



Article

A Graph Database Representation of Portuguese Criminal-Related Documents

Gonçalo Carnaz ^{1,*} , Vitor Beires Nogueira ¹  and Mário Antunes ^{2,3} ¹ Informatics Department, University of Évora, 7002-554 Évora, Portugal; vbn@uevora.pt² Computer Science and Communication Research Centre (CIIC), School of Technology and Management, Polytechnic of Leiria, 2411-901 Leiria, Portugal; mario.antunes@ipleiria.pt³ INESC TEC, CRACS, 4200-465 Porto, Portugal

* Correspondence: d34707@alunos.uevora.pt

Abstract: Organizations have been challenged by the need to process an increasing amount of data, both structured and unstructured, retrieved from heterogeneous sources. Criminal investigation police are among these organizations, as they have to manually process a vast number of criminal reports, news articles related to crimes, occurrence and evidence reports, and other unstructured documents. Automatic extraction and representation of data and knowledge in such documents is an essential task to reduce the manual analysis burden and to automate the discovering of names and entities relationships that may exist in a case. This paper presents *SEMCrime*, a framework used to extract and classify named-entities and relations in Portuguese criminal reports and documents, and represent the data retrieved into a graph database. A 5WH1 (Who, What, Why, Where, When, and How) information extraction method was applied, and a graph database representation was used to store and visualize the relations extracted from the documents. Promising results were obtained with a prototype developed to evaluate the framework, namely a name-entity recognition with an F-Measure of 0.73, and a 5W1H information extraction performance with an F-Measure of 0.65.



Citation: Carnaz, G.; Nogueira, V.B.; Antunes, M. A Graph Database Representation of Portuguese Criminal-Related Documents. *Informatics* **2021**, *8*, 37. <https://doi.org/10.3390/informatics8020037>

Academic Editor: Antony Bryant

Received: 29 April 2021

Accepted: 29 May 2021

Published: 4 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: knowledge representation; graph databases; natural language processing; criminal-related documents; cybersecurity; criminal domain, police reports

1. Introduction

Data abounds in organizations, as a result of their massive online presence and the corresponding information sharing and storage. IDC (<https://www.idc.com/> (Accessed on: 1 June 2021).) reports that data volume increased by a factor of 300 (from 130 to 40,000 exabytes) in the period of 2005 to 2020, which has forced companies to adopt innovative mechanisms to deal with data representation, and storage. In the same direction, Statista website (<https://www.statista.com/statistics/871513/worldwide-data-created/> (Accessed on: 1 June 2021).) refers to an increasing of data that is created, captured, and consumed through the WWW (World Wide Web) from 2010 to 2024, which ascends to 149 zettabytes. The increasing amount of data and the different formats that are used, such as relational databases or text documents, brings challenges related to data heterogeneity and the corresponding computational needs, with the purpose of achieving data homogenization. Regarding textual documents, they are written in different languages and grammar, which brings specific needs concerning data processing and representation. However, emergent and more robust computational methods, such as Natural Language Processing (NLP), are being used to deal with the automatic processing of documents [1,2], which facilitates and simplifies the understanding of written texts for further knowledge representation. The use of NLP is an asset for organizations, and has the potential to be implemented in several sectors, such as academic research, industrial, government, and sports [3].

Text documents, photos, videos, e-mail messages, and social media activity are examples of pieces of information that can be collected in seized computer systems and used as evidence in criminal investigations. Evidences are enumerated in police reports or in online newspapers, highlighting the police departments need for computational methods to process data and information flows. Moreover, data representation should also comprise past and ongoing criminal investigations usually stored in databases and within different formats, which together with news published in newspapers, is able to produce relevant information and knowledge.

To have a perspective of the amount of data that criminal police departments deal with, we have analyzed the Portuguese Internal Security Report, regarding all crimes investigated during the past years. The 2017–2018 annual report points out an amount of 333,223 general crimes, and 13,981 of violent/dangerous crimes, which leads to a total of 347,204 crimes. Considering that several reports and forensic evidences are produced in textual form for each reported crime, the amount of data produced overgrows to numbers depicted and challenge human and digital processing. As an illustration, consider for instance the reports of road accident with fatal victims prepared by the Criminal Investigation Nucleus of the Republican National Guard (in Portuguese: NICAIV/GNR).

