corpus for training and testing to implement this classifier. Therefore, we have extracted texts from daily newspapers and blogs that mention narcotics terms. Because of the lack of annotated texts in the Portuguese language related to the study domain, a manual annotation process was performed, by annotating the correct entity using a narcotics list, with current and street names. The sentence below illustrates how the documents have been annotated.

> *Foram ainda apreendidas 5500 doses de <narcotics> liamba </narcotics>, 323 plantas de <narcotics> canábis </narcotics>, 16 doses de <narcotics> haxixe </narcotics> e 12 doses de <narcotics> MDMA </narcotics> (mais conhecido por <narcotics> ecstasy </narcotics>) e 930 euros.*

We have divided the corpus into *narcotics.train* (80%) and *narcotics.test* (20%). For classifier generation, we have used the *Apache OpenNLP* tool (for training and evaluation) with a Java wrapper that enables us to annotated criminal-related documents, and outputs the *NE Pairs* as output. Table 5.5 shows the training outputs.

Sentences	3453
Tokens	85975
Narcotics entities	1051

Table 5.5: OpenNLP Training Results.

For training purposes, we have followed the same path of the other classifiers, such as the *COMMON Classifier*, by training them with three different learning algorithms: the MaxEnt, Perceptron, and Naïve Bayes. Table 5.6 shows the performance measures regarding the trained and evaluations for the narcotics classifier by using the MaxEnt, Perceptron, and Naïve Bayes learning algorithms:

	MaxEnt			Perceptron			Naïve Bayes		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Narcotics Classifier	0.758	0.383	0.519	0.770	0.531	**0.634**	0.494	0.821	0.611

Table 5.6: Narcotics Classifier performance measures.

From the analysis of the evaluation measures, we realized that the Perceptron algorithm was the best classifier using the F-Measure to compare algorithms: *Precision* is equal to **77%**, in a sample of 100 entities annotated text; *Recall* measures revealed that our system decreased in percentage. Thus obtaining **53%**, when missed positives tags were taken into account and mistakenly replaced with negative ones (the false negatives), introducing noise in the extracted data; *F-Measure* that is a trade-off between *Precision* an *Recall* which is around **63%**, almost the same at around 56%, between *Precision* and *Recall*. The results obtained were revealed to be quite promising.

NER Classifier for Crime Types

We have identified words or compound words during a manual analysis of criminal-related documents, such as drug dealing, assault, or homicide, to be candidate entities for crime type name. Also, in Portugal, the *Law 59/2007 of 4 of September in the Portuguese Penal System* [7], enumerates the crimes, articles,

[7] See www.dre.pt/home/-/dre/640142/details/maximized [Accessed: December 2019].

and the corresponding penalization. This law allows the construction of a list of entities that enable the identification and classification of crime types.

Therefore, from this law, we have created a gazetteer of crime types with around 252 names used on our NER classifier for training purposes. The listing 5.2 shows an excerpt of the crime types gazetteer.

```
corrupção activa
peculato
peculato de uso
participação económica em negócio
violação de domicílio por funcionário
concussão
recusa de cooperação
abuso de poder
violação de segredo por funcionário
abandono de funções
assassínio
terrorismo
sequestro
falsificação de documento
```

Listing 5.2: Types of Crimes Gazetteer (Excerpt).

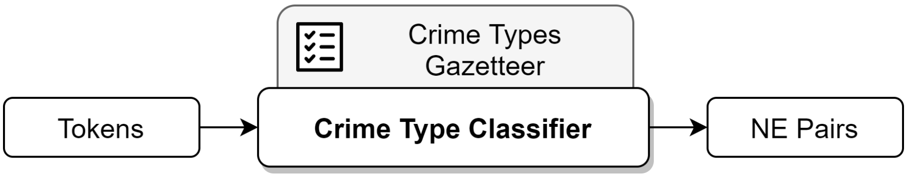


Figure 5.5: Crime Type NER Classifier Scheme.

Figure 5.5 shows the proposal for our a gazetteer-based approach that performs a matching that detects tokens or sequences of tokens that match a fixed pattern to identify a crime type name, by a lookup into a gazetteer using regular expressions. We established the following steps:

- create a pattern from each term in the gazetteer;
- the pattern is compared with each sentence (in lower case) to find a positive match;
- remove overlapping entities;
- the entities are annotated.

For technical purposes, we have used the *OpenNLP* toolkit; more specifically, the class *RegexNameFinder* to extract specific patterns based on the gazetteer terms. The sentence below illustrates the annotations with the *Crime Type* NER classifier:

O suspeito foi detido em Coimbra pelas 14 horas pelo <crimetype> homicídio </crimetype> de Luis Silva.

NER Classifier for Criminal Role

During the manual analysis of documents, we found that sometimes persons and organizations were identified by their roles. For example, instead of using a person's name, we have the term "suspect" to identify it. Since extracting information in these situations can be particularly challenging, we resorted to the classifier proposed in this section, to help on this issue.

Therefore, after identifying this need, we have established that Criminal Role entities formally identify and classify a person or organization in the criminal context. For example, the sentence *"the suspect was arrested by the Police."* allows identifying "suspect" as a person and an associated role, and "police" as an organization and associated role. Emphasizing the omission of the person's name is recurrent in the analyzed documents, either for anonymity or because the person or organization's name is unknown at the investigation time. For example, the listing 5.3 shows an excerpt of the criminal role gazetteer.

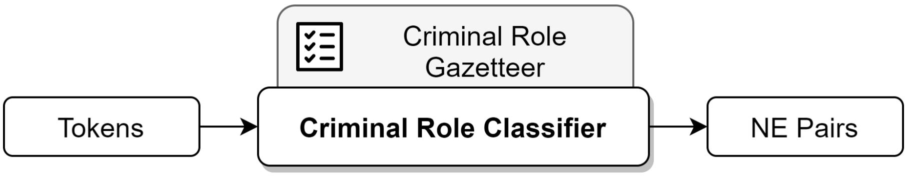


Figure 5.6: Criminal Role NER Classifier Scheme.

Figure 5.6 shows the proposal for our a gazetteer-based approach that performs a matching that detects tokens or sequences of tokens that match a fixed pattern to identify a role, by a lookup into a gazetteer using regular expressions. We established the following steps:

- create a pattern from each term in the gazetteer;
- the pattern is compared with each sentence (in lower case) to find a positive match;
- remove overlapping entities;
- the entities are annotated.

```
coautor
autor
suspeito
homicida
inspector
assassino
traficante
acusado
delinquente
infrator
```

Listing 5.3: Criminal Role Gazetteer (excerpt).

A prototype was implemented to identify and classify the *Criminal Role* entity by using a Criminal Role Gazetteer. For technical purposes, we have used the *OpenNLP* toolkit; more specifically, the class *RegexNameFinder* used to extract specific patterns based on the gazetteer terms. The sentence below illustrates the annotations with the *Criminal Role* NER classifier:

> *O <criminalrole> suspeito </criminalrole> foi detido em Coimbra pelas 14 horas. como <criminalrole> coautor </criminalrole> do homicídio.*

5.2.3 Ensemble Method for Classifier Outputs

From previous sections, the **NER-SEMCrime** distinguishes different NEs by using different NER classifiers. Thus, we propose a method that combines the different NER classifiers to permit a unified representation. This reduces the probability of the same annotation appearing with similar or different NE types. To enable the fusion, we have considered that:

- if two or more NE have the same entity type, the method selects the NE from the core classifier. For example, if two classifiers identify an entity type as a person, we select the classifier trained for that purpose, in this case, the *COMMON-NER* classifier.

The sentence below illustrates an document excerpt with NEs annotated after processed by the ensemble method:

> *Após investigação da <organization> Polícia Judiciaria </organization>, o suspeito <person> Luis Silva </person> foi indiciado pelos crimes de <crimetype> roubo </crimetype>. Os crimes foram cometidos em <place> Coimbra </place>, durante <date> Setembro </date>, com auxilio do veiculo de matricula <licenseplates> XX-X0-X1 </licenseplates>. O <criminalrole> suspeito </criminalrole> era consumidor de <narcotics> cocaina </narcotics>.*

5.3 Experimental Setup

We conducted several experiments over the criminal-related documents dataset to validate the proposed classifiers, supported by a prototype developed in Java programming language that enables the classifiers annotation and outputs the NE pairs. We manually annotated the documents with the help of external annotators, who identified in each sentence the NE and entity types in each sentence. For example, in listing 5.4, we have a sample of an annotated sentence retrieved from a PGdLNews, where the NE is mentioned between <NE> and </NE> tags.

```
<sent1 desc=``Ao abrigo do disposto no art. 86.º, n.º 13, al. b) do Código de
    Processo Penal, a Procuradoria-Geral Distrital de Lisboa torna público que foi
    detido e presente ao Juiz de Instrução Criminal para primeiro interrogatório
    judicial, um arguido, fortemente indiciado pela prática de um crime de tráfico
    de estupefacientes.''>
        <NE>
                <organization>Procuradoria-Geral Distrital de Lisboa</organization>
                <crimetype>tráfico de estupefacientes</crimetype>
                <criminalrole>arguido</criminalrole>
                <criminalrole>Juiz de Instrução Criminal</criminalrole>
        </NE>
</sent1>
```

Listing 5.4: PGdLNews NEs Annotated (extract).

Table 5.7 describes the annotated documents, with a total of 163 documents, 1528 sentences, and 53583 tokens. Table 5.8 shows the matrix with the results obtained for the instances matched with relevant/non-relevant and retrieved/non-retrieved classes for the annotated documents.

	# Documents	Sentences	Tokens
Criminal News	80	718	22745
PGdLisboa News	80	547	20479
Criminal Investigation Reports	3	263	10359

Table 5.7: Annotated Criminal-Related Documents.

	Criminal News		**PGdLNews**		**Criminal Investigation Reports**	
	Relevant	**Not Relevant**	**Relevant**	**Not Relevant**	**Relevant**	**Not Relevant**
Retrieved	1048	166	819	140	514	188
Not Retrieved	175	21349	128	19382	106	9563

Table 5.8: Criminal-Related Documents Matrix.

From the results described in table 5.8, we measure *Precision*, *Recall*, *F-Measure* and *Accuracy* for each annotated group of documents. Therefore, table 5.9 enumerates the evaluation measures related to NER task.

	Precision	Recall	F-Measure (F1)
Criminal News	0.846	0.659	0.712
PGdLisboa News	0.850	0.679	0.716
Criminal Investigation Reports	0.728	0.829	0.771
Avg.	**0.808**	**0.722**	**0.733**

Table 5.9: Criminal-Related Documents Evaluation.

The experimental results demonstrate its effectiveness and have a good performance, thus, validating the **NER-SEMCrime** module for being integrated into the framework. We pointed out that the *Recall* result for the Criminal Investigation Reports is quite above the other results.

5.4 Concluding Remarks

This chapter proposes a unified use of several NER classifiers with different objectives, answering the research question enumerated in Chapter 1, *"Can we identify relevant entities related to the criminal domain?"*.

We stated that adding this module to our framework allows for identifying and classifying entities related to the criminal domain or other relevant entities. Along with this chapter, we improved NER models, and proposed new NEs, like narcotics or crime types, divided them into dependent and independent NEs to obtain a module applied to the Criminal domain in the Portuguese language. The reached outcomes were:

- three novel NER classifiers, namely narcotics, crime types, and criminal roles;
- an increase of classifier performance, namely the *COMMON Classifier* that identifies and classifies the common NE, such as persons, locations, or organizations;
- proposed a NER classifier for entities that followed a certain pattern, like license plates, using a configuration file that enabled the addition of new patterns;
- an ensemble method to combine NER outputs of a unified annotation of a certain corpus, reducing this way, the issues related to annotation of the same token;

Experimental results demonstrate the approach's effectiveness, allowing us to have an optimistic perspective for applying the module in our framework. Of course, we can improve the obtained results for the proposed models, like transforming the models based on gazetteers into trained models with the aid of learning algorithms, decreasing the issues related to gazetteers. By domain analysis, the necessary entities are widespread, given that it depends on the context. For example, if a given crime has boats, airplanes, and tax havens as resources, we need entities that classify these resources. In other words, this module needs to be fitted to the context where it is applied.

6

Criminal-Related Data Representation in a Graph Database

"Machines take me by surprise with great frequency."

Alan Turing

This chapter illustrates the final steps to represent the data retrieved from criminal-related documents into a Neo4j graph database. We proposed a module, named as Neo4j Criminal-Related Documents Representation, divided into Criminal Information Extraction, which enables the identification and classification of NEs, criminal terms, and semantic roles. 5W1H Information Extraction Method identifies and classifies the 5W's questions based on the 5W1H information, the crime type, and criminal terms. Finally, we need to populate and enrich the Neo4j graph database using a module named as Graph Database Population and Enrichment. An experimental setup was performed by using criminal-related documents and an application example.

The criminal-related documents followed syntactic rules and linguistic conventions similar to those we can find in academic, news, and business contexts. To understand each text, we need to identify the semantic relation between words that are useful for several applications, such as information retrieval, construction, and the extension of lexical resources (Specia and Motta, 2006). We opted for a *5W1H* method to answer the questions What?, Who?, Where?, When?, Why?, and How? (nevertheless, the last one was discarded from this proposal). The figure 6.1 shows the ***Neo4j Criminal-Related Documents Representation***.

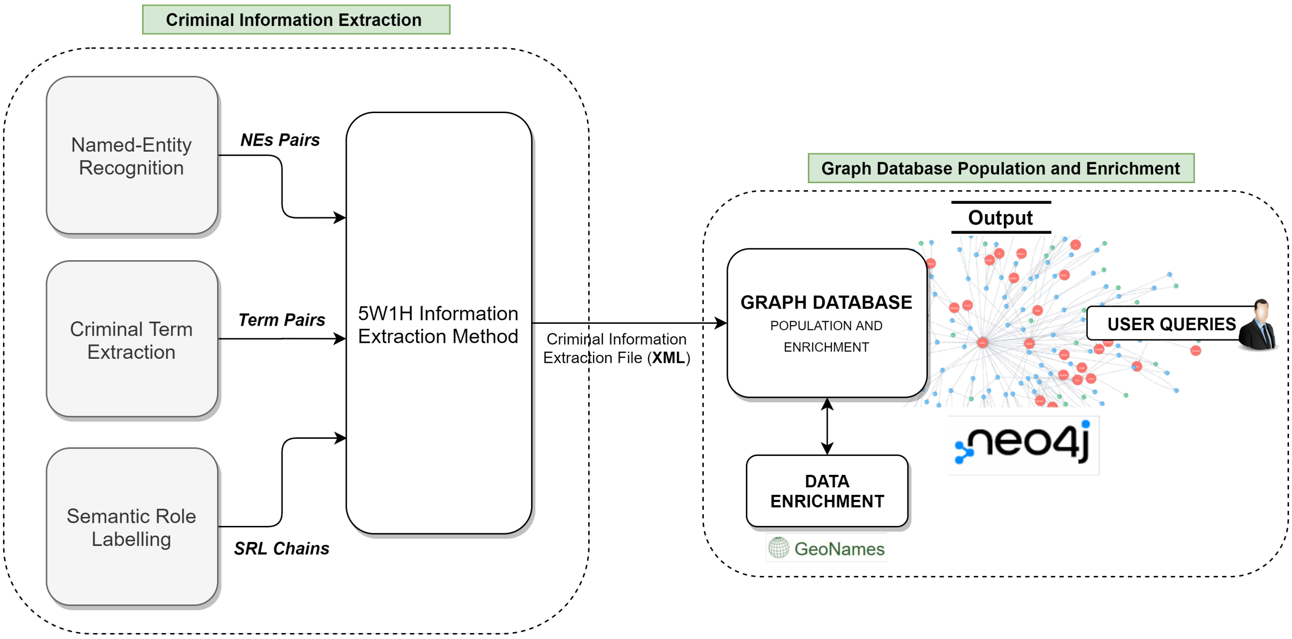

Criminal Information Extraction
Named-Entity Recognition
Criminal Term Extraction
Semantic Role Labelling
NEs Pairs
Term Pairs
SRL Chains
5W1H Information Extraction Method
Criminal Information Extraction File (XML)
Graph Database Population and Enrichment
Output
GRAPH DATABASE
POPULATION AND ENRICHMENT
USER QUERIES
neo4j
DATA ENRICHMENT
GeoNames

Figure 6.1: Neo4j Criminal-Related Documents Representation.

The module is divided into:

- the *Criminal Information Extraction*:
  - *Named-Entity Recognition*: this module identifies and classifies NE, that are relevant to the domain. As output, *NE Pairs* are generated (for further information, see Chapter 5);
  - *Criminal Term Extraction*: identifies and classifies domain-specific terms related to domain that were missed in extraction by other modules and are relevant to be extracted;
  - *Semantic Role Labelling*: assigns semantic roles to sentence constituents;
  - *5W1H Information Extraction Method*: extracts events and related elements in order to answer the *5Ws* (Who, What, When, Where, and Why) information;
- the *Graph Database Population and Enrichment*:
  - *Graph Database Population and Enrichment*: enables the population and enrichment of a *Neo4j* [1] graph database. This module uses a Data Enrichment task to enrich the location NE with GeoNames [2] geographical database information;

[1] See neo4j.com [Accessed: 1 April 2020].
[2] See www.geonames.org [Accessed: 1 May 2020].

The remainder of this chapter is organized as follows: section 6.1 introduces a novel module to extract terms related to the criminal domain, named as criminal terms; section 6.2 explains the use and adaptation of the SRL to identify and classify the semantic roles; section 6.3 proposes a novel *5W1H Information Extraction Method* applied to criminal-related documents; section 6.4 introduces the graph database population and enrichment by enumerating the modeling decisions, and population and enrichment method; the experimental setup was proposed in section 6.4.3, and finally, in section 6.5, we add an application example using a criminal investigation report evaluated by our proposal and by a domain expert.

6.1 Criminal Term Extraction Module

The introduction of the *Criminal Term Extraction* module permits the extraction of domain-specific terms that are common in criminal-related documents. This necessity is a consequence of the manual analysis of documents and domain expert interviews, where terms were identified as part of the criminal context, such as "buscas domiciliárias" (in English: "home searches"). These terms were named as *Criminal Terms*. Without this module, domain-specific terms would be lost during information extraction.

To resolve this challenge, have collected a list of terms related to the criminal domain by executing the following steps:

- to analyze the criminal-related documents by using the Linguakit [3] Keyword Extraction to identify a list of candidate terms. The terms frequency reveals their importance in the domain;
- to query the IATE [4] terminological database for terms related to Portuguese criminal law and to retrieve a list of candidate terms;
- to validate the candidate terms, by a domain expert.

After the analysis, a gazetteer was created with a list of around 400 criminal terms. Listing 6.1 shows an excerpt of the gazetteer.

```
causa provável
segurança nacional
investigação criminal
inquérito judiciário
ato terrorista
ato de terrorismo
arrombamento
escalamento
crime continuado
procedimento penal
entrega temporária
```

Listing 6.1: Criminal Terms Gazetteer (excerpt).