Criminal investigators spend time by manually processing and analysing police reports and journal articles. According to conversations with police investigators, it is suggested to use criminal-related articles delivered by online newspapers, as the narrative of facts and actors' descriptions are similar to the police reports. Even slang is used on criminal-related news, like "horse", which is slang for "heroin".

IBM™ i2 Notebook, which is incomplete on processing Portuguese language, is used to graphically map the entities and relations in a predefined Microsoft Excel template. This process is time consuming and inefficient to automatically process the documents. The proposed framework aims to reduce the time spent, and to automatically process the documents. It supports graphical features to produce a graph with the extracted entities and its relationships.

The major motivations that support our proposed framework, described in this paper, can be described as follows:

- Despite the Portuguese language is spoken and written by 250 million people (<https://www.up.pt/portuguesuporto/o-portugues-no-mundo/> (Accessed on: 1 June 2021)), to the best of authors' knowledge there is not a comprehensive set of tools to automatically process criminal-related documents;
- The police data repositories are populated with police reports that are manually processed by police investigators, which constitutes a time-consuming and not error-free task. Therefore, an approach that takes advantage of computational methods to automatically deal with such data is a path to be followed;
- The police reports and online newspapers are available in different formats, namely, structured, semi-structured, and unstructured data, which represents a challenge for the identification and classification of possible entities and relationships between them.

We describe a framework based on a graph database, which provides an unified approach to automatically extract and represent the name-entity relations that can be found in criminal-related documents. The framework is named *SEMCrime*, which means the fusion of the words "*SEMantic*" and "*Crime*". The framework processes data retrieved from criminal-related documents in a semantic form, and populates a *Neo4j* graph database with the retrieved data. The contributions provided by this research work are as follows:

- A systematic approach that ties together the criminal investigation and the computer science domains, focused on the analysis of criminal-related documents in the Portuguese language;
- An end-to-end framework to deal with several phases ranging from data extraction to knowledge representation into a graph database. These phases can be summarized as follows:

- Input: a set of documents that are retrieved from police departments and open sources (online news about crimes), in Portable Document Format (.pdf), Microsoft Word (.doc) and HTML format;
 - Document preprocessing: enables a set of tasks for document processing and Natural Language Processing;
 - Graph database representation: enables the semantic understanding of data retrieved using Named Entity Recognition (NER), Criminal-Term Extraction, Semantic Role Labelling (SRL), and *5W1H* information extraction methods. Finally, the graph database population and enrichment of data retrieved and analyzed in posterior tasks.
- An information extraction method based on an *5W1H* (Who, What, Why, Where, When, and How) approach to understand the semantic relations observed in the extracted entities of the processed documents;
 - A set of machine learning models to identify and classify name-entity pairs related with the criminal domain;
 - An approach to identify and classify terms linked to the criminal domain that can be populated in the documents;
 - A graph database implemented in *Neo4j* (<https://www.neo4j.com/> (Accessed on: 1 June 2021)) to accommodate the named-entities and relations supported by modelling decisions;
 - A prototype to evaluate the framework deployed and a performance assessment extract;
 - A dataset built by a set of documents, such as police reports, criminal and PGdLisboa (Procuradoria-Geral Distrital de Lisboa, in English: District Attorney of Lisbon) news.

2. Literature Review

To support our work, we performed a literature review to analyse related studies. In the following paragraphs, related works are enumerated and ordered by date, with focus on frameworks and NLP applied to the criminal domain that retrieve data from different sources and represent it into a knowledge database.

In 2003, the Coplink [4] project introduced a cooperation between the Tucson Police Department and the University of Arizona-Artificial Intelligence Lab. The main objective was to develop a set of knowledge management technologies addressing challenges such as information sharing and collaboration between police departments and agencies to promote criminal intelligence analysis and knowledge management by each department. Two relevant features were:

- Domain-Specific Detect Concept Space: that identifies documents related to domain-specific concepts or terms and performs a co-occurrence analysis to identify the relationships among indexed terms after filtering and indexing
- Coplink Detect Module: was designed to recommend similar cases to users and identify police officers with similar information needs.

In 2007, the Jigsaw [5] Named-Entity Recognition (NER) tool (<https://www.cc.gatech.edu/gvu/ii/jigsaw> (Accessed on: 1 June 2021)) proposes to identify and classify named-entities in English language, such as persons, locations, objects or actions retrieved from police documents. The key objective of this tool is to establish relationships between entities across document collections, supported by external tools, such as GATE (<https://www.gate.ac.uk/> (Accessed on: 1 June 2021)), LingPipe (<http://www.alias-i.com/> (Accessed on: 1 June 2021)), OpenCalais (<https://www.refinitiv.com/en/products/intelligent-tagging-text-analytics> (Accessed on: 1 June 2021)) and Illinois-NER (<https://cogcomp.seas.upenn.edu/page/software/> (Accessed on: 1 June 2021)).

In 2011, the Police Intelligence Analysis Framework [6] (PIAF) was created to analyze witnesses statements regarding post-incident investigations by using computational methods, such as fusion algorithms. The architecture is based on a front-end (graphic user interface console) and a back-end (server). The server is divided into Software Architecture; Fusion Algorithms, Data Access, and Analysis. For evidence storage, a database

was included on the server. The PIAF uses an entity matching [7] approach to identify incomplete information on witnesses statements about entities, such as persons. The entity matching with the help of fusion algorithms matches the incomplete information with a known entities dataset.

In 2012, Albertetti et al. [8] proposed a data warehouse based system to represent retrieved data from police reports, which is supported by a five-step process, varying from data identification through testing and analyzing tasks. This system is supported by a relational tool for crime analyzes, organizing the police reports into a data warehouse. Authors used a method, named CORDIET [9], for retrieving information from unstructured text supported by use cases provided by the Amsterdam-Amstelland Police, GasthuisZusters Antwerpen (GZA) hospitals, and Katholieke Universiteit (KU) Leuven. Authors evaluated the system with real police datasets using Pentaho™ Data Integration Suite (<https://www.pentaho.com/product/data-integration/>) (Accessed on: 1 June 2021)) tool.

The Combined Websites and Textual Document Framework (CWTDF) [10] was proposed for investigating crime suspects, by retrieving and analyzing web pages to discover crimes, with the help of web mining techniques. A set of crime communities were built using text mining processing, which allows the creation of community profiles and criminal networks that are able to detect crime hotspots indirectly linked to each other.

Hossain et al. [11] proposed a system to build stories based on entity networks, by using a corpus and an entity extraction task to identify and classify named-entities. The Coreference task was added to disambiguate pronominal mentions and references to the same person. Finally, to model the entities detected after entity extraction and disambiguate (in Coreference), and to find the links between two entities, the authors used the *Concept Lattice Generation*, which enables to construct stories to find explanations from links between two entities.

In 2014, Project Multi-Modal Situation Assessment and Analytics Platform, named by MOSAIC [12] was proposed to automatically detect and recognize crimes in specific environments. The architecture was based on semantic information and classification of data sources, where the unstructured and structured data are integrated into a standard data format using ontologies. This platform combines an NLP pipeline, a knowledge base, and a crime pattern detection, which were designed to retrieve entities and events from text documents and websites. CAPER [13] introduces a platform for organized crime detection and prevention, by sharing, exploiting, and analyzing Open-Source Intelligence (OSINT) data. For a semantic representation, the authors used two ontologies (the *European LEAs Interoperability Ontology* and the *Multi-Lingual Crime Ontology*). ePOOLICE [14] proposed a system for detecting criminal threats retrieved from OSINT using an ontology, and knowledge graphs to create an Environmental Knowledge Repository. The data is retrieved using crawlers that extract data from OSINT. Filtering and classification methods were applied to detect named entities.

Environment Knowledge Repository stores the relevant information retrieved from the source environment, which permits the further analysis using several approaches, such as sentiment analysis or conceptual graphs [15].