We followed an approach that uses regular expressions combined with the criminal terms gazetteer to implement our module. The module uses the *Criminal Terms Gazetteer*, and a *Pattern Rules* to create a pattern that seeks the term in the input sentence by using the term contained in the gazetteer, and outputs the *Term Pairs*, such as <**criminalterm**> buscas domiciliárias </**criminalterm**> (term pairs

[3] See linguakit.com [Accessed: 1 May 2020].
[4] See iate.europa.eu [Accessed: 1 July 2020].

means the combination of the criminal label term and the extracted criminal term). For development, we used the *OpenNLP* toolkit; more specifically, the class *RegexNameFinder* was used to extract specific patterns. Figure 6.2 shows an representation of our approach.

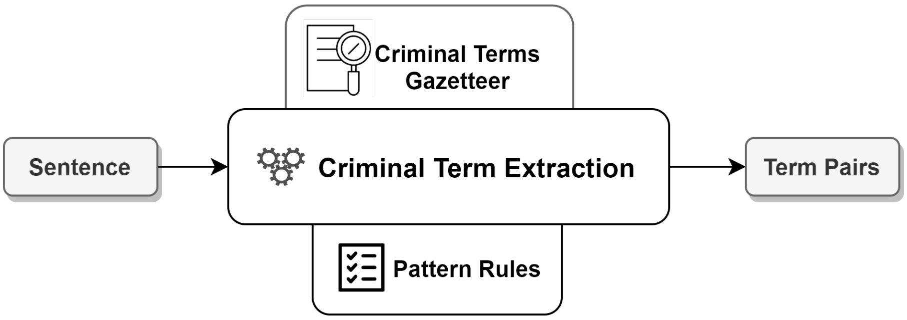


Figure 6.2: Criminal Term Extraction Scheme.

6.2 Semantic Role Labelling Module

The introduction of a semantic role labeling task in our *Criminal Information Extraction Module* is to identify and classify the verb-argument structure of each sentence in criminal-related documents. Using this task, we can understand sentences semantically. For example, given a sentence *"O Rui Silva assaltou o Banco de Portugal, pelas 14 horas"*, the output goal is:

O Rui Silva $_{\textbf{Arg0}}$ asssaltou $_{\textbf{V}}$ o Banco de Portugal $_{\textbf{Arg1}}$, pelas 14 horas $_{\textbf{ArgM-TMP}}$.

The semantic roles annotated in the sentence above, the subject [O Rui Silva $_{\textbf{Arg0}}$], the verb/predicate [assaltou $_{\textbf{V}}$], the object [o Banco de Portugal $_{\textbf{Arg1}}$], and finally, the temporal marker [pelas 14 horas $_{\textbf{ArgM-TMP}}$].

However, the development of a SRL tool from the ground up is out of the scope of this work. Nevertheless, from the analysis of Chapter 3, we can derive some restrictions that could help us fit any of the described systems as part of our framework, such as source code available, initial performance measures results that must be relevant, and focused on the Portuguese language. From the constraints enumerated before, we selected the *NLPNET* (Fonseca and Rosa, 2013) to support our SRL task.

6.2.1 NLPNET system

The *NLPNET* system has been chosen to support our SRL task. This tool is freely available at *NLPNET* GitHub repository (`www.github.com/erickrf/nlpnet`), as a Python library for NLP tasks based on neural networks. The system was created by (Silla and Kaestner, 2004) (Fonseca and Rosa, 2013) to enable an SRL task for Brazilian Portuguese, inspired by the SENNA system, with the intent of also pursuing the goal of reducing the computational cost of the algorithms employed. For training purposes, the text corpus used for training and evaluation was built from Brazilian newspapers, the PropBank-Br (Duran and Aluísio, 2012). *NLPNET* process is divided into three steps, namely the predicate identification, the argument identification and classification. Table 6.1 shows the results obtained by the authors.

The table 6.1 describes the identification and classification sub-tasks. Despite the achieved performance measures, the tool has room to improve the performance measures. Therefore, it is an exciting path to be

Description	Precision (P)	Recall (R)	F-Measure (F1)	Accuracy
Identification	0.764	0.724	0.744	—
Classification	—	—	—	0.875

Table 6.1: *NLPNET* System Evaluation Measures (Fonseca and Rosa, 2013).

followed - a new SRL model with better results.

NLPNET Output Adaptation

The *NLPNET* system offers a standalone script to execute the proposed models for a given text input, however, the output needs some teaks before the integration on our framework. Namely, the output does not follow an output format easy to analyze. Therefore, we have defined the following output: P identifies the predicate; $A_0...n$ defines the semantic role arguments (see table 2.6); and V is the verb of the sentence. Listing 6.2 shows the output template:

```
<P>(...)</P><A0...n>(...)</A0...n><V>(...)</V><A0...n>(...)</A0...n>
```

Listing 6.2: SRL Output Format.

A method was developed to accomplish the following requirements: read all criminal-related documents and processed sentence by sentence; produce an output file that follows the established format. Listing 6.3 shows the output with the defined format, that demonstrates each input sentence in criminal-related documents as a sequence of semantic roles between sentence tag (<sentence> (...) </sentence>), thus, forming the **SRL Chains**.

```
<sentence>
<P>assaltou</P> <A0>O Rui Silva</A0> <V>assaltou</V> <A1>o Banco de Portugal</A1><
    ArgM-TMP>pelas 14 horas</ArgM-TMP>
</sentence>
```

Listing 6.3: SRL Output (Sentence Example).

6.3 5W1H Information Extraction Method

The 5W1H information extraction method relies on the answers to the *5W's* questions presented in sentences of the criminal-related documents. Two other extractions were introduced related to the criminal domain: the crime type and the criminal terms. Our approach is inspired by the works of (Wang, 2012b) that uses SRL, NER approaches to represent Chinese News into an ontology, using a list of trigger words for events filtering. In (McCracken et al., 2006) the use of a SRL and a rule-based approach was used to create an event model to understand the summary reports, and then to map results into a domain ontology; Zhou et al. (2016) uses an SRL combining set of heuristic rules to extract events in Chinese texts. However, our work is different from others, as it provides a new view of dealing with criminal-related documents written in Portuguese language.

The key contributions of this section are:

- to extract the event type and elements to answer the *5W1H* information in the Portuguese language applied to the criminal domain;

- to set a triple format that easily enables the use of several tools or knowledge bases, like graph databases;
- to extract the crime type and criminal terms to enable comprehension of the domain by adding information that is, only, connected to the criminal domain.

We need to give some context to our proposal: crime is usually a sequential chain of events spread across multiple documents. These events represent several situations and actions that can lead to a crime, based on transitional or permanent events that link entities (e.g., people or organizations) (Braz, 2019), representing several situations and actions that can lead to a main criminal event. We can surpass this issue by identifying the questions that criminal police seek to answer during a criminal investigation and are referred to in each police investigation report, by following the suggestions shaped by Braz (2019) - the author describes the investigation technique, supported by the identification of the answers to 5WH1 information. This technique is also used in journalism to assemble all information about a story. However, we have applied this technique to the criminal domain to synthesize the criminal information contained in each sentence by filtering them into small portions of texts that try to answer each *5W1H* question, enumerated below:

- *Who*: who was involved? (like a person or organization). We use the term *Whom* that identifies the object of the sentence - the recipient;
- *What*: what happened? (describes an event by its type);
- *When*: when did it take place? (like time or date);
- *Where*: where did it take place? (like a location);
- *Why*: why did it happen? (describes the motivation for a certain crime).

For example, the following normalized sentence describes a crime committed by two persons over an organization at a specific time or place. Here we can easily denote an event type, named "assalto" (in English "robbery") that is the core element and its connection to persons, organization, place, or a given time.

> *"O Rui Silva* ***WHO*** *e o Pedro Silva* ***WHO*** *assaltaram* ***WHAT*** *o Banco de Portugal* ***WHOM*** *em Coimbra* ***WHERE****, pelas 14 horas* ***WHEN****, por necessitarem de dinheiro para estupefacientes* ***WHY****."*

The *1H* question, known as *How?*, was omitted from this work, because it is the following step on data representation. After all, only after the representation, we can reconstruct the steps that justify a given crime, its "modus operandi".

6.3.1 The Extraction Method

This method relies upon the SRL, NER, and *Criminal Term Extraction* modules, to semantically analyze the criminal-related documents. We established the answers that will lead to understanding the meaning of each document (sentence by sentence) by using the verb (predicate) and the semantic roles nested with NE and criminal terms to identify the actors that are linked to a defined event.

Table 6.2 describes relations between *5W1H* information, the *SRL Chains*, *NE Pairs*, and *Term Pairs* outputs, thereby:

- the *SRL Chains* details the meaning of each sentence, by identifying and classifying the event type, using the *Predicate* to answer *WHAT* question. The semantic roles (Arg_0, Arg_1, ...) are used to identify the roles of each element and permits the identification and classification of the others *5Ws*. This outputs are associated with:
  - the *NE Pairs*;
  - the *Term Pairs*.

The combination of each output allows us to fine-tune in detecting where each NE makes part of the semantic roles.

5W1H Information	SRL Chains	Other Extractions	NEs Term Pairs	Description
What	PREDICATE	—	—	identifies the event type
Who	ARG0	—	Persons, Organizations, Locations Time/Date, Numeric, License Plates Emails Accounts, Narcotics, Crime Types Criminal Role	event Executant
Whom	ARG1	—	Persons, Organizations, Locations Time/Date, Numeric, License Plates Emails Accounts, Narcotics, Crime Types Criminal Role	event Recipient
When	AM-TMP	—	Time/Date	time or date of the event
Where	AM-LOC	—	Locations	location of the event
Why	AM-CAU	—	—	give an explanation for the event
—	All semantic tags	**Crime Type**	Crime Type Names	identifies a crime type name in event
—	All semantic tags	**Criminal Term**	Criminal Term	identifies a criminal term in event

Table 6.2: 5W1H Information/Other Extractions/SRL Chains/NE Pairs Matching.

The *Neo4j Criminal-Related Documents Representation* is divided into: the *5W1H Information Identification* and *5W1H Information Classification*:

- the *5W1H Information Identification* begins with *SRL Chains*, *NE Pairs*, and *Term Pairs* as inputs, then the three main points that we must take into consideration:
  - *Event Type Trigger*: event type identification that expresses an occurrence or action. POS tags, such as verbs, nouns, or adjectives (occasionally) that can be used for this identification. However, we focus on *Predicate* tag (*SRL Chains*) that identifies the term that will work as a trigger;
  - *Event Arguments*: for each predicate is generally built by a set of arguments with specific roles in the event, like location or temporal markers. We identify these arguments by using the predicate and SRL semantic roles, like:
    - Arg0: identifies the subject/executant marker on an event;
    - Arg1: identifies the object/recipient marker on an event;
    - AM-LOC: identifies the location marker on event;
    - AM-TMP: identifies the temporal marker on event;
    - AM-CAU: identifies the cause associated with an event;
    - Predicate: identifies the event type;
    - All semantic roles: used to identify the crime types or criminal terms related to the event
  - *Entity Markers*: the set SRL arguments have mentions to NE. Each entity is a reference to a thing that exists in the real world, such as persons or places. This enables the entities

identification and classification that compose each event's argument, using the NER results, such as persons, organizations, places, time, narcotics, and type of crime.

- the *5W1H Information Classification* classifies the *5W1H* Information depending on the existence of of NE. We rely on this phase on hand-crafted rules to enable the extraction of relationships between events and classified *5W1H* Information.

The figure 6.3 describes the method using an diagram.

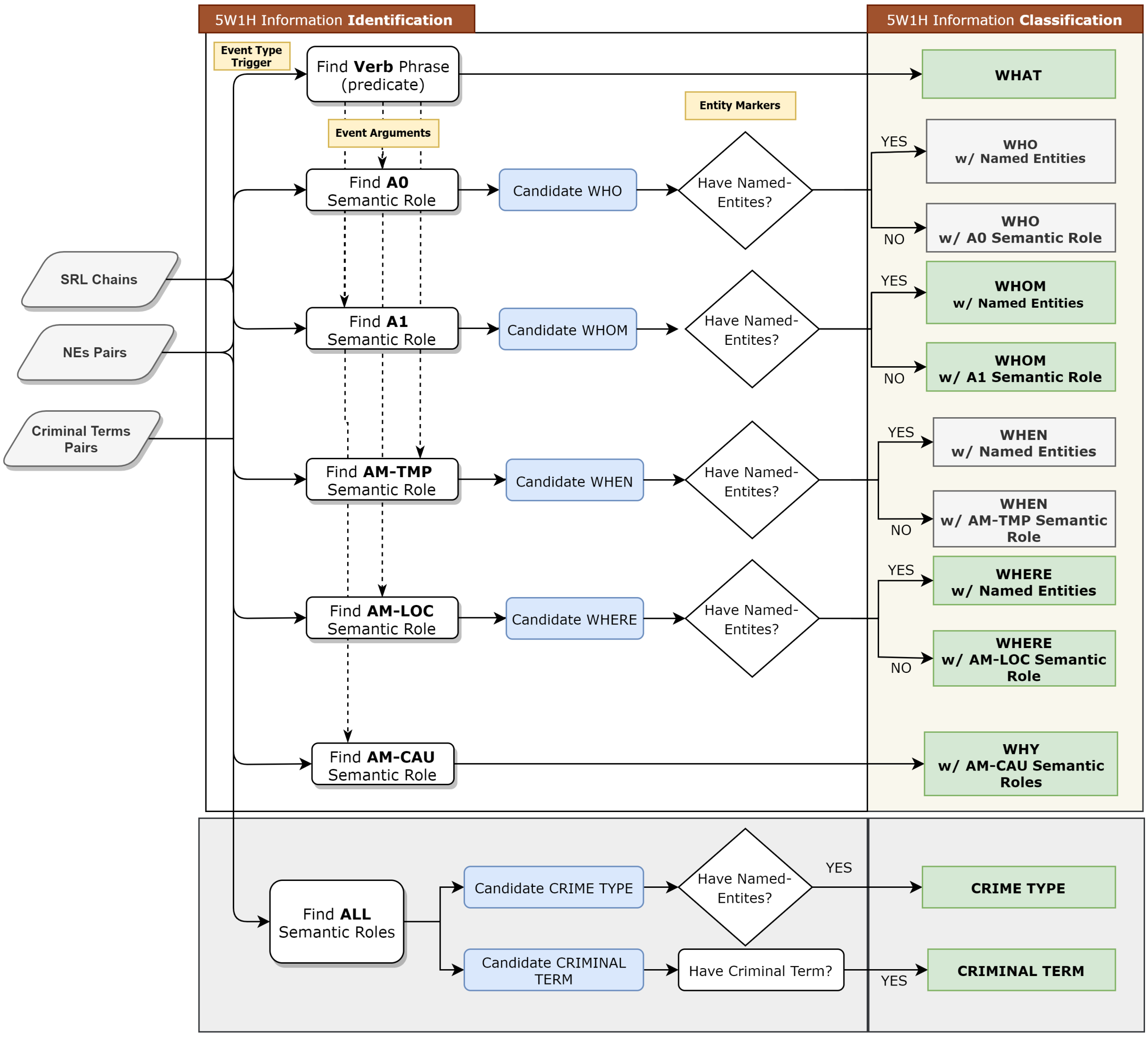


Figure 6.3: 5W1H Information Extraction Method Pipeline.

We established a set of hand-crafted rules to extract the information, and fill the Criminal Information Extraction Method file (XML format) (see listing 6.4). The extraction rules are enumerated on Table 6.3. For example, the triple is extracted by is Subject = Document$_{\mathbf{ID}}$ (ID is the document order by sequential numbers), and the Relation = WHO, and Object = NE identified by the NER module. This leads to a triple like: [Subject = Document01_Event01, Relation=WHO, Object="Pedro Silva"].

The Listing 6.4 shows the XML Schema for the *5W1H Information Extraction Method* output file. The given XML document have the following structure:

	Triple	
Subject	Relation	Object
Document$_{\mathbf{ID}}$_Event$_{\mathbf{ID}}$	**WHAT**	Predicate
	WHO	Named-Entities | A0
	WHOM	Named-Entities | A1
	WHERE	Named-Entities | AM-LOC
	WHEN	Named-Entities | AM-TMP
	WHY	AM-CAU
	CRIMETYPE	Named-Entity
	CRIMINALTERM	Criminal Term

Table 6.3: Hand-Crafted Rules Using a Triple (S,P,O) Representation.

- the root element "DocumentID" describes the criminal-related document by the concatenation of Document Name and ID;

- the element "DocumentID_EventID" describes the joint of "DocumentID" and "EventID" (separated by underscore) to identify the events per document. The data is retrieved from the subject of the triple;

- the child elements of element "DocumentID_EventID" identifies the main core of extractions, the 5W's, Crime Type, and Criminal Term. The data is retrieved from the object of the triple, such as named-entities;

- the attribute actortype is used to store the entity type of each named-entity identified.

```
<?xml version="1.0" encoding = "UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="DocumentID">
  <xs:element name="DocumentID_EventID">
        <xs:complexType>
               <xs:sequence>
                 <xs:element name="WHO" type="xs:string"/>
                 <xs:attribute actortype="Entity Type" type="xs:string"/>
                 <xs:element name="WHOM" type="xs:string"/>
                 <xs:element name="WHEN" type="xs:string"/>
                 <xs:element name="WHERE" type="xs:string"/>
                 <xs:element name="WHAT" type="xs:string"/>
                 <xs:element name="WHY" type="xs:string"/>
                 <xs:element name="CRIMETYPE" type="xs:string"/>
                 <xs:element name="CRIMINALTERM" type="xs:string"/>
               </xs:sequence>
        </xs:complexType>
  </xs:element>
</xs:element>
</xs:schema>
```

Listing 6.4: Criminal Information Extraction Method File - XML Schema.

6.3.2 Experimental Setup

We have established an experimental setup to evaluate our *5W1H Information Extraction Method*. This method used results obtained by other modules already proposed during this thesis. Furthermore, this setup was supported by a prototype developed in Java. Our initial experiments used a single sentence to simplify the understanding of our method, and also, a group of 20 (manually annotated) criminal-related documents were evaluated using *Precision (P)*, *Recall (R)* and *F-Measure (F1)* metrics.

To demonstrate the method and prototype results, a sentence was submitted to the prototype. Figure 6.4 shows graphically how the method works and established the linkages between data extracted, such as event type, relations, and NEs.

> *"O Rui Silva e o Pedro Silva assaltaram o Banco de Portugal em Coimbra, pelas 14 horas."*
>
> *In English*: "Rui Silva and Pedro Silva robbed the Bank of Portugal in Coimbra, at 2 pm".

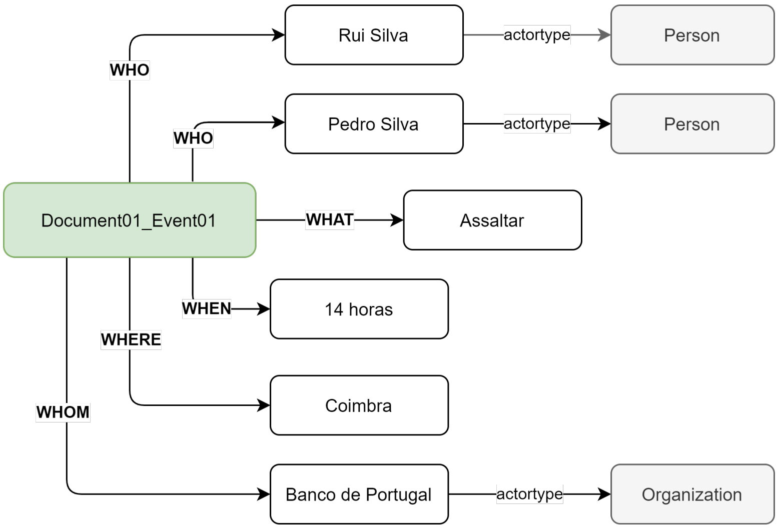


Figure 6.4: Graph Representation of the Example Sentence.

The graph illustrates the links between events and entities, gathered from the answers for the *5W1H* information. In addition, this approach enables us to understand the event in terms of its elements, as for the role played in the event using the Arg_0 and Arg_1 of SRL arguments.

Finally, the triples are represented in a XML file, following the requirements of XML Schema, in the listing 6.4.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Document01>
   <Event id="Document01_Event01">
      <WHAT>assaltar</WHAT>
      <WHEN>14 horas</WHEN>
      <WHO actortype="Person">Pedro Silva</WHO>
      <WHO actortype="Person">Rui Silva</WHO>
      <WHERE>Coimbra</WHERE>
      <WHOM actortype="Organization">Banco de Portugal</WHOM>
   </Event>
</Document01>
```

Listing 6.5: 5W1H Information Extraction Method XML File output.

We need to evaluate the proposed method using the evaluation measures *Precision*, *Recall* and *F-Measure*. To our knowledge, there is no suitable Portuguese annotated corpus for an evaluation fitted to this task. Therefore, we have annotated a set of criminal-related documents with the help of external annotators. In the Listing 6.6, we can check an excerpt from an annotated document.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!-- Manual Annotated Document: Example -->
<CrimeNews01>
<Sentence num="01" description="O Pedro Henriques assaltou o Banco de Portugal em
   Coimbra.">
  <WHERE> Coimbra </WHERE>
  <WHO actortype="Person"> Pedro Henriques </WHO>
  <WHOM actortype="Organization"> Banco de Portugal </WHOM>
  <WHAT> assaltou </WHAT>
</Sentence>
--- more
</CrimeNews01>
```

Listing 6.6: Annotated Criminal-Related Document in XML Format .

The Table 6.4 shows the evaluation measures of 20 criminal-related documents regarding the evaluation of the *5W1H Information Extraction Method*.

	Precision (P)	**Recall (R)**	**F-Measure (F1)**
Criminal-Related Documents	0.732	0.634	0.653

Table 6.4: Evaluation Measures for 5W1H Information Extraction Method.

The Table 6.5 shows the score of *5W's* elements, independently, using the method over 20 criminal-related documents.

	Who			What			Where			When			Why		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Criminal-Related Documents	0.686	0.683	0.675	0.727	0.672	0.689	0.723	0.615	0.662	0.817	0.575	0.646	0.708	0.625	0.583

Table 6.5: Evaluation Measures of each 5W's questions.

By analyzing the table 6.5, we can see that the results obtained for a first approximation are encouraging. From the *5W's* elements analyzed, the *Why* element performed worst for the *F1* (below 0.60), the others elements achieved a better results regarding *F1* (above 0.60). The *What* and *Who* elements are comparatively easier to identify against the other elements. Finally, this method is limited to the results obtained in *SRL* and *NER* methods. If we analyzed the combined measures, the results obtained here are not far below.

6.4 Graph Database Population and Enrichment

We have gradually introduced computational methods by using a well-defined pipeline to retrieve information from unstructured data in criminal-related documents into the *Neo4j* graph database to convert data into actionable knowledge. The population and enrichment of the graph database were the input file is in the form of XML, containing *5W1H* information that semantically represents the knowledge extracted from criminal-related documents. Notwithstanding, no schema or guidelines could transform unstructured data into graph databases. For the data creation, we used *Neo4j* embedded in Java applications [5] and the *Cypher Query Language* (for further information, see Chapter 2).

The key contributions in this section are:

- explores the feasibility of using a graph database to store data extracted from criminal-related documents, using property graphs to allow for a more flexible model to store our data;
- to apply the *Neo4j* database to the criminal domain, more specifically, the representation of the semantics of Portuguese criminal-related documents;
- to propose a method that automatically populates a graph database with events and entities;
- to propose a method to enrich nodes with external knowledge bases, more specifically, the *GeoNames* geographical database;
- the ability to cluster related information querying across many different criminal-related documents.

6.4.1 Modeling Decisions

Our approach follows the labeled property graph concepts made up of nodes, properties, and relationships. First, we needed to establish the modeling decisions (Bruggen and Mohanta, 2014) (Robinson et al., 2013) regarding the graph schema. This modeling decision has in mind that a graph database is a schema-free approach, meaning that the schema is implicitly defined by the stored data that is not formally described (Daniel et al., 2016), which gave us the flexibility to introduce other graph elements when need. Therefore, we defined each element that composed a graph database, such as *nodes*, *properties* and *relations*. We used the Unifed Modelling Language (UML) and followed the name conventions and recommendations [6] to define these modelling decisions. We adopted a strategy, already followed in section 6.3, where event handling was defined by a set of questions that we wanted to ask for data - the *5W1H* information. Also, crime types and criminal terms mentions were identified as a node.

Our work is significantly inspired by the Simple Event Model (SEM) (Van Hage et al., 2011) ontology and how the concepts and relations were defined into the ontology and their modeling decisions. With this in mind, we have provided it with the necessary database elements for semantic annotation of criminal-related documents. We have defined the nodes by determining whose nodes have answered the *5W1H* information (plus crime type and criminal terms). Therefore, to represent the semantics contained in criminal-related documents, we defined a set of six nodes: Event, Event Type, Actor, Time, Location, and Crime. Figure 6.5 shows the UML with modeling decisions to support our graph database module.

Clusters of nodes were created from the graph database population. Sometimes, these nodes identify the same information. For example, the **:Actor**, **:Location**, **:Time**, or **:Crime** nodes are subject to multiple

[5] See `www.neo4j.com/docs/java-reference/current/java-embedded` [Accessed: 1 May 2020].

[6] See `www.neo4j.com/docs/cypher-manual/4.0/syntax/naming` [Accessed: 1 Jul 2020].

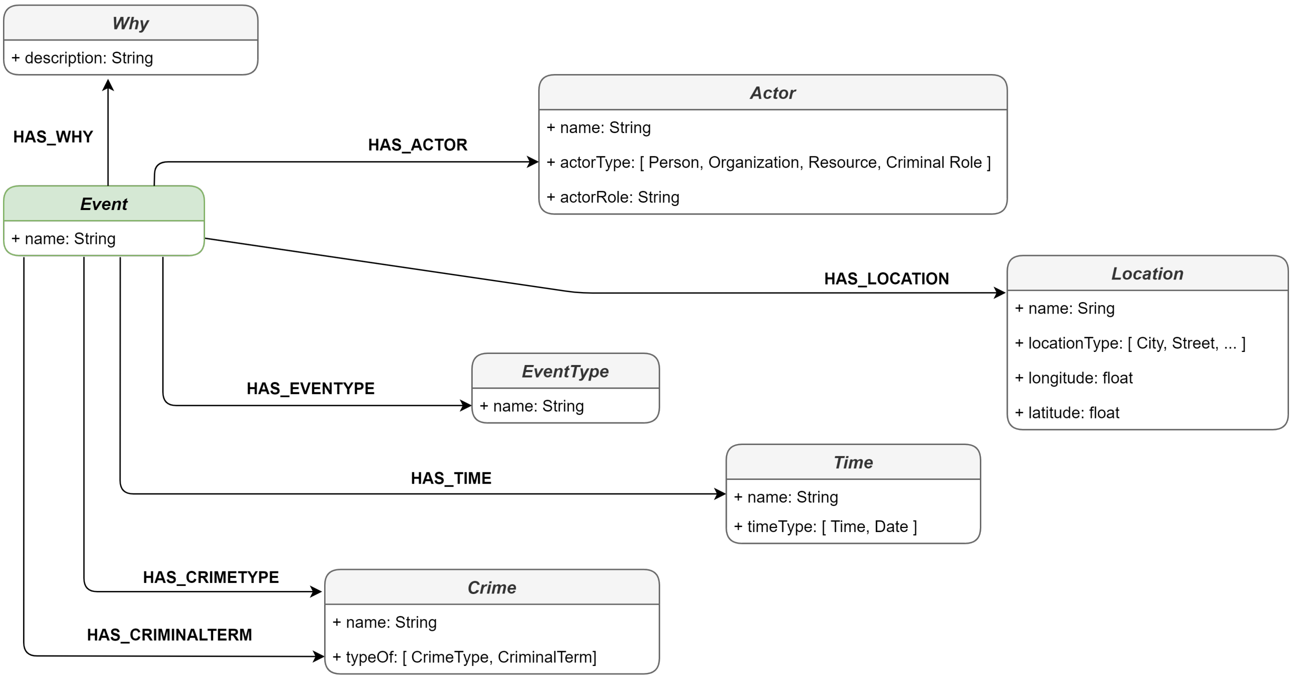


Figure 6.5: UML with Modelling Decisions.

mentions (the same name property). Therefore, we established a set of rules to eliminate this redundancy and enable the connection between nodes. Thus, we introduce the following rules:

- **Rule 1**: the **:Actor** is merged with another **:Actor** node whenever the property **name** is equal in the same or multiple documents;
- **Rule 2**: the **:Location** is merged with another **:Location** node whenever the property **name** is equal in the same or multiple documents;
- **Rule 3**: the **:Time** is merged with another **:Time** node whenever the property **name** is equal in the same or multiple documents;
- **Rule 4**: the **:Crime** is merged with another **:Crime** node whenever the property **name** is equal in the same or multiple documents.

Figure 6.6 shows an example of the rules 1 and 2 applied to a group of events from the same document. The events are linked by the **:Actor** and **:Location** nodes.

In the following subsections, we describe each node, their properties and relations.

Event Node

The event node, uses **Event** as **Label** and is referred as **:Event**. This node acts as a central node and it was introduced to identify each event using a unique identifier. The **name** property identifies the event ID (an unique key), a new **:Event** added to the graph database. This unique key was borrowed from the previous chapter, stored in XML element by the name Event (with an ID).

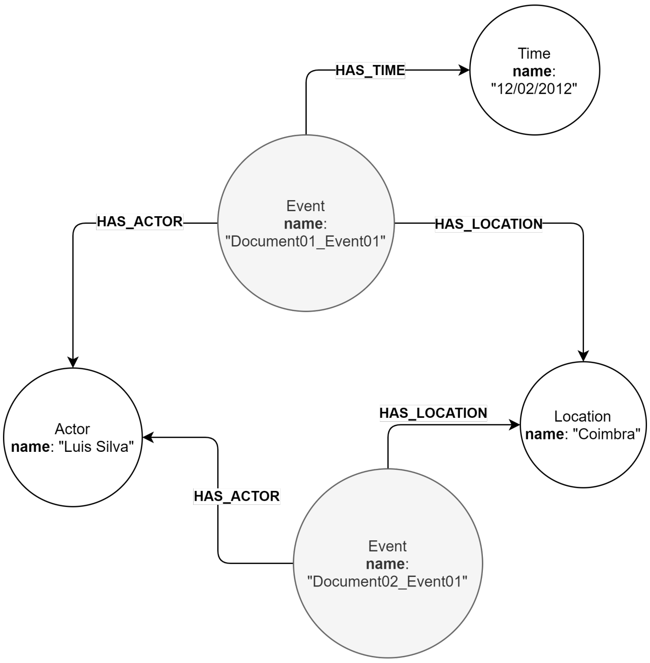


Figure 6.6: Graph Database Diagram using the Proposed Rules (Example).

EventType Node

The event type node, uses **EventType** as **Label** and is referred as **:EventType**. This node tries to answer an important question, the *What*, in the criminal domain. The *What* determines *"what happened?"* in a context within an entity set, such as persons or organizations, and gathered on a specific time and place. The **name** property defines the event type, such as "Homicide", "Assault", or "Kidnapping", which is retrieved from each event. Other event types were retrieved, such as "phone call" or "talk to". Conceptually, a specific event type is associated with a Portuguese verb category by sharing a common behavior or meaning. To normalize the verbs, we have applied the lemmatization tool (introduced in Chapter 4), such as "assaltaram" to "assaltar" or "telefonaram" to "telefonar".

Regarding to relations, we defined:

- the ***HAS_EVENTTYPE*** relation between **:Event** and **:EventType**.

Actor Node

The actor node, uses **Actor** as **Label** and is referred as **:Actor**. This node aims to answer a set of questions, such as *"Who participated or did something"*. These questions, when answered, highlight the involved persons, organizations, or resources. Thus, inanimate actors, like weapons or account banks, are identified in this node as resources. Therefore, the *Who* established by **:Actor** node, and how an actor is participating in an event, and its role in it. This node (an active entity of the domain that can activate or perform events) holds data (retrieved from the corpus) that are part of a given event, actively or passively. The value **resource** for the **actorType** property defines the resources used in an event, meant to describe *"with what"* is something being done. This can range from physical objects like a boat to animate entities when involved in a situation as an instrument. For this, we have defined three properties to **:Actor** node:

- **name**: defines the name of the actor, such as a person name;
- **actorType**: defines the type of actor that node represents in this context, like persons, organizations, or resources.
- **actorRole**: assigned the role of the actor in a specific event, like a homicide or police officer.

Regarding to relations, we defined:

- the ***HAS_ACTOR*** relation between **:Event** and **:Actor**. We add an property, named as *roleType*, that is used to identify the "Executant" or "Recipient" in an event.

Time Node

The time node, uses **Time** as **Label** and is referred as **:Time**. This node was introduced to answer the *When* question - something that happened at a specific time and date. The definition of a **:Time** node enables the mapping of time or date into a node. To be noted, temporal information is an important element in expressing criminal events, thus enabling the timeline's location of events. We have defined properties to the **:Time** node:

- **name**: defines the name attributed to the **:Time** node, in a time or date format;
- **timeType**: defines the type of time that node represents in this context, like time or date.

Regarding to relations, we defined:

- the ***HAS_TIME*** relation between **:Event** and **:Time**.

Location Node

The **Location** node, uses **Location** as **Label** and id referred as **:Location**. This node is meant to describe *Where* something has happened, like a city or village, or, more specifically, the crime scene location. It can be a public location or a place in a city (Coimbra). Thus, some places are locations where an event happened, or it is somehow connected. We have the following properties for the **:Place** node:

- **name**: defines the name of the location;
- **locationType** property is suitable to assign the type of location;
- **Longitude**: stores the value related to longitude value;
- **Latitude**: stores the value related to latitude value;

The **:Location** node could be enriched with georeferencing data, therefore, we add two properties to **:Location** node that assigns **latitude** and **longitude** properties, like the World Geodetic System, the WGS-84 with longitude, latitude (x, y). These two properties will be populated by an enrichment task done in section 6.4.2. This node can represent concrete and symbolic places, such as "inside house".

Regarding to relations, we defined:

- the ***HAS_LOCATION*** relation between **:Event** and **:Location**.

Crime Node

The **Crime** node, uses **Crime** as **Label** and is referred as **:Crime**. This node is introduced to identify the criminal-related terms that are present in events. In this domain, the representation of terms related to the criminal domain is useful to support expert decision-making and correlation between events and entities and outcomes. We have two properties to **:Crime** node:

- **name**: defines the name attributed to certain type of crime or criminal action;
- **typeOf**: defines the crime type, or criminal term.

Regarding to relations, we defined:

- the ***HAS_CRIMINALTERM*** relation between **:Event** and **:Crime**;
- the ***HAS_CRIMETYPE*** relation between **:Event** and **:Crime**.

6.4.2 Population and Enrichment tasks

These tasks are used to populate and enrich the database with the extractions performed in section 6.3.1, and the modelling decisions established in section 6.4.1, thereby:

- for the database population, we used as input the XML file that carries the semantic information extracted from the criminal-related document, a list of mappings between *5W's* (plus crime type and criminal term information) and the graph database elements. See Table 6.6 with the mappings;
- the database enrichment is performed by using *GeoNames API* as a geographical database to enrich the **:Location** node properties.

Population task

The **population** task analyses the XML file, and using the mappings defined, creates the nodes, relations, and properties of the database. Figure 6.7 shows an example of mapping. On the left side, we have a portion of the XML file with Event ID and WHAT elements. They map into the right-side database, creating two nodes: the :Event and EventType nodes, filling the name properties. The relation is established using the HAS_EVENTTYPE. Technically, the structure is built in temporary memory for each event and committed to the *Neo4j* database.

5W1H Information	Other Extractions	Definition	Graph Database Elements
WHO	—	actor involved in event (animated and inanimated)	:Actor node has_actor relation
WHAT	—	event type	:EventType node has_eventtype relation
WHERE	—	location of the event	:Location node has_location relation
WHEN	—	time or date that an event occurred	:Time node has_time relation
WHY	—	reason or cause	:Why node has_why
—	**EVENT**	event identification	:Event node
—	**CRIMINAL TERM**	criminal term identification	:Crime node has_criminalterm relation
—	**CRIME TYPE**	crime type identification	:Crime node has_crimetype relation

Table 6.6: Mapping Information Extracted and the Graph Database Elements.

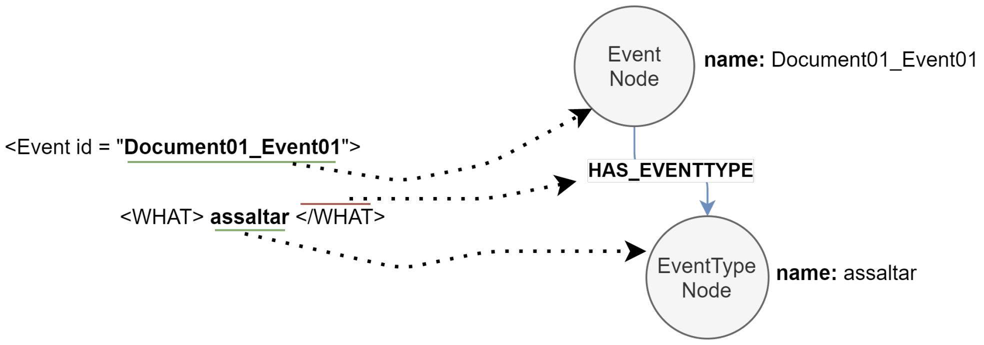


Figure 6.7: Mapping Example.

Enrichment task

The *enrichment task* adds new information related to location concept, by using the *GeoNames* geographical database, by using the *GeoNames API*. The data enrichment task performed a series of processing steps in order to enrich the terms againts *GeoNames API* results:

- identifying the **name** property from the list of **:Location** nodes;
- sends a API request using **name** property for relevant geographical data, and, in our case, longitude and latitude (geography coordinates). The API response to add **latitude** and **longitude** properties that identify the exact correspondences regarding terms used in API request :Location node;

We have select these two properties to be added by this enrichment task because they add geographical coordinates that enable its use in external tools for hotspot map development. Figure 6.8 shows the Enrichment Data diagram that describes the method to enrich the **Location** node properties, the longitude, and latitude coordinates.

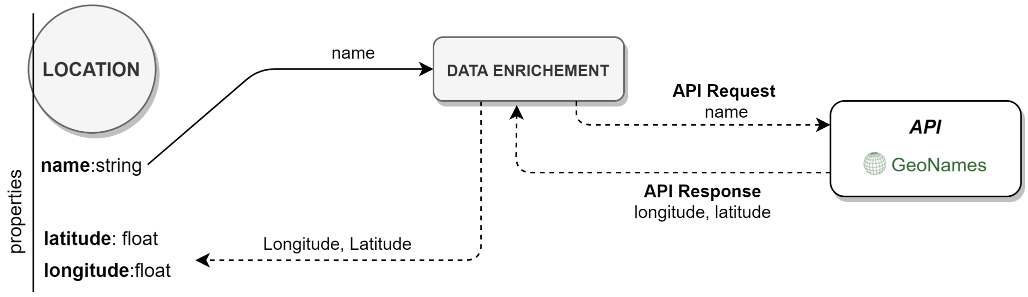


Figure 6.8: Enrichment Data Scheme.

6.4.3 Experimental Setup

An experimental setup was proposed to evaluate our proposal using a prototype that combines the *Neo4j* database (version 3.5.12) and Java Programming language. To illustrate our proposal, we start by using an example with two sentences.

> *"O Rui Silva e o Pedro Silva assaltaram o Banco de Portugal em Coimbra, pelas 14 horas. O Rui Silva telefonou ao Pedro Silva, com recurso ao telemóvel Nokia, com o nº 959999000."*

We perform a manual analysis of the sentences, and schematic represented the events and connects entities (this exercise is similar to the ones performed by domain experts when analyzing criminal investigation reports) by using the defined mappings. Thus, we can verify that:

- two events were identified, namely "Document01_Event01" and "Document01_Event02";
- two event types were identified in a dictionary form, namely "Telefonar" and "Assaltar";
- the entities that we extracted from sentences were "Rui Silva", "Pedro Silva" as a person, "Mobile Phone" as a resource, and "Banco de Portugal" as an organization;
- the location and time were identified, namely "Coimbra" and "14 horas";
- the relations were established by linking the event to each entity, location, and time;
- the two actors ("Pedro Silva" and "Rui Silva") shared the same events, which allows the connection between the events.

The result is schematically represented in figure 6.9. Then, we used the prototype to evaluate the two sentences. The figure 6.10, shows the results, the graph database after been populated, obtaining 10 (ten) nodes and relations.

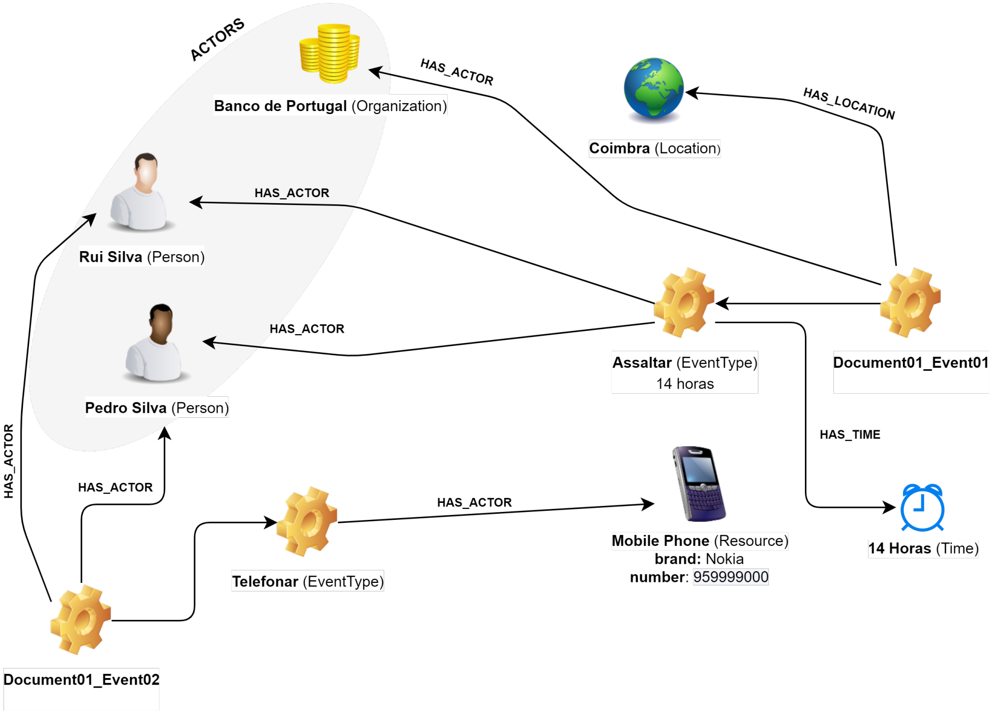


Figure 6.9: Sentences Diagram.

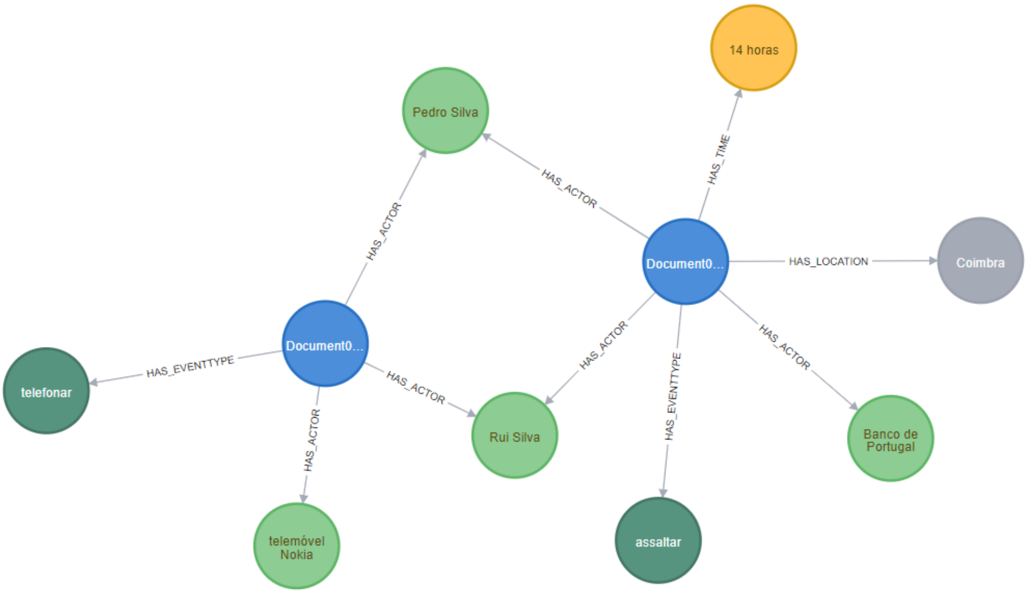


Figure 6.10: Neo4j Graph Database (Example).

We proceed with the experiment using a total of 20 criminal-related documents. Table 6.7 enumerates the amount of graph elements (nodes and relations) with six different relationship types. Another value extracted from this prototype was the number of responses of *API GeoNames* for location names, obtaining 97 responses with latitude and longitude coordinates.

	#Count
Nodes	889
Relations	911
Enrichment task (Geonames)	97

Table 6.7: Amount of graph elements populated and enrichment task (geonames).

The figure 6.11 shows an excerpt of the nodes and relations created by executing our graph population and enrichment method.

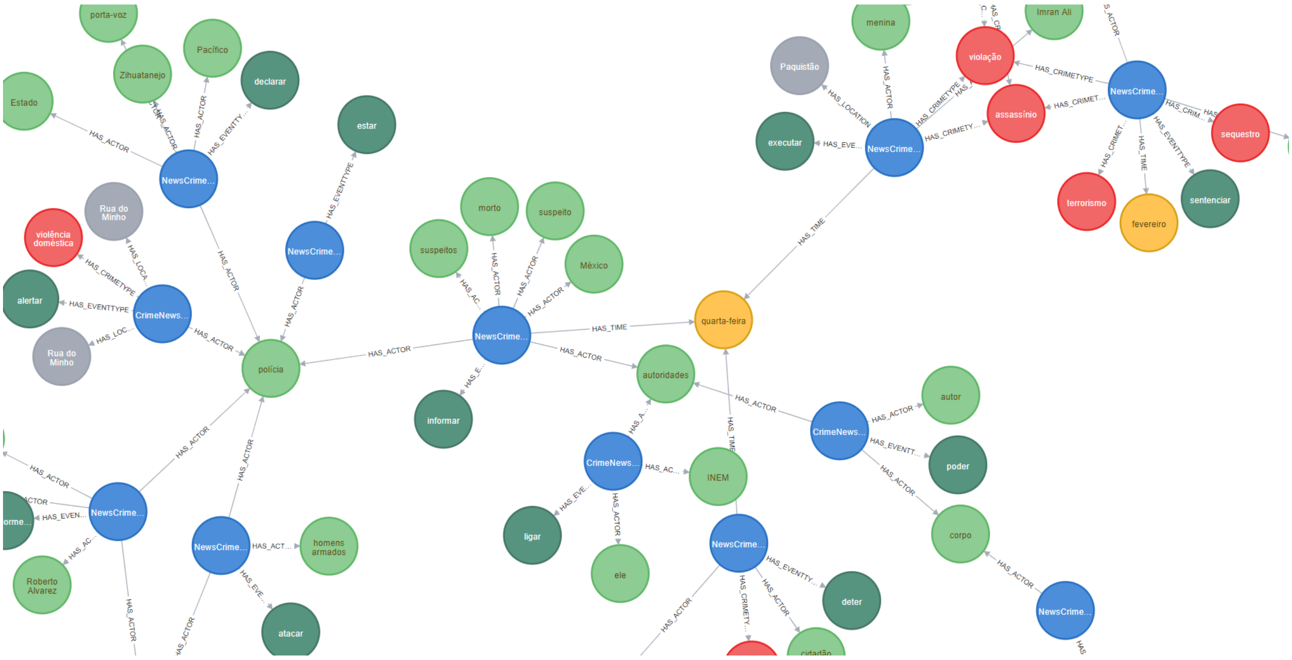


Figure 6.11: Neo4j Graph Database screenshot.

6.5 Application Example: a Criminal Investigation Report

The objective of this section is to evaluate our proposal against a real criminal investigation report. Therefore:

1. a criminal investigation report was analyzed with a domain expert, where each person and location names, phone numbers, or license plates were changed or masked. Notice: all names, phone numbers, locations, or other personal identification information are fictional (in this example, and all examples along with the thesis) for data confidentiality;

2. the report was submitted into our framework;

3. the domain expert analyzed the criminal investigation report using the IBM™ i2 Analyst's Notebook tool.

The criminal investigation report, refereed in (2), was submitted into *SEMCrime* framework in Microsoft™ Word format, thereby:

- the criminal investigation report was submitted to our framework to be processed by the *Preprocessing Criminal-Related Documents* and outputs an semi-structured format (XML), as showed in listing A.2;
- the results obtained after NLP pipeline, detecting 27 sentences and 828 tokens;
- the report was submitted to *Criminal Information Extraction* module producing an output file;
- the *Neo4J* graph database was populated and enriched.

Referring the *Graph Database Population and Enrichment* module that creates an graph database with 155 nodes and 159 relations graph elements. The figure 6.17 shows an excerpt of the *Neo4J* graph database.

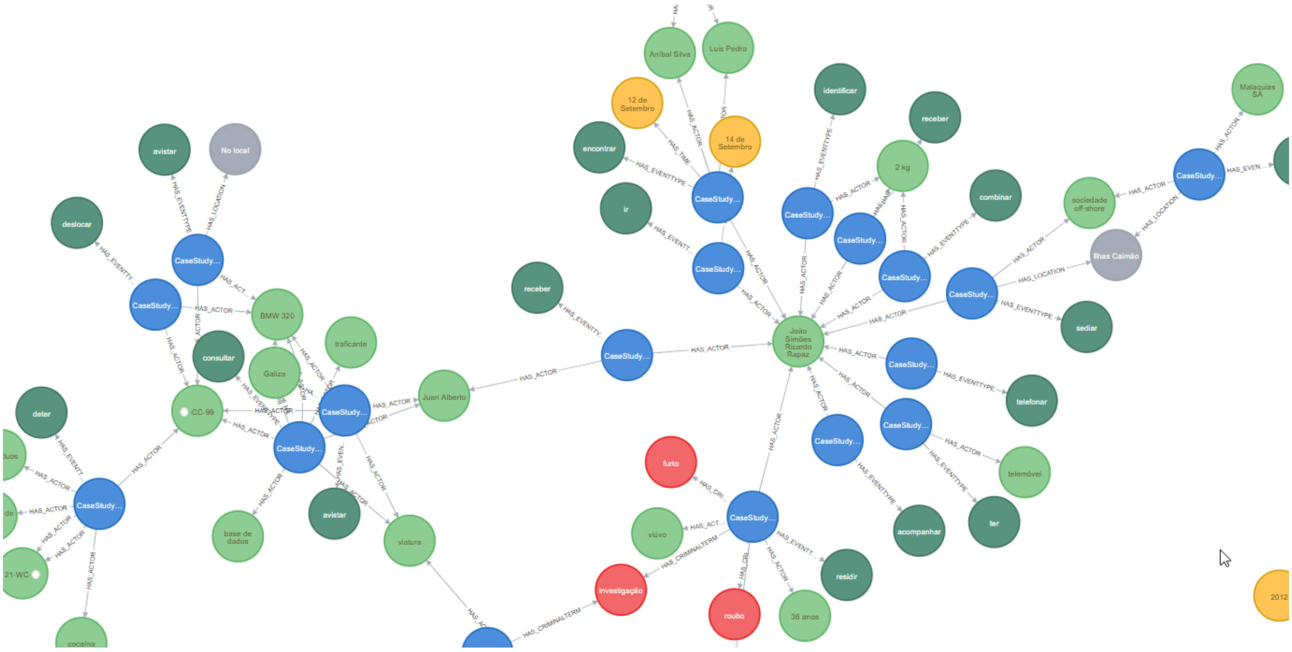


Figure 6.12: Neo4j Graph Database Screenshot (Application Example).

In (3), a domain expert analyzes the report with the IBM™ i2 Analyst's Notebook tool executing similar proceedings for reports analysis performed in a police department (see Section 1.2 that details the problem). Therefore, an Microsoft™ Excel sheet was built with entities and relations and then imported into IBM™ i2 Analyst's Notebook tool (see output results in A.2).

The figure 6.13, and figure 6.14 shows partial results of our proposal and IBM™ i2 Analyst's Notebook output. We marked the figures with numbers with similar findings from the two approaches. As we can see, our approach was able to identify the same entities and the relationships between them, which may differ in the name (depends on the domain expert's perspective and could differ from each other). From the obtained results, we can infer the viability of the *SEMCrime* framework to extract and represent the data retrieved from criminal-related documents.

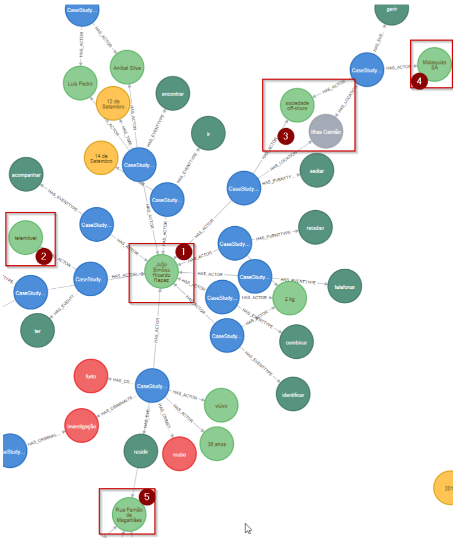


Figure 6.13: Application Example Screenshot.

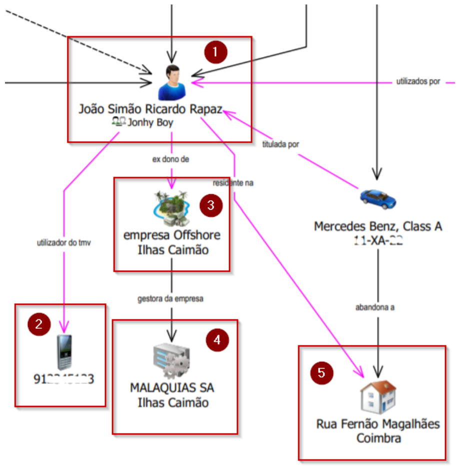


Figure 6.14: IBM™ i2 Analyst's Notebook Screenshot.

Cypher Queries and Algorithms

After the graph database population, we can start using it by querying the database using the *Cypher Query Language* statements. Therefore, we need to identify a set of answers that we want to get from the graph database. Then, a specific question on the data will be derived and answered by querying the graph data model introduced in Section 6.4.1.

Question 1: *Who are the persons enumerated in this report ?*

```
MATCH (a:Actor)
WHERE a.actorType="Person"
return a
```

Listing 6.7: Question 1 Cypher Query.

To answer the question 1, we have created the query described in listing 6.7 that asks for all *Actor* that have the *actorType* equal to *Person*. The output of the query is shown in figure 6.15.

Question 2: *Who is/are the actor(s) that are related to a consumption of narcotic drugs (in Portuguese "consumo de estupefacientes")?*

```
MATCH (e:Event) - [r:HAS_CRIMINALTERM] -> (c:Crime),
(e:Event) - [r1:HAS_ACTOR] -> (p:Actor)
WHERE c.name="consumo de estupefacientes"
return e,r,c,r1,p
```

Listing 6.8: Question 2 Cypher Query.

To answer the question 2, we have created the query described in listing 6.8 that asks for the *Event* that related an criminal term action and a *Actor*. The output of the query is shown in figure 6.16.

Question 3: *What is the shortest path between "João Simões Ricardo Rapaz" and "Malaquias SA" ?*

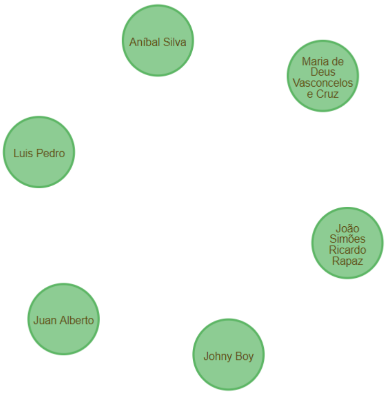


Figure 6.15: Cypher Query Output (Question 1).

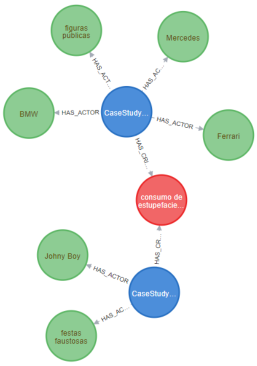


Figure 6.16: Cypher Query Output (Question 2).

To achieve the shortest path between the principal suspect "João Simões Ricardo Rapaz" and an off-shore organization named "Malaquias SA". We used an inbuilt shortest-path search algorithm present in *Neo4j*, like the *single shortest path algorithm*. The algorithm parameters were set up with start and end nodes, without specifying the direction of the relations, and the path is a maximum of 10 relationships long (this means that from the start node to end node, we have a maximum of 10 relations between each other).

```
MATCH (start:Actor{name:'João Simões Ricardo Rapaz'}),(end:Actor{name:'Malaquias SA
    '}),
p = shortestPath((start)-[*..10]-(end))
RETURN p
```

Listing 6.9: Single Shortest Path algorithm execution.

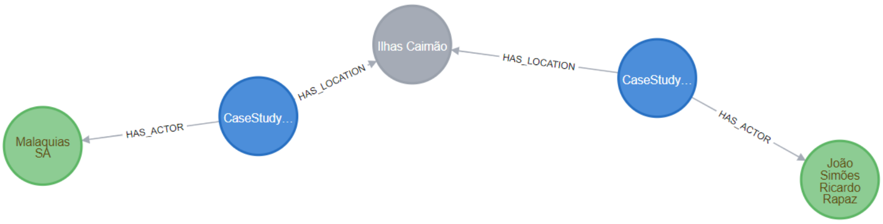


Figure 6.17: Cypher Query Output (Question 3).

This is a path-finding problem, where the algorithm tries to find the degrees of separation between actors, such as persons and organizations. As we can see, the obtained result by the algorithm, we verify the path between the suspect actor, *João Simões Ricardo Rapaz*, and the organization *Malaquias SA*, passing through the *Cayman Islands* (in Portuguese: Ilhas Caimão).

6.6 Concluding Remarks

This chapter describes a method for data retrieval from criminal-related documents and the representation into a *Neo4j* Graph Database. To complete this, we combine a set of modules, like the *Criminal Term Extraction* module, a NER sub-module to identify and classify the relevant NEs, and *Semantic Role Labelling* module for semantic roles identification and classification. This combination was used to develop the *5W1H Information Extraction Method*, associated with crime type and criminal terms identification. The results of this method permit us to answer the following research question *"Can we extract information from criminal-related documents by using a 5W1H approach?"* and the possibility to extract information using *5W1H* information was accomplished. Therefore, this approach has the following insights:

- understand the criminal-related documents semantically, by using a sequence of events and related entities;
- followed the same process that domain experts use to investigate a particular crime: the answer to *5Ws*. Furthermore, this approach allows us the adaptation of the analyzed criminal investigation reports and news to the same *5W1H* approach, since they were written taking into account this information;
- presented the initial steps to extract events and entities in criminal-related documents in their unstructured form. Thus, it allows a basis for future related work regarding the information that may not have been extracted in this approach and improving it.

The research question *"Is it possible to propose a graph-based framework that allows us to represent information extracted from documents related to criminal investigations?"* is answered by representing the data extracted from criminal-related documents in the *Neo4j* graph database. In general, this representation focuses on how we can provide relevant information for police departments with different timeline investigations, such as historical or ongoing investigations, stored in a *Neo4j* graph database, thereby:

- the use of graph theory concepts to define nodes, relations, and properties to store criminal-related information of a given group of documents;
- the graph database allows for a more flexible and scalable representation for this domain and permits to reduce the semantic gap between heterogeneous sources to answer end-user queries.

We consider that, after the semantic task that elects concepts and relations that occur in criminal-related documents, *Neo4j* is flexible and powerful enough to store and represent such knowledge. Furthermore, the structure of linked data can be an advantage in the criminal field, since this is a dynamic environment where data can emerge from the different criminal investigations.

Finally, this chapter accomplishes its goal of representing criminal-related information into a graph database, where the modeling decisions enable the instantiating of the extracted data following the *5W1H* information, crime types, and criminal terms. We also added a task that enables the enrichment of data by using *GeoNames* geographical database to populate the **:Location** node with geo-coordinates (longitude and latitude) as properties. Hence, to retrieve the full power of graph databases, queries are recommended to extract sub-graphs to answer user questions, like the queries performed over the application example database. Even the simple ones can help domain experts search for answers regarding relevant questions during a criminal investigation.

7

Conclusions and Future Work

> "Study without desire spoils the memory, and it retains nothing that it takes in."
>
> *Leonardo da Vinci*

Throughout the study carried out in this thesis, we developed a framework, named as *SEMCrime*, to retrieve and understand the unstructured data that populates the criminal-related documents written in the Portuguese language. Using several computational methods applied to the criminal domain. For that to be possible, we have:

- focused on the Portuguese language, without discarding what has been done in other languages;
- studied and analyzed the approaches applied to the criminal domain and related work;
- surveyed which ETL, NLP, GDB approaches existed and, of those, which can be proposed, used or adapted for the task at hand;

Retrospectively, our proposal was ambitious, mainly because, the framework is applied to a domain without a solid background and relevant related work to the Portuguese language. Despite the works already published and applied to other cases such as the English language. Therefore, this is a unique approach to an emerging problem in police departments or related. During the resolution of the problem, we found several challenges related to the fact that:

- the criminal-related documents had different content structures and file formats;
- the extracted plain text, due to extraction or existing problems, may contain errors or "garbage";
- the existence of abbreviations and acronyms related to the domain;
- the existence of entities related to the domain, such as narcotics or crime types that are not identified and classified by the NER approaches;
- the use of domain-specific terms related to the criminal domain, such as "Pulseira Eletrónica" (in English: ankle bracelet);
- like other written text, the criminal-related documents need to be semantically understood.

We have studied related works to help us represent the data retrieved from criminal-related documents into a graph database. To perform such work, we identified the key tasks, like the analysis of the criminal-related documents and their content; we have adapted an ETL and NLP approach to our proposal, thereby:

- the documents were analyzed, and data was retrieved, performing tasks to clean, transform, normalize and load into a semi-structured format, producing a computer-readable format (XML format);
- abbreviations and acronyms were normalized to its extended form, like the token "PSP" refers to an acronym of "Polícia de Segurança Pública";
- a NLP pipeline was introduced, performing tasks like tokenization or sentence splitting.

From the tasks enumerated before, we verify that the criminal-related document's syntax does not differ from other domains, except on the use of abbreviations or acronyms related to the domain. Nevertheless, these tasks are useful for the following phases. To understand the meaning of documents, we have applied a set of tasks to enable a semantic approach, thereby:

- an NER module was used to identify and classify the NEs relevant to the domain. In this phase, we have proposed classifiers that identify NE related to the domain;
- to identify and classify the domain-specific terms related to the criminal domain, we added a module to perform such task using a gazetteer of criminal terms;
- the SRL was adapted to our approach to enabling the identification of the semantic roles.

Since our ultimate goal was to understand documents, we applied the *5W1H* approach in each sentence, which permits the identification and classification of the *5Ws* responses in the Portuguese language. This introduced a novel method that:

- identifies and classifies the *Who, When, Where, What, and Who* questions; using this method, we tried to find the 5W's answers in each sentence, which outputs a network of entities and relations.

We identify and classify the *5W's* questions by combining the NER, SRL, and *Criminal Term Extraction* approaches to extract information from criminal-related documents, named as *5W1H Information Extraction Method*. The output file permits the population and enrichment of the *Neo4j* graph database. This fact allows us to identify the following:

- the proposed method leads to a knowledge representation of the unstructured data presented in criminal-related documents;
- allows a cluster of events and entities and corresponding user queries;
- an enrichment task has also been added to provide GPS coordinates to location nodes.

We have developed prototypes for framework evaluation, obtaining promising results with room for improvements. The application example obtaining similar results as the manual analysis, such as identifying and classifying entities and event detection, which is promising for further evaluations. However, the events may differ from our proposal because it depends on the domain expert's perspective that could retain the sentence verb or introduce a new expression to create the relations. Overall, our approach performs well to represent a cluster of relations and entities that populate the criminal-related documents without discarding new improvements.

7.1 Published Papers

Several papers have been peer-reviewed and published in conference proceedings and submitted for publication during with this research work. The papers have been authored by the thesis author or co-authored with supervisors. Therefore, some of the work described in this thesis has aroused from these papers, such as:

- Carnaz, G., Nogueira, V.B., Antunes, M. (2021) **A Graph Database Representation of Portuguese Criminal-Related Documents**. Informatics 2021, 8, 37;
- Carnaz G., Beires Nogueira V., Antunes M., Ferreira N. (2020) **An Automated System for Criminal Police Reports Analysis**. In: Madureira A., Abraham A., Gandhi N., Silva C., Antunes M. (eds) Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018). SoCPaR 2018. Advances in Intelligent Systems and Computing, vol 942. Springer, Cham;
- Carnaz, G., Nogueira, V., and Antunes, M. (2019). **Knowledge representation of crime-related events: a preliminary approach**. In 8th Symposium on Languages, Applications and Technologies (SLATE 2019).Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik;
- Carnaz, G., Nogueira, V., and Antunes, M. (2019). **Simple event model ontology population using an information extraction system: A preliminary proposal**. In 9th Workshop in Informatics of the University of Évora (JIUE'19);
- Carnaz G., Quaresma P., Beires Nogueira V., Antunes M., Fonseca Ferreira N.N.M. (2019) **A Review on Relations Extraction in Police Reports**. In: Rocha Á., Adeli H., Reis L., Costanzo S. (eds) New Knowledge in Information Systems and Technologies. WorldCIST'19 2019. Advances in Intelligent Systems and Computing, vol 930. Springer, Cham;

- Carnaz G., Vitor B. Nogueira, Antunes M., Fonseca Ferreira N.N.M (2018). **Named entity recognition for portuguese police reports**. In 8th Workshop in Informatics of the University of Évora (JIUE'18);

- Carnaz, G., Bayot, R., Nogueira, V. B., Gonçalves, T., and Quaresma, P. (2017). **Extracting and representing entities from open sources of information in the Agatha project**. In 21th International Conference on Applications of Declarative Programming and Knowledge Management, INAP 2017, Wurzburg, Germany;

- Carnaz, G. and Caldeira, C. (2017). **Aplicação da ontologia PROV-O ao crime de branqueamento de capitais**. In 7th Workshop in Informatics of the University of Évora (JIUE'17);

- Carnaz, G., Nogueira, V., Antunes, M. (2017). **Ontology-Based Framework Applied to Money Laundering Investigations**. In 7th Workshop in Informatics of the University of Évora (JIUE'17);

7.2 Future Work

The research work carried out proposed contributions for the criminal domain. However, some aspects can be explored in future work:

- adds a new module that introduces the co-reference resolution. This module will help us to determine whether two expressions in the Portuguese language refer to the same entity in the real world in order to improve the identification and classification of entities and their relationships;

- the SRL module could be improved in performance. Therefore, the training of a new model could provide for better results in our information extraction method;

- to enable the answer of the *1H* question, the *How*;

- to improve the NER module with new NE related to the domain or relevant to domain;

- to train the *Criminal Role* and *Crime Type* NEs with a supervised approach, replacing the approaches carried out with the rule-based approaches and gazetteers;

- to propose an annotated corpus related to the criminal domain for a NER task, based on the dataset proposed in this work, and continuing to increase its volume and performing curation tasks;

- to understand the slang used by criminals and written in criminal-related documents. This recognition is essential to understand the text, its semantics. Given the use of expressions that can remove the context of the text and make it impossible to represent it correctly;

- to extract biographical information about persons and organizations (this includes the affiliations, such as gang affiliations).

A

Application Example

"Prediction is very difficult, especially about the future."

Niels Bohr

This appendix groups the documents and outputs related to the section Application Example: a Criminal Investigation Report. We initially listed the same criminal investigation report. First, the content of the original documents, and second, after being processed by Criminal-Related Documents Preprocessing module. Finally, the output of application example in the IBM™ i2 Notebook Analyst's.

A.1 Criminal Investigation Report

The criminal investigation report listed below was subjected to a process of anonymization, manually performed. The person names, phone numbers, licence plates, locations, and among others. They were anonymized by changing the original names to different ones (fictional names), without loosing the document context and meaning. The document was provided to us in the Microsoft™ Word format. Listing A.1 shows the content of the document:

```
Exmo. Senhor Coordenador de Investigação Criminal
Inspector: Luis Alberto.
NUIPC: 1222SSHH5678G.
Data da denúncia: 01 Julho 2020.
Relatório de Investigação Criminal.
Chegou ao conhecimento desta polícia, em 23 de julho de 2015, que um indivíduo
   conhecido por Jonhy Boy, residente na Rua Fernão de Magalhães, na localidade de
   Coimbra se dedica ao tráfico de droga, fazendo disso modo de vida, sem que lhe
   seja conhecida profissão, exibindo evidentes sinais exteriores de riqueza.
   Utiliza uma viatura da marca Mercedes Classe A, com a matrícula XX-XA-XX.
Também são referidas a existência de festas faustosas, por ele promovidas,
   frequentadas por diversas figuras públicas conectadas com o consumo de
   estupefacientes. Para além disso, a horas tardias, são vistas diversas viaturas
   topo de gama, por exemplo BMW, Mercedes e Ferrari, algumas delas pertencentes a
   indivíduos proprietários de estabelecimentos noturnos, locais conectados com
   consumo de estupefacientes.
Iniciada a investigação apurou-se que na morada Rua Fernão de Magalhães, reside João
    Simões Ricardo Rapaz, viúvo, 36 anos de idade, com antecedentes por furto e
   roubo, não lhe sendo conhecidas declarações de rendimentos nos últimos 5 anos.
   Informações complementares permitiram apurar que casou no ano de 2012, com Maria
    de Deus Vasconcelos e Cruz, 54 anos, que faleceu entretanto, no ano de 2017, em
    circunstâncias estranhas após uma queda da residência.
Anteriormente era proprietário de uma sociedade off-shore sediada nas Ilhas Caimão,
   que por sua vez geria uma empresa de importacao e exportacao de nome Malaquias,
   SA, sediada nas mesmas Ihas. Mais de apurou que tinha o telemóvel com o número
   91XXXXXXX.
Face aos indícios colhidos, que apontavam que o indivíduo se dedicaria ao tráfico de
    estupefacientes, foram realizadas as vigilâncias policiais infra, centradas na
   sua residência, e de onde resultou:
- No dia 12 de Setembro de 2015, o suspeito encontrou-se com Luis Pedro e Aníbal
   Silva. Mais tarde, com a ajuda da PSP de Coimbra, foi possível saber que estes
   dois se dedicavam a pequenos furtos.
- No dia 14 de Setembro, João Simões Ricardo Rapaz saiu da residência no Mercedes de
    matrícula NN-XA-NN, acompanhado de um rapariga loura, cujo a identificação não
   foi possível de fazer.
- Para além disso no dia 23, foi avistado um indivíduo tez escura, acompanhado de 2
   indivíduos corpulentos, cada um deles com volume ao lado direito de cintura,
   presumindo-se que fossem portadores de arma de fogo. Consultado a base de dados,
    verificou-se que o indivíduo era um conhecido traficante espanhol, da Galiza,
   com o nome de Juan Alberto, que se deslocava numa viatura da marca BMW 320 de
   matrícula XX-CC-XX.
Face aos estes fortes indícios, procedeu-se à interseção telefónica dos telefones
   que são por ele utilizados, com os números 91XXXXXXX, 96XXXXXXX, escutas essas
   que se iniciaram a 1 de Novembro de 2015 e terminaram em 12 de Dezembro de 2015,
    e de onde resultou:
```

```
- O João Simões Ricardo Rapaz recebeu uma chamada para o 91XXXXXXX de Juan Alberto,
    onde combinavam a entrega de farinha, no local do costume.
- O João Simões Ricardo Rapaz telefonou para o 92XXXXXXX, e pediu que lhe
    entregassem o Mercedes carregado. E que as banadas estavam prontas.
- O João Simões Ricardo Rapaz recebeu uma chamada de número não identificado, onde
    combinava a entrega de 2 kg de neve.
Dando continuidade à investigação, procedeu-se a vigilâncias na casa do suspeito e
    nos locais frequentados por este, apurou-se:
- Houve um transporte de um camião com matrícula XX-11-XX , vindo de Espanha. O
    suspeito seguiu atrás no Mercedes e foi descarregado num armazém em Aveiro.
    Percebeu-se que o indivíduo João Alberto, é o dono do armazém e amigo de João
    Rapaz. No local foram avistadas três viaturas, nomeadamente um BMW 320 de
    matrícula XX-CC-XX.
Tendo em conta este cenário , foram envolvidos 30 inspectores que procederam a
    buscas no armazém, local onde foram detectados 30 pacotes de cocaína. Foram
    detidos os indivíduos, tendo sido apreendidas armas e as viaturas com as
    matrículas XX-CC-XX, XX-WC-XX e XX-CC-XX, bem como 3 milhões de euros em
    numerário.
Foram realizadas buscas domiciliárias, que permitiram a apreensão de 6 milhões de
    euros, um quadro do pintor Picasso, DVD com imagem de vídeo vigilância relativas
     ao alegado acidente da mulher do João Simões Ricardo Rapaz e que contradizem a
    versão de acidente por ele apresentada.
```

Listing A.1: Criminal Investigation Report.

Listing A.2 shows the XML file, after being processed by the *Criminal-Related Documents Preprocessing* module:

```
<?xml version="1.0" encoding="UTF-8"?>
<ApplicationExample>
    <docname>Application Example</docname>
    <authors>Inspector Luis Alberto</authors>
    <crimeprocessnumber>1222SSHH5678G</crimeprocessnumber>
    <publicationdate>01 Julho 2020</publicationdate>
    <title>Criminal Investigation Report - Application Example</title>
    <documentbody>
                Chegou ao conhecimento desta polícia, em 23 de julho de 2015, que um
                    indivíduo conhecido por Johny Boy, residente na Rua Fernão de
                    Magalhães, na localidade de Coimbra se dedica ao tráfico de droga,
                     fazendo disso modo de vida, sem que lhe seja conhecida profissão,
                     exibindo evidentes sinais exteriores de riqueza.
                Utiliza uma viatura da marca Mercedes Classe A, com a matrícula XX-XA-
                    XX.
                Também são referidas a existência de festas faustosas, por ele
                    promovidas, frequentadas por diversas figuras públicas conectadas
                    com o consumo de estupefacientes.
                Para além disso, a horas tardias, são vistas diversas viaturas topo de
                     gama, por exemplo BMW, Mercedes e Ferrari, algumas delas
                    pertencentes a indivíduos proprietários de estabelecimentos
                    noturnos, locais conectados com consumo de estupefacientes.
```

Iniciada a investigação apurou-se que na morada Rua Fernão de Magalhães, reside João Simões Ricardo Rapaz, viúvo, 36 anos de idade, com antecedentes por furto e roubo, não lhe sendo conhecidas declarações de rendimentos nos últimos 5 anos.

Informações complementares permitiram apurar que casou no ano de 2012, com Maria de Deus Vasconcelos e Cruz, 54 anos, que faleceu entretanto, no ano de 2017, em circunstâncias estranhas após uma queda da residência.

Anteriormente era proprietário de uma sociedade off-shore sediada nas Ilhas Caimão, que por sua vez geria uma empresa de importacao e exportacao de nome Malaquias, SA, sediada nas mesmas Ihas. Mais de apurou que tinha o telemóvel com o número 96XXXXXXX.

Face aos indícios colhidos, que apontavam que o indivíduo se dedicaria ao tráfico de estupefacientes, foram realizadas as vigilâncias policiais infra, centradas na sua residência, e de onde resultou:.

No dia 12 de Setembro de 2015, o suspeito encontrou-se com Luis Pedro e Aníbal Silva. Mais tarde, com a ajuda da PSP de Coimbra, foi possível saber que estes dois se dedicavam a pequenos furtos.

No dia 14 de Setembro, João Simões Ricardo Rapaz saiu da residência no Mercedes de matrícula NN-XA-NN, acompanhado de um rapariga loura, cujo a identificação não foi possível de fazer.

Para além disso no dia 23, foi avistado um indivíduo tez escura, acompanhado de 2 indivíduos corpulentos, cada um deles com volume ao lado direito de cintura, presumindo-se que fossem portadores de arma de fogo.

Consultado a base de dados, verificou-se que o indivíduo era um conhecido traficante espanhol, da Galiza, com o nome de Juan Alberto, que se deslocava numa viatura da marca BMW 320 de matrícula XX-CC-XX.

Face aos estes fortes indícios, procedeu-se à interseção telefónica dos telefones que são por ele utilizados, com os números 91XXXXXXX , 96XXXXXXX, escutas essas que se iniciaram a 1 de Novembro de 2015 e terminaram em 12 de Dezembro de 2015, e de onde resultou:.

O João Simões Ricardo Rapaz recebeu uma chamada para o 91XXXXXXX de Juan Alberto, onde combinavam a entrega de farinha, no local do costume.

O João Simões Ricardo Rapaz telefonou para o 92XXXXXXX, e pediu que lhe entregassem o Mercedes carregado.

E que as banadas estavam prontas.

O João Simões Ricardo Rapaz recebeu uma chamada de número não identificado, onde combinava a entrega de 2 kg de neve.

Dando continuidade à investigação, procedeu-se a vigilâncias na casa do suspeito e nos locais frequentados por este, apurou-se:

Houve um transporte de um camião com matrícula OH-NN-NN , vindo de Espanha.

O suspeito seguiu atrás no Mercedes e foi descarregado num armazém em Aveiro.

Percebeu-se que o indivíduo João Alberto, é o dono do armazém e amigo de João Rapaz. No local foram avistadas três viaturas,

```
                 nomeadamente um BMW 320 de matrícula NN-CC-NN.
              Tendo em conta este cenário , foram envolvidos 30 inspectores que
                 procederam a buscas no armazém, local onde foram detectados 30
                 pacotes de cocaína.
              Foram detidos os indivíduos, tendo sido apreendidas armas e as
                 viaturas com as matrículas XX-CC-XX, XX-WC-XX e XX-CC-XX, bem como
                  3 milhões de euros em numerário.
              Foram realizadas buscas domiciliárias, que permitiram a apreensão de 6
                  milhões de euros, um quadro do pintor Picasso, DVD com imagem de
                 vídeo vigilância relativas ao alegado acidente da mulher do João
                 Simões Ricardo Rapaz e que contradizem a versão de acidente por
                 ele apresentada.
       </documentbody>
</ApplicationExample>
```

Listing A.2: Criminal Investigation Report - Application Example (.XML).

A.2 Application Example: IBM™i2 Notebook Network

The figure A.1 shows the output file of our case study after being generated by IBM i2 Notebook.

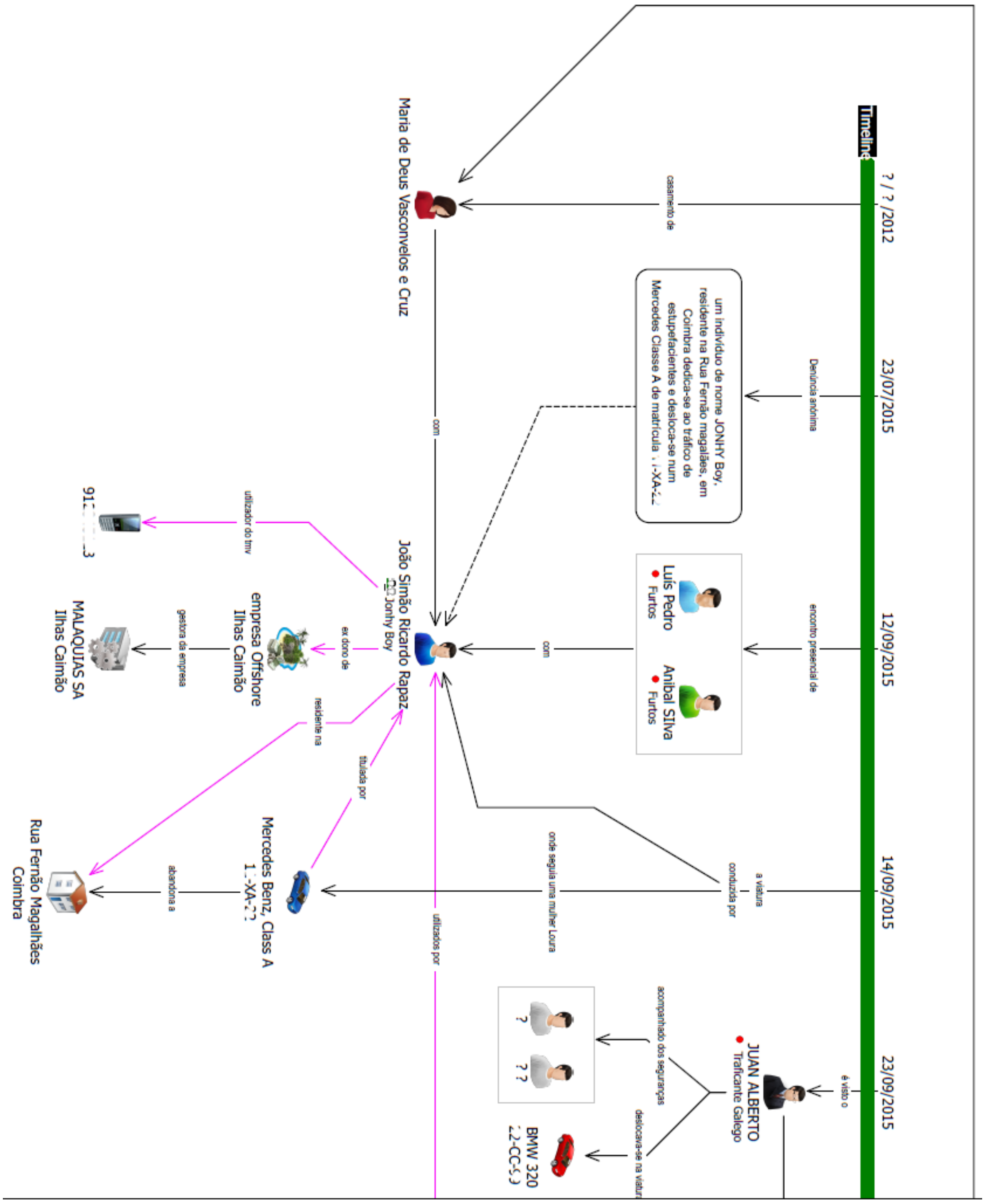
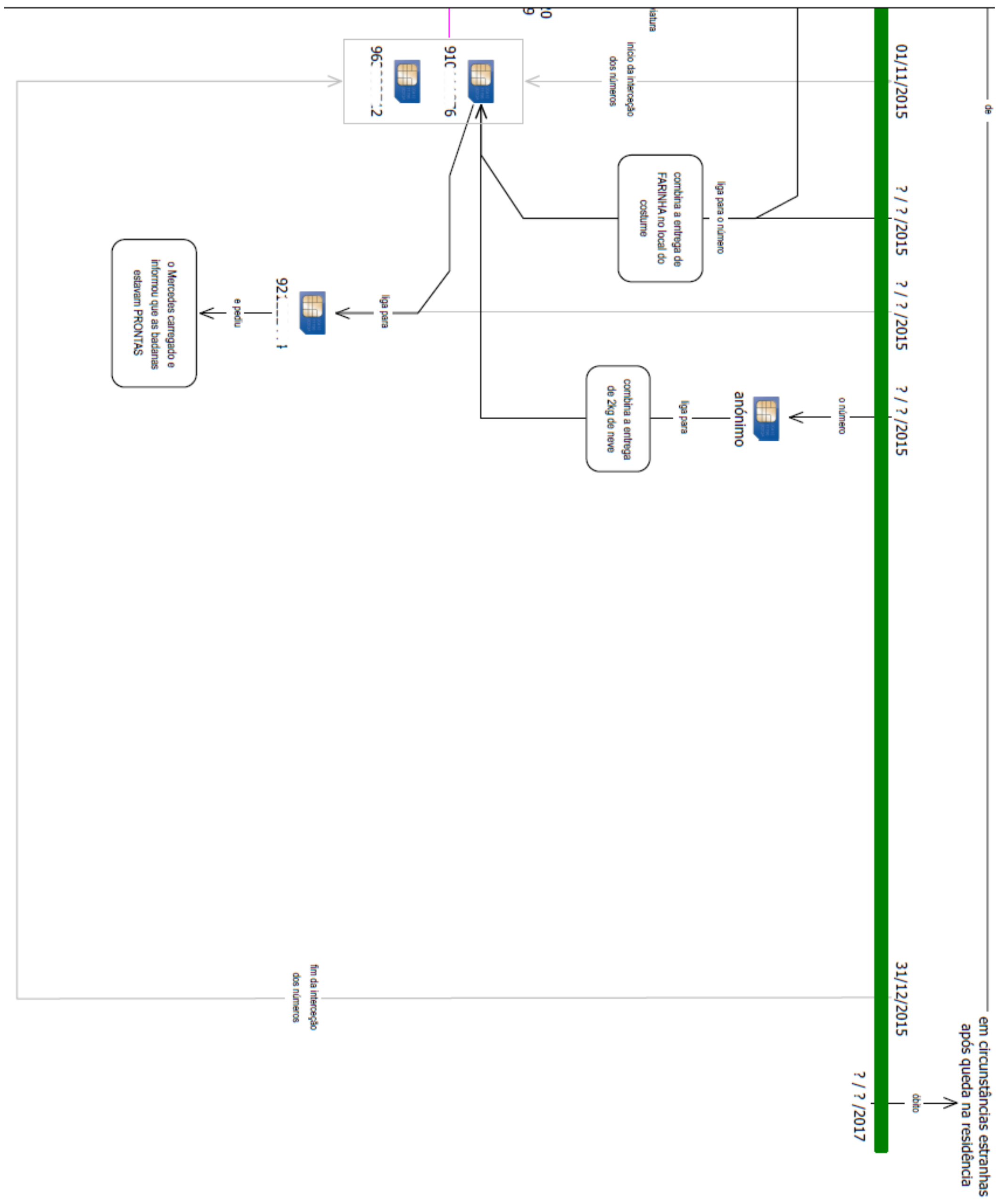


Figure A.1: IBM™i2 Notebook Output File (Page 1).



de
01/11/2015
? / ? /2015
? / ? /2015
? / ? /2015
31/12/2015
? / ? /2017
óbito
em circunstâncias estranhas após queda na residência
iatura
início da interceção dos números
910[redacted]6
96[redacted]2
liga para o número
combina a entrega de FARINHA no local do costume
o número
anónimo
liga para
combina a entrega de 2kg de neve
liga para
92[redacted]4
e pediu
o Mercedes carregado e informou que as badanas estavam PRONTAS
fim da interceção dos números

Figure A.2: IBM™i2 Notebook Output File (Page 2).

Bibliography

Adderley, R., Seidler, P., Badii, A., Tiemann, M., Neri, F., and Raffaelli, M. (2014). Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights. *The Fourth International Conference on Advances in Information Mining and Management*, 1(c):36–40. (Cited on pages 42, 45, and 63.)

Adhvaryu, N. and Balani, P. (2015). Survey : Part-Of-Speech Tagging in NLP. *International Journal of Research in Advent Technology*, 1(1):102–107. (Cited on page 19.)

Afonso, S. (2004). Árvores deitadas: Descrição do formato e descrição das opções de análise na floresta sintáctica. *Texto produzido no âmbito da Floresta Sintá (c) tica*. (Cited on page 91.)

Akerkar, R. and Joshi, M. (2003). Natural language int erface using shallow parsing. *Int. J. Comput. Sci. Appl.*, 5(3):70–90. (Cited on page 20.)

Akhgar, B., Bayerl, P. S., and Sampson, F. (2017). *Open Source Intelligence Investigation: From Strategy to Implementation*. Springer. (Cited on page 8.)

Al-Zaidy, R., Fung, B. C. M., and Youssef, A. M. (2011). Towards discovering criminal communities from textual data. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 172–177, New York, NY, USA. ACM. (Cited on pages 47 and 50.)

Albertetti, F. and Stoffel, K. (2012). From police reports to data marts: a step towards a crime analysis framework. *Int. Work. Comput. Forensics*. (Cited on pages 2, 8, 41, 45, and 63.)

Allan, J. (1996). Incremental relevance feedback for information filtering. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 270–278. (Cited on page 41.)

Alrehamy, H. H. and Walker, C. (2017). Semcluster: unsupervised automatic keyphrase extraction using affinity propagation. In *UK Workshop on Computational Intelligence*, pages 222–235. Springer. (Cited on page 24.)

Alshammari, H. Z. and Alghathbar, K. S. (2017). Clogvis: Crime data analysis and visualization tool. In *Proceedings of the Second International Conference on Advanced Wireless Information, Data, and Communication Technologies*, pages 1–7. (Cited on pages xiv, 61, and 62.)

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*. (Cited on page 55.)

Amaral, D. O. F. d. et al. (2013). O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa. *Pontifícia Universidade Católica do Rio Grande do Sul*. (Cited on page 55.)

Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39. (Cited on page 35.)

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association. (Cited on page 22.)

Arora, J. and Mishra, P. K. (2014). Analysis of stock crime using graph mining. *International Journal of Computer Applications*, 97(14). (Cited on page 61.)

Arulanandam, R., Savarimuthu, B. T. R., and Purvis, M. A. (2014). Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference - Volume 155*, AWC '14, pages 31–38, Darlinghurst, Australia, Australia. Australian Computer Society, Inc. (Cited on pages 48 and 50.)

Asher, R. E. and Moseley, C. (2018). *Atlas of the World's Languages*. Routledge. (Cited on page 16.)

Atkin, H. (2011). Criminal intelligence: Manual for analysts. *United Nations Office on Drugs and Crime (UNODC)*. (Cited on page 8.)

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, 7(1):1–16. (Cited on page 25.)

Azeez, J. and Aravindhar, D. J. (2015). Hybrid approach to crime prediction using deep learning. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1701–1710. (Cited on page 61.)

Bansal, S. K. and Kagemann, S. (2015). Integrating big data: A semantic extract-transform-load framework. *Computer*, 48(3):42–50. (Cited on page 45.)

Beamer, B., Rozovskaya, A., and Girju, R. (2008). Automatic semantic relation extraction with multiple boundary generation. In *AAAI*, pages 824–829. (Cited on page 58.)

Bick, E. (2000). The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. (Cited on pages 25 and 91.)

Bick, E. (2006a). Functional aspects in portuguese ner. In *International Workshop on Computational Processing of the Portuguese Language*, pages 80–89. Springer. (Cited on page 53.)

Bick, E. (2006b). *Functional Aspects in Portuguese NER*, pages 80–89. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on page 58.)

Bick, E. (2007). Automatic semantic role annotation for portuguese. In *Proceedings of TIL 2007-5th Workshop on Information and Human Language Technology*, pages 1713–1716. (Cited on page 56.)

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. (Cited on page 47.)

Branco, A. and Silva, J. (2006). A suite of shallow processing tools for portuguese: Lx-suite. In *Demonstrations*. (Cited on page 52.)

Branco, A. and Silva, J. (2007). Very high accuracy rule-based nominal lemmatization with a minimal lexicon. *APL XXI, Lisbon*. (Cited on page 52.)

Branco, A. H. and Silva, J. (2003). Tokenization of portuguese: resolving the hard cases. *Department of Informatics, University of Lisbon*. (Cited on page 51.)

Braz, J. (2019). *Investigacao Criminal*. Almedina Brasil. (Cited on pages 9, 10, 11, 71, and 110.)

Brewster, B., Andrews, S., Polovina, S., Hirsch, L., and Akhgar, B. (2014). Environmental scanning and knowledge representation for the detection of organised crime threats. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 8577 LNAI:275–280. (Cited on pages xiv, 42, and 45.)

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, page 152–155, USA. Association for Computational Linguistics. (Cited on page 20.)

Brown, L. (1993). *New shorter Oxford English dictionary on historical principles*. Clarendon. (Cited on page 20.)

Bruggen, R. v. and Mohanta, P. (2014). *Learning Neo4j*. Packt Publishing Ltd. (Cited on page 116.)

Bsoul, Q., Salim, J., and Zakaria, L. Q. (2016). Effect Verb Extraction on Crime Traditional Cluster. *World Appl. Sci. J.*, 34(9):1183–1189. (Cited on pages 49 and 50.)

Camara Junior, A. T. d. (2013). Processamento de linguagem natural para indexação automática semântico-ontológica. *Revista Ibero-Americana de Ciência da Informação*, 9(2):569. (Cited on pages 48 and 50.)

Cardoso, N. (2008). Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. (Cited on page 58.)

Cardoso, N. (2012). Rembrandt-a named-entity recognition framework. In *quot; In Nicoletta Calzolari; Khalid Choukri; Thierry Declerck; Mehmet U? ur Do? an; 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)*. (Cited on pages xiv, 55, and 57.)

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC*, pages 239–242. (Cited on page 52.)

Carvalho, P., Gonçalo Oliveira, H., Santos, D., Freitas, C., and Mota, C. (2008). Segundo harem: Modelo geral, novidades e avaliaçao. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*. (Cited on pages xiii, 28, and 29.)

Casanovas, P., Arraiza, J., Melero, F., González-Conejero, J., Molcho, G., and Cuadros, M. (2014). Fighting Organized Crime Through Open Source Intelligence: Regulatory Strategies of the CAPER Project. *Frontiers in Artificial Intelligence and Applications*, 271:189–198. (Cited on pages 42 and 45.)

Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013. (Cited on page 41.)

Cavanillas, J. M., Curry, E., and Wahlster, W. (2016). *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe*. Springer. (Cited on page 2.)

Chakma, K. and Das, A. (2018). A 5w1h based annotation scheme for semantic role labeling of english tweets. *Computación y Sistemas*, 22(3). (Cited on page 60.)

Chau, M., Xu, J. J., and Chen, H. (2002). Extracting meaningful entities from police narrative reports. *The University of Arizona*. (Cited on pages 46 and 50.)

Chaves, M. (2008). Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*. (Cited on pages xiv, 54, and 56.)

Chávez, J. V. and Li, X. (2011). Ontology based etl process for creation of ontological data warehouse. In *2011 8th International Conference on Electrical Engineering, Computing Science and Automatic Control*, pages 1–6. IEEE. (Not cited.)

Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., and Schroeder, J. (2003). Coplink: Managing law enforcement data and knowledge. *Communications of the ACM*, 46(1):28–34. (Cited on pages 2, 40, 45, and 63.)

Chu, X., Ilyas, I. F., Krishnan, S., and Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on Management of Data*, pages 2201–2206. ACM. (Cited on page 75.)

Chávez, J. V. and Li, X. (2011). Ontology based etl process for creation of ontological data warehouse. In *2011 8th International Conference on Electrical Engineering, Computing Science and Automatic Control*, pages 1–6. (Cited on page 45.)

Codd, E. F. (2002). A relational model of data for large shared data banks. In *Software pioneers*, pages 263–294. Springer. (Cited on pages 34 and 36.)

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297. (Cited on page 31.)

Costa, S. (2012). Saberes e práticas dos órgãos de polícia criminal na gestão da cena do crime. *A ciência na luta contra o crime: potencialidade e limites*, pages 69–97. (Cited on page 11.)

Cunningham, P. and Delany, S. J. (2007). k-nearest neighbour classifiers. *Multiple Classifier Systems*, 34(8):1–17. (Cited on page 49.)

Curran, J. R. and Clark, S. (2003). Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 164–167. (Cited on page 22.)

da Silva, J. R. M. F. (2007). *Shallow processing of Portuguese: From sentence chunking to nominal lemmatization*. PhD thesis, Master's thesis. (Cited on page 52.)

da Silva Teixeira, A. J., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P. (2008). *Computational Processing of the Portuguese Language: 8th International Conference, PROPOR 2008 Aveiro, Portugal, September 8-10, 2008, Proceedings*, volume 5190. Springer. (Cited on page 25.)

Daniel, G., Sunyé, G., and Cabot, J. (2016). Umltographdb: Mapping conceptual schemas to graph databases. In Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., and Saeki, M., editors, *Conceptual Modeling*, pages 430–444, Cham. Springer International Publishing. (Cited on page 116.)

Das, A., Ghosh, A., and Bandyopadhyay, S. (2010). Semantic role labeling for bengali using 5ws. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, pages 1–8. IEEE. (Cited on page 60.)

Das, P. and Das, A. K. (2017). A two-stage approach of named-entity recognition for crime analysis. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. (Cited on pages 49 and 50.)

Das, P., Das, A. K., Nayak, J., Pelusi, D., and Ding, W. (2019). A graph based clustering approach for relation extraction from crime data. *IEEE Access*, 7:101269–101282. (Cited on pages xiv, 62, and 63.)

Dasgupta, T., Dey, L., Saha, R., and Naskar, A. (2018). Automatic curation and visualization of crime related information from incrementally crawled multi-source news reports. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 103–107. (Cited on page 62.)

Davis, L. (1991). *Handbook of genetic algorithms*. CumInCAD. (Cited on page 41.)

de Holanda Maia, M. R. and Xexéo, G. B. (2011). Part-of-speech tagging of portuguese using hidden markov models with character language model emissions. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. (Cited on page 52.)

Dean, G., Fahsing, I. A., Glomseth, R., and Gottschalk, P. (2008). Capturing knowledge of police investigations: towards a research agenda. *Police Practice and Research: An International Journal*, 9(4):341–355. (Cited on page 2.)

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113. (Cited on page 13.)

Deng, D., Li, G., Feng, J., Duan, Y., and Gong, Z. (2015). A unified framework for approximate dictionary-based entity extraction. *The VLDB Journal*, 24(1):143–167. (Cited on page 41.)

Duran, M. S. and Aluísio, S. M. (2012). Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*, pages 1862–1867. (Cited on pages 22 and 108.)

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365. (Cited on page 31.)

Eftimov, T., Seljak, B. K., and Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488. (Cited on pages 22 and 41.)

Ejem, R. (2017). Relation extraction in police records. *Univerzita Karlova, Matematicko-fyzikální fakulta*. (Cited on pages 49 and 50.)

Ekbal, A., Mondal, S., and Bandyopadhyay, S. (2007). Pos tagging using hmm and rule-based chunking. *The Proceedings of SPSAL*, 8(1):25–28. (Cited on page 20.)

El-sappagh, S. H. A., Hamed, A., Bastawissy, E., and Ahmed, A. M. (2011). A proposed model for data warehouse ETL processes. *J. King Saud Univ. - Comput. Inf. Sci.*, 23(2):91–104. (Cited on page 15.)

Elyezjy, N. T. (2015). Investigating crimes using text mining and network analysis. *Investigating Crimes Using Text Mining and Network Analysis*. (Not cited.)

Elyezjy, N. T. and Elhaless, A. M. (2015). Investigating crimes using text mining and network analysis. *International Journal of Computer Applications*, 126(8):19–25. Published by Foundation of Computer Science (FCS), NY, USA. (Cited on pages 43 and 45.)

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics. (Cited on page 59.)

Feldman, A. (2006). Tagging Portuguese with a Spanish Tagger Using Cognates. (Cited on page 51.)

Feldman, R., Sanger, J., et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press. (Cited on page 14.)

Fernandes, I., Cardoso, H. L., and Oliveira, E. (2018). Applying deep neural networks to named entity recognition in portuguese texts. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 284–289. IEEE. (Cited on page 61.)

Figueira, H., Mendes, A., Mendes, P., and Pinto, C. (2011). O novo acordo ortográfico e os correctores automáticos. In *Lusofonia tempo de reciprocidades: actas/IX Congresso da Associação Internacional de Lusitanistas, Madeira, 4 a 9 de agosto de 2008*, pages 65–78. Ediçôes Afrontamento. (Cited on page 82.)

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370. (Cited on page 22.)

Fonseca, E. B., Chiele, G., Vanim, A., and Vieira, R. (2015a). Reconhecimento de entidades nomeadas para o português usando o opennlp. *Anais do ENIAC 2015, 2015, Brasil.* (Cited on page 92.)

Fonseca, E. R. and Rosa, J. L. G. (2012). An architecture for semantic role labeling on portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 204–209. Springer. (Cited on page 57.)

Fonseca, E. R. and Rosa, J. L. G. (2013). A two-step convolutional neural network approach for semantic role labeling. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. (Cited on pages xvii and 24.)

Fonseca, E. R. and Rosa, J. L. G. (2013). A two-step convolutional neural network approach for semantic role labeling. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. (Cited on pages xviii, 57, 108, and 109.)

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015b). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2. (Cited on page 52.)

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. (Cited on pages 20 and 31.)

Frické, M. (2019). The knowledge pyramid: the dikw hierarchy. *KO KNOWLEDGE ORGANIZATION*, 46(1):33–46. (Cited on pages 12 and 13.)

Furtado, V., Ayres, L., Oliveira, M. D., Vasconcelos, E., Caminha, C., Orleans, J. D., and Belchior, M. (2010). Collective intelligence in law enforcement – The WikiCrimes system. *Inf. Sci. (Ny).*, 180(1):4–17. (Cited on page 47.)

Galante, L. M. T. B. (2016). A inteligência na prevenção criminal: caso de estudo dos furtos em interior de residência na área do posto territorial da gnr da ericeira. Master's thesis, Instituto Superior de Ciências Policiais e Segurança Interna. (Cited on page 3.)

Gamallo, P. and Garcia, M. (2015). Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*, pages 711–722. Springer. (Cited on page 59.)

Gamallo, P. and Garcia, M. (2017). Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, 9(1):19–28. (Cited on page 59.)

Gangemi, A. (2005). Ontology design patterns for semantic web content. In Gil, Y., Motta, E., Benjamins, V. R., and Musen, M. A., editors, *The Semantic Web – ISWC 2005*, pages 262–276, Berlin, Heidelberg. Springer Berlin Heidelberg. (Not cited.)

Garcia, M. and Gamallo, P. (2011). Evaluating various linguistic features on semantic relation extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 721–726. (Cited on page 58.)

Garcia, M., Gamallo, P., Gayo, I., and Cruz, M. A. P. (2014). Pos-tagging the web in portuguese. national varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53:95–101. (Cited on page 52.)

Geepalla, E. and Abuhamoud, N. (2019). Analysis cdr for crime investigation using graph-based method (neo4j). In *International Conference on Technical Sciences (ICST2019)*, volume 4, page 06. (Cited on pages xiv and 62.)

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. (2019). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? (Cited on page 75.)

Gianola, L. (2020). *Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique*. PhD thesis, Université de Cergy-Pontoise. (Cited on pages 50, 63, and 65.)

Gleick, J. and Calil, A. (2013). *A informação: Uma história, uma teoria, uma enxurrada*. Companhia das Letras. (Cited on page 2.)

Golomb, S. W. and Taylor, H. (1984). Constructions and properties of costas arrays. *Proceedings of the IEEE*, 72(9):1143–1163. (Cited on pages xiii and 33.)

Gonçalves, T., Silva, C., Quaresma, P., and Vieira, R. (2006). Analysing part-of-speech for portuguese text classification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 551–562. Springer. (Cited on page 52.)

Gorawski, M. and Gorawska, A. (2014). Research on the stream etl process. In *International Conference: Beyond Databases, Architectures and Structures*, pages 61–71. Springer. (Cited on page 45.)

Gottschalk, P. (2006). *Knowledge Management Systems in Law Enforcement: Technologies and Techniques: Technologies and Techniques*. IGI Global. (Not cited.)

Gribkovskaia, I., Halskau Sr, Ø., and Laporte, G. (2007). The bridges of königsberg—a historical perspective. *Networks: An International Journal*, 49(3):199–203. (Cited on pages xiii and 33.)

Grzymala-Busse, J. W. (1993). Selected algorithms of machine learning from examples. *Fundam. Inform.*, 18:193–207. (Cited on page 41.)

Haerder, T. and Reuter, A. (1983). Principles of transaction-oriented database recovery. *ACM computing surveys (CSUR)*, 15(4):287–317. (Cited on page 35.)

Hajic, J., Vidová-Hladká, B., and Pajas, P. (2001). The prague dependency treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 105–114. (Cited on page 21.)

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., and Gipp, B. (2018). Giveme5w: main event retrieval from news articles by extraction of the five journalistic w questions. In *International Conference on Information*, pages 356–366. Springer. (Cited on pages xiv and 60.)

Helbich, M., Hagenauer, J., Leitner, M., and Edwards, R. (2013). Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. *Cartography and Geographic Information Science*, 40(4):326–336. (Cited on pages 48 and 50.)

Hirschman, L. (1998). The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech and Language*, 12(4):281 – 305. (Cited on page 28.)

Hossain, M. S., Butler, P., Boedihardjo, A. P., Ramakrishnan, N., and Tech, V. (2012). Storytelling in Entity Networks to Support Intelligence Analysts. In *Conf. Knowl. Discov. Data Min.* (Cited on pages 42 and 45.)

Hosseinkhani, J., Chaprut, S., and Taherdoost, H. (2012). Criminal network mining by web structure and content mining. *Advances in Remote Sensing, Finite Differences and Information Security. In Proceedings of the 11th WSEAS International Conference on Information Security and Privacy (ISP '12)*, pages 210–215. (Cited on pages 42 and 45.)

Hu, H., Wen, Y., Chua, T.-S., and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, 2:652–687. (Cited on pages 13 and 14.)

Hull, R. and King, R. (1987). Semantic database modeling: Survey, applications, and research issues. *ACM Computing Surveys (CSUR)*, 19(3):201–260. (Cited on page 34.)

IACA, S., Elder, J., IACA, S., Bruce, C. W., Santos, R. B., Rodriguez, E., Los Angeles County, C., Steiner, F., Police, A. F., and Wyckoff, L. (2014). Definition and types of crime analysis. *Citeseer*. (Cited on page 8.)

Indurkhya, N. and Damerau, F. J. (2010). *Handbook of natural language processing*, volume 2. CRC Press. (Cited on pages xiii, 16, and 17.)

Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I., and Wijayasiri, A. (2015). Crime analytics: Analysis of crimes through newspaper articles. In *2015 Moratuwa Engineering Research Conference (MERCon)*, pages 277–282. (Not cited.)

Jiang, L., Cai, H., and Xu, B. (2010). A domain ontology approach in the etl process of data warehousing. In *2010 IEEE 7th International Conference on E-Business Engineering*, pages 30–35. (Cited on page 45.)

Jungnickel, D. and Jungnickel, D. (2005). *Graphs, networks and algorithms*. Springer. (Cited on page 33.)

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition. (Cited on pages 16, 17, 20, and 22.)

Kakish, K. and Kraft, T. A. (2012). Etl evolution for real-time data warehousing. In *Proceedings of the Conference on Information Systems Applied Research ISSN*, volume 2167, page 1508. (Cited on pages 14 and 15.)

Kanimozhi, K. and Venkatesan, M. (2015). Unstructured data analysis-a survey. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(3):223–225. (Cited on page 13.)

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., Alam, M., Shiraz, M., and Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, 2014. (Cited on page 14.)

Khan, W., Daud, A., Nasir, J. A., and Amjad, T. (2016). A survey on the state-of-the-art machine learning models in the context of nlp. *Kuwait journal of Science*, 43(4). (Cited on page 17.)

Kind, S. S. (1994). Crime investigation and the criminal trial: a three chapter paradigm of evidence. *Journal-Forensic Science Society Harrogate*, 34:155–155. (Cited on page 8.)

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525. (Cited on page 51.)

Knap, T., Kukhar, M., Macháč, B., Škoda, P., Tomeš, J., and Vojt, J. (2014). Unifiedviews: An etl framework for sustainable rdf data processing. In *European Semantic Web Conference*, pages 379–383. Springer. (Cited on page 45.)

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480. (Cited on page 48.)

Köpcke, H. and Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197–210. (Cited on page 41.)

Krenek, J., Kuca, K., Krejcar, O., Maresova, P., Sobeslav, V., and Blazek, P. (2014). Artificial neural network tools for computerised data modeling and processing. In *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 255–260. IEEE. (Cited on page 49.)

Ku, C. H., Iriberri, A., and Leroy, G. (2008). Crime Information Extraction from Police and Witness Narrative Reports. In *2008 IEEE Conference on Technologies for Homeland Security*, pages 193–198, Waltham, MA, USA. IEEE. (Cited on pages 46 and 50.)

Lafferty, J., McCallum, A., and Pereira, F. C. (2014). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289. (Cited on page 31.)

Lagos, N., Segond, F., Castellani, S., and O'Neill, J. (2010). Event extraction for legal case building and reasoning. In *International Conference on Intelligent Information Processing*, pages 92–101. Springer. (Cited on page 58.)

Leech, G. N. (1992). 100 million words of english: the british national corpus (bnc). *Seoul National University Language Education Center*. (Cited on page 25.)

Lezcano, L., Sánchez-Alonso, S., and Roa-Valverde, A. J. (2013). A survey on the exchange of linguistic resources. *Program*. (Cited on page 26.)

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., and Zhang, C. (2019). Cleanml: A benchmark for joint data cleaning and machine learning [experiments and analysis]. *arXiv preprint arXiv:1904.09483*. (Cited on page 75.)

Li, X. and Mao, Y. (2015). Real-time data etl framework for big real-time data analysis. In *2015 IEEE International Conference on Information and Automation*, pages 1289–1294. IEEE. (Cited on page 45.)

Liddy, E. D. (2001). *Natural language processing*. Encyclopedia of Library and Information Science. (Cited on page 17.)

Liew, A. (2007). Understanding data, information, knowledge and their inter-relationships. *Journal of knowledge management practice*, 8(2):1–16. (Cited on page 12.)

Liew, A. (2013). Dikiw: Data, information, knowledge, intelligence, wisdom and their interrelationships. *Business Management Dynamics*, 2(10):49. (Cited on page 13.)

Lockard, C., Dong, X. L., Einolghozati, A., and Shiralkar, P. (2018). Ceres: Distantly supervised relation extraction from the semi-structured web. *arXiv preprint arXiv:1804.04635*. (Cited on page 30.)

López, R. and Pardo, T. A. (2015). Experiments on sentence boundary detection in user-generated web content. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 227–237. Springer. (Cited on page 51.)

Lydia, G. and Seetha, H. (2019). Deep Learning Based Crime Investigation Framework. *International Journal of Scientific & Technology Research*, 8(11). (Cited on page 61.)

M. Ramirez-Alcocer, U., Tello-Leal, E., and A. Mata-Torres, J. (2019). Predicting Incidents of Crime Through LSTM Neural Networks in Smart City Domain. In *The Eighth International Conference on Smart Cities, Systems, Devices and Technologies*, page 32 to 37, Nice, France. (Cited on page 61.)

Majeed, F., Mahmood, M. S., and Iqbal, M. (2010). Efficient data streams processing in the real time data warehouse. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 57–61. (Cited on page 44.)

Maltby, D. (2011). Big data analytics. In *74th Annual Meeting of the Association for Information Science and Technology (ASIST)*, pages 1–6. (Cited on page 14.)

Mamede, N. J., Baptista, J., Trancoso, I., and das Graças Volpe Nunes, M., editors (2003). *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, volume 2721 of *Lecture Notes in Computer Science*. Springer. (Cited on page 53.)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition. Technical report, and productivity. Technical report, McKinsey Global Institute. (Cited on page 14.)

Maršík, J. and Bojar, O. (2012). Trtok: a fast and trainable tokenizer for natural languages. *The Prague Bulletin of Mathematical Linguistics*, 98:75–85. (Cited on page 51.)

Martin-Rodilla, P., Hattori, M. L., and Gonzalez-Perez, C. (2019). Assisting forensic identification through unsupervised information extraction of free text autopsy reports: The disappearances cases during the brazilian military dictatorship. *Information*, 10(7):231. (Cited on pages 49 and 50.)

Martins, B., Silva, M., and Chaves, M. (2007). O sistema cage no harem-reconhecimento de entidades geográficas em textos em língua portuguesa. *quot; In Diana Santos; Nuno Cardoso (ed) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área Linguateca 2007*. (Cited on pages xiv, 53, and 54.)

Mase, H. and Tsuji, H. (1999). Text summarizing method and system. US Patent 5,978,820. (Cited on page 42.)

Mata, F., Torres-ruiz, M., Guzmán, G., Quintero, R., Zagal-flores, R., Moreno-ibarra, M., and Loza, E. (2016). A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes : A Case Study of Mexico City. *Mobile Information Systems*, 2016:11. (Cited on pages 43, 45, and 63.)

Matos, J. (2010). *Gramática moderna da língua portuguesa: para o conhecimento e aperfeiçoamento dos aspetos fundamentais da estrutura e funcionamento da língua*. Gramática: moderna da língua portuguesa. Escolar Editora. (Cited on page 80.)

McCallum, A., Freitag, D., and Pereira, F. C. (2000). Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598. (Cited on page 31.)

McColl, R. C., Ediger, D., Poovey, J., Campbell, D., and Bader, D. A. (2014). A performance evaluation of open source graph databases. In *Proceedings of the first workshop on Parallel programming for analytics applications*, pages 11–18. (Cited on page 35.)

McCracken, N., Ozgencil, N. E., and Symonenko, S. (2006). Combining techniques for event extraction in summary reports. In *AAAI 2006 Workshop Event Extraction and Synthesis*, pages 7–11. (Cited on pages 59 and 109.)

McNamee, P., Mayfield, J. C., and Piatko, C. D. (2011). Processing named entities in text. *Johns Hopkins APL Technical Digest*, 30(1):31–40. (Cited on pages xvii, 21, and 28.)

Mendes, A., Amaro, R., and Bacelar, M. F. (2003). Morphological Tagging of a Spoken Portuguese Corpus Using Available Resources. (Cited on page 51.)

Miller, J. J. (2013). Graph database applications and concepts with neo4j. In *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*. (Cited on page 34.)

Mírian Bruckschen, J. G., Souza, R.-n. V., and Rigo, S. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 14: Sistem, page 436. Linguateca. (Cited on page 58.)

Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877. (Cited on page 30.)

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*. (Cited on pages xiv, 54, and 55.)

Mota, C. and Santos, D. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter : Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM, page 436. Linguateca. (Cited on page 57.)

Murphy, K. P. et al. (2006). Naive bayes classifiers. *University of British Columbia*, 18. (Cited on page 49.)

Nadeau, D. and Sekine, S. (2006). A survey of named entity recognition and classification. (Cited on pages 21 and 22.)

Nath, R. P. D., Hose, K., Pedersen, T. B., and Romero, O. (2017). Setl: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, 68:17–43. (Cited on page 45.)

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA). (Cited on page 20.)

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135. (Cited on page 53.)

Nothman, J., Honnibal, M., Hachey, B., and Curran, J. R. (2012). Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 228–232. Association for Computational Linguistics. (Cited on page 59.)

Nouh, M., Nurse, J. R. C., and Goldsmith, M. (2016). Towards Designing a Multipurpose Cybercrime Intelligence Framework. *2016 European Intelligence and Security Informatics Conference*, pages 60–67. (Cited on pages 43 and 45.)

Oliveira, H. G., Costa, H., and Gomes, P. (2010). Extracçao de conhecimento léxico-semântico a partir de resumos da wikipédia. *Actas do II Simpósio de Informática*. (Cited on page 58.)

Ong, T., Pradhananga, R., Holve, E., and Kahn, M. G. (2017). eGEMs A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. *J. Electron. Heal. Data Methods*, 5(1). (Cited on page 15.)

Onnoom, B., Chiewchanwattana, S., Sunat, K., and Wichiennit, N. (2015). An ontology framework for recommendation about a crime scene investigation. *14th International Symposium on Communications and Information Technologies, ISCIT 2014*, pages 176–180. (Cited on pages 42, 45, and 63.)

Osborne, M. (2000). Shallow Parsing as Part-of-Speech Tagging. In *Proc. CoNLL-2000 LLL-2000*, pages 145–147. (Cited on page 20.)

Otero, P. G. and González, I. (2012). Deppattern: a multilingual dependency parser. In *International Conference on Computational Processing of the Portuguese Language (PROPOR 2012), Coimbra, Portugal*, pages 659–670. Citeseer. (Cited on page 53.)

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2018). Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4):431–448. (Cited on page 2.)

Palmer, D. D. (2000). Tokenisation and sentence segmentation. *Handbook of natural language processing*, pages 11–35. (Cited on page 18.)

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106. (Cited on page 22.)

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103. (Cited on page 22.)

Pereira, C. (2013). Análise criminal e sistemas de informação. Master's thesis, Instituto de Estudos Superiores Militares. (Cited on page 11.)

Perez-Ortiz, J. A. and Forcada, M. L. (2001). Part-of-speech tagging with recurrent neural networks. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, pages 1588–1592. IEEE. (Cited on page 20.)

Pinheiro, V., Furtado, V., Pequeno, T., Nogueira, D., and Aplicada, I. (2010). Natural Language Processing Based on Semantic Inferentialism for Extracting Crime Information from Text. (Cited on pages 47 and 50.)

Pinheiro, V., Pequeno, T., Furtado, V., Assunção, T., and Freitas, E. (2008). Sim: um modelo semântico-inferencialista para sistemas de linguagem natural. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, pages 353–358. ACM. (Cited on page 47.)

Pinheiro, V., Pequeno, T., Furtado, V., and Nogueira, D. (2009). Semantic inferentialist analyser: Um analisador semântico de sentenças em linguagem natural. In *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology. STIL, Brasil*. (Cited on page 47.)

Pires, A. R. O. (2017). Named entity extraction from portuguese web text. Master's thesis, Faculdade de Engenharia da Universidade do Porto. (Cited on page 56.)

Pirovani, J. P. and de Oliveira, E. (2018). Portuguese named entity recognition using conditional random fields and local grammars. In *LREC*. (Cited on page 56.)

Pirovani, J. P. C. and de Oliveira, E. S. (2015). Extração de nomes de pessoas em textos em português: uma abordagem usando gramáticas locais. In *Computer on the Beach*, pages 1–10. (Cited on page 56.)

Poelmans, J., Elzinga, P., Neznanov, A. A., Dedene, G., Viaene, S., and Kuznetsov, S. O. (2012). Human-centered text mining: A new software system. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7377 LNAI:258–272. (Cited on page 41.)

Prins, R. (2004). Beyond n in n-gram tagging. In *Proceedings of the ACL Student Research Workshop*, pages 61–66. (Cited on page 20.)

Qazi, N., Zhang, L., Blomqvist, E., Stoffel, F., Aichroth, P., and Weigel, C. (2017). Applying data science to criminal intelligence analysis. Technical report, Middlesex University London. (Cited on page 2.)

Quaresma, P., Nogueira, V. B., Raiyani, K., and Bayot, R. (2019). Event extraction and representation: A case study for the portuguese language. *Information*, 10(6):205. (Cited on page 57.)

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. (Cited on page 31.)

Rahem, K. R. and Omar, N. (2015). Rule-based named entity recognition for drug-related crime news documents. *Journal of Theoretical & Applied Information Technology*, 77(2). (Cited on pages 49 and 50.)

Raiyani, K., Gonçalves, T., Quaresma, P., and Nogueira, V. B. (2019). Automated event extraction model for linked portuguese documents. In *Text2Story@ ECIR*, pages 13–20. (Cited on page 59.)

Randolph, J. J. (2009). A guide to writing the dissertation literature review. *Practical assessment, research & evaluation*, 14(13):1–13. (Cited on page 40.)

Ribaux, O., Girod, A., Walsh, S. J., Margot, P., Mizrahi, S., and Clivaz, V. (2003). Forensic intelligence and crime analysis. *Law, Probability and Risk*, 2(1):47–60. (Not cited.)

Robertson, S. (2000). Evaluation in information retrieval. In *European Summer School on Information Retrieval*, pages 81–92. Springer. (Cited on page 29.)

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph databases*. " O'Reilly Media, Inc.". (Cited on pages xiii, 34, 35, 36, and 116.)

Rodrigues, M., Dias, G. P., and Teixeira, A. (2010). Knowledge extraction from minutes of portuguese municipalities meetings. *Proc. of the FALA*. (Cited on page 58.)

Rodrigues, R. and Gomes, P. (2015). Rapport—a portuguese question-answering system. In *Portuguese Conference on Artificial Intelligence*, pages 771–782. Springer. (Not cited.)

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2014). Lemport: a high-accuracy cross-platform lemmatizer for portuguese. In *OASIcs-OpenAccess Series in Informatics*, volume 38. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. (Cited on pages 52 and 85.)

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2018). Nlpport: A pipeline for portuguese nlp (short paper). In *OASIcs-OpenAccess Series in Informatics*, volume 62. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. (Cited on pages 51, 53, 56, 81, 83, and 93.)

Rodriguez, M. A. and Neubauer, P. (2010). Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology*, 36(6):35–41. (Cited on page 34.)

Rodriguez, M. A. and Neubauer, P. (2012). The graph traversal pattern. In *Graph data management: Techniques and applications*, pages 29–46. IGI Global. (Cited on page 34.)

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386. (Cited on page 32.)

Ruohonen, K. (2013). Graph theory. tampereen teknillinen yliopisto. originally titled graafiteoria, lecture notes translated by tamminen, j., lee, k. *C. and Piché, R.* (Cited on page 33.)

Russell, S. and Norvig, P. (2002). *Artificial intelligence: a modern approach*. Prentice Hall. (Cited on page 30.)

Sanderson, M. (2010). Manning christopher d., raghavan prabhakar, schütze hinrich, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi+ 482 pages. *Natural Language Engineering*, 16(1):100–103. (Cited on page 19.)

Santos, A. P., Ramos, C., and Marques, N. C. (2012). Extraçao de relaçoes em tıtulos de notıcias desportivas. In *INFORUM*. (Cited on page 58.)

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)*. (Cited on page 28.)

Santos, R. P. T. d. (2014). Automatic semantic role labeling for european portuguese. Master's thesis, University of Algarve. (Cited on page 57.)

Sarawagi, S. et al. (2008). Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377. (Cited on page 23.)

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. (Cited on page 50.)

Schaffer, J. (2015). What not to multiply without necessity. *Australasian Journal of Philosophy*, 93(4):644–664. (Cited on page 70.)

Schraagen, M. (2017). Evaluation of Named Entity Recognition in Dutch online criminal complaints. *Comput. Linguist. Netherlands J.*, 7:3–15. (Cited on pages 49 and 50.)

Seidler12, P., Badii, A., and Adderley, R. (2012). Computational modelling for law enforcement intelligence analysis. *University of Surrey, Institute of Advance Studies*. (Cited on page 2.)

Sena, C. F. L., Glauber, R., and Claro, D. B. (2017). Inference approach to enhance a portuguese open information extraction. In *Proceedings of the 19th International Conference on Enterprise Information Systems*, volume 1, pages 442–451. (Cited on page 59.)

Sequeira, J., Gonçalves, T., and Quaresma, P. (2012). Semantic role labeling for portuguese–a preliminary approach–. In *International Conference on Computational Processing of the Portuguese Language*, pages 193–203. Springer. (Cited on page 56.)

Shabat, H. A. and Omar, N. (2015). Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East J. Sci. Res.*, 23(6):1215–1221. (Cited on pages 48 and 50.)

Sharma, S. (2017). Activation functions in neural networks. *Towards Data Science*, 6. (Cited on page 32.)

Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*. (Cited on page 31.)

Sheela, J. and Vadivel, A. (2016). Criminal event detection and classification in web documents using ann classifier. *Int. J. Signal Process. Syst.*, 4(5):382–388. (Cited on page 49.)

Siaw, N. H., Kulathuramaiyer, N., and Labadin, J. (2013). Natural language semantic event extraction pipeline. In *4th International Conference on Computing and Informatics (ICOCI 2013)*. (Cited on page 59.)

Silla, C. N. and Kaestner, C. A. A. (2004). An analysis of sentence boundary detection systems for english and portuguese documents. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 135–141, Berlin, Heidelberg. Springer Berlin Heidelberg. (Cited on pages 50 and 108.)

Simões, A. and Almeida, J. J. (2001). jspell. pm: um módulo de análise morfológica para uso em processamento de linguagem natural. *Associação Portuguesa de Linguística (APL)*. (Cited on page 52.)

Sinclair, J. (2005). Corpus and text-basic principles. *Developing linguistic corpora: A guide to good practice*, pages 1–16. (Cited on page 25.)

Singh, S. (2018). Natural language processing for information extraction. (Cited on page 57.)

Skoutas, D. and Simitsis, A. (2007). Ontology-based conceptual design of etl processes for both structured and semi-structured data. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(4):1–24. (Cited on page 44.)

Song, J., Bao, Y., and Shi, J. (2010). A triggering and scheduling approach for etl in a real-time data warehouse. In *2010 10th IEEE International Conference on Computer and Information Technology*, pages 91–98. (Cited on page 44.)

Souza, E. N. P. and Claro, D. B. (2014). Extração de relações utilizando features diferenciadas para português. *Linguamática*, 6(2):57–65. (Cited on page 58.)

Souza, F., Nogueira, R. F., and de Alencar Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649. (Cited on page 61.)

Sowa, J. F. (2008). Chapter 5 conceptual graphs. In [van Harmelen], F., Lifschitz, V., and Porter, B., editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 213 – 237. Elsevier. (Cited on page 42.)

Specia, L. and Motta, E. (2006). A hybrid approach for extracting semantic relations from texts. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 57–64, Sydney, Australia. Association for Computational Linguistics. (Cited on page 106.)

Stampouli, D., Roberts, M., Powell, G., and L??pez, T. S. (2011). Implementation of a police intelligence analysis framework. *International Journal of Security and its Applications*, 5(4):13–22. (Cited on pages xiii, 41, and 45.)

Stasko, J., Görg, C., Liu, Z., and Singhal, K. (2007). Jigsaw: Supporting investigative analysis through interactive visualization. *VAST IEEE Symposium on Visual Analytics Science and Technology 2007, Proceedings*, 1(March):131–138. (Cited on pages 41 and 45.)

Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis*. Sebtel Press. (Cited on page 32.)

Suchanek, F. M., Ifrim, G., and Weikum, G. (2006). Leila: Learning to extract information by linguistic analysis. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 18–25. (Cited on page 58.)

Sunkara, V. K. M. (2019). A data driven approach to identify journalistic 5ws from text documents. Master's thesis, University of Nebraska. (Cited on pages xiii and 21.)

Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification*, pages 207–235. Springer. (Cited on page 49.)

Taba, L. S. and de Medeiros Caseli, H. (2014). Automatic semantic relation extraction from portuguese texts. In *LREC*, pages 2739–2746. (Cited on page 58.)

Tanev, H., Piskorski, J., and Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *International Conference on Application of Natural Language to Information Systems*, pages 207–218. Springer. (Cited on page 60.)

Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., and Steinberger, R. (2009). Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamática*, 1(2):55–66. (Cited on page 60.)

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer. (Cited on page 21.)

Technology, A. I., Asharef, M., Omar, N., and Albared, M. (2012). Arabic Named Entity Recognition in Crime. *J. Theor. Appl. Inf. Technol.*, 44(1):1–6. (Cited on pages 47 and 50.)

Van Campenhout, J. and Cover, T. (1981). Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, 27(4):483–489. (Cited on page 31.)

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the simple event model (sem). *Journal of Web Semantics*, 9(2):128–136. (Cited on page 116.)

Vassiliadis, P. (2003). Extraction, transformation, and loading. (Cited on pages 14 and 15.)

Vassiliadis, P., Simitsis, A., and Baikousi, E. (2009). A taxonomy of etl activities. In *Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP*, DOLAP '09, pages 25–32, New York, NY, USA. ACM. (Cited on page 14.)

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference*, pages 1–6. (Cited on pages 34 and 35.)

Vijayarani, S. and Janani, R. (2016). Text mining: open source tokenization tools–an analysis. *Advanced Computational Intelligence*, 3(1):37–47. (Cited on page 51.)

Wall, D. S. (2018). How big data feeds big crime. *Global History: A Journal of Contemporary World Affairs*. (Cited on page 73.)

Wang, W. (2012a). Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web*, pages 197–202. (Cited on page 60.)

Wang, W. (2012b). Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 197–202, New York, NY, USA. Association for Computing Machinery. (Cited on page 109.)

Wang, W., Zhao, D., and Wang, D. (2010). Chinese news event 5w1h elements extraction using semantic role labeling. In *2010 Third International Symposium on Information Processing*, pages 484–489. IEEE. (Cited on page 60.)

Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer. (Cited on pages 47 and 50.)

Wen, Y., Fan, C., Chen, G., Chen, X., and Chen, M. (2019). A survey on named entity recognition. In *International Conference in Communications, Signal Processing, and Systems*, pages 1803–1810. Springer. (Cited on page 22.)

Weston, J. and Karlen, M. (2011). Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12:2493–2537. (Cited on page 19.)

Wiedemann, G., Yimam, S. M., and Biemann, C. (2018). A multilingual information extraction pipeline for investigative journalism. *arXiv preprint arXiv:1809.00221*. (Cited on pages 44, 45, 63, and 72.)

Wijeratne, S., Doran, D., Sheth, A., and Dustin, J. L. (2015). Analyzing the social media footprint of street gangs. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 91–96. (Cited on pages 43 and 45.)

Wunderwald, M. (2011). Newsx event extraction from news articles. *Diplom, Dresden University of Technology*. (Cited on page 9.)

Xu, J. J. and Chen, H. (2005). CrimeNet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems*, 23(2):201–226. (Not cited.)

Xu, W., Yuan, C., Li, W., Wu, M., and Wong, K.-F. (2006). Building document graphs for multiple news articles summarization: An event-based approach. In *International Conference on Computer Processing of Oriental Languages*, pages 181–188. Springer. (Cited on page 59.)

Yang, B. and Mitchell, T. (2016). Joint extraction of events and entities within a document context. *arXiv preprint arXiv:1609.03632*. (Cited on page 59.)

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52. (Cited on page 19.)

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480. (Cited on page 22.)

Zhou, W., Zhang, Y., Su, X., Li, Y., and Liu, Z. (2016). Semantic role labeling based event argument identification. *International Journal of Database Theory and Application*, 9(6):93–102. (Cited on pages 59 and 109.)

Zilio, L., Wilkens, R., and Fairon, C. (2018). Passport: A dependency parsing model for portuguese. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 479–489, Cham. Springer International Publishing. (Cited on page 53.)

Zins, C. (2007). Conceptual Approaches for Defining Data , Information ,. *Journal of the American Society for Information Science and Technology*, 58(January):479–493. (Cited on page 12.)

UNIVERSIDADE DE ÉVORA
INSTITUTO DE INVESTIGAÇÃO
E FORMAÇÃO AVANÇADA

Contactos:
Universidade de Évora
Instituto de Investigação e Formação Avançada — IIFA
Palácio do Vimioso | Largo Marquês de Marialva, Apart. 94
7002 - 554 Évora | Portugal
Tel: (+351) 266 706 581
Fax: (+351) 266 744 677
email: iifa@uevora.pt