In 2015, Wijeratne et al. [16] suggested an architecture to discover criminal gangs structure, functioning and modus operandi, by analyzing the social media footprints in Chicago, Illinois (EUA). The architecture detects and retrieves tweets associated with gangs from Twitter, using a Slang Term Dictionary for slang detection mentioned by Chicago gangs. It performs data processing through machine learning and NLP models, such as NER or sentiment analysis. For more in-depth analysis, a question-answering task was added to perform questions like “*Who is the gang user A affiliated with?*”. Sowa [17] proposed a crime detection framework to discover relationships between offenders and communities by extracting information from police documents in the Arabic language.

In 2016, Mata et al. [18] proposed a mobile information system based on crowd-sensed and official crime data, like criminal reports. This approach aims to find safe routes for app users 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 mobile application. It uses a criminal ontology to process data semantically and a Bayes classifier.

In 2018, Wiedemann et al. [19] suggests an information extraction pipeline to automatically process collections of unstructured textual data for investigative journalism. The data sources, named by Hoover, were developed by the European Investigative Collaborations (EIC) network, focusing on large data leaks and heterogeneous datasets. Authors based their work on a configurable pipeline that takes part in the available tools in the following modules: preprocessing, dictionaries, and regular expression patterns, temporal expressions, NER, and term extraction.

Regarding the NLP approaches, we have surveyed a set of research papers that were applied to the criminal domain, which are detailed below.

In 2011, Al-zaidy et al. [20] proposed a set of NLP methods to discover criminal communities, by establishing their relations and extracting relevant information from criminal data, such as e-mails, chat logs, or any textual data. Pinheiro et al. [21] proposed an information extraction system using the Portuguese language, based on a SIM (Semantic Inferential Model) [22] which drives semantic analysis applied on public safety areas, like the police departments. The designed system supports a collaborative web-based system of crime registering, named WikiCrimes [23]. This system performs a morphological and syntactic analysis using NLP tasks to produce syntactic trees as an output.

In 2013, Camara et al. [24] propose to create a system for indexing information of documents using NLP mechanisms at a semantic level, supported by an ontology. The dataset was built with data extracted from forensics reports released by the Federal Police Forensics Experts in the Portuguese-Brazilian language.

In 2014, Arulanandam et al. [25] proposed a system to extract information from online newspapers focused on the “hidden” information related to theft crime. To enable these extractions, a NER task was applied, like CRFs classifiers.

In 2015, Shabat et al. [26] proposed to extract crime information from the web, based on machine learning models and crime NER task, using classification algorithms, e.g., Naïve Bayes, SVM, and K-Nearest Neighbor.

In 2017, Ejem et al. [27] propose, in its Master Thesis, an approach regarding relation extraction between NEs in police reports. These police reports are from the Anti-drug Department of the Police of the Czech Republic.

In 2019, Martin-Rodilla et al. [28] describe the analysis of 3000 textual reports in São Paulo during the Brazilian dictatorship by using unsupervised and supervised approaches, supported by the Linguakit Suite, the Stanford CoreNLP (<https://stanfordnlp.github.io/CoreNLP/> (Accessed on 1 June 2021)) tool, SIEMÊS [29] algorithms, by identifying NEs and relevant terms. This approach tries to set people’s information and automate the study of correlations between actors.

In 2020, Gianola [30] proposed in her Ph.D. thesis an adaptation of NLP methods applied to witness interviews in the French language using official documents from Gendarmerie Nationale (www.gendarmerie.interieur.gouv.fr/ (Accessed on 1 June 2021)). For its identification was based on a rule-based approach, by using Unitex/GramLab (www.unitexgramlab.org/pt (Accessed: 1 June 2021)) tool to support this approach.

Summary

The works previously described have provided valuable insights for SEMCrime framework development, proposed in this paper, which are summarized below:

- The partnerships that have been established between universities and other public institutions, such as police departments, allowed the access and use of classified data, originated from police investigations. Another approach followed by some authors was to combine data from distinct official and public sources;
- Several authors proposed different pipelines that were developed from scratch, or have been configured from others already available. The main purpose was to deal

with data extraction, such as structured (relational databases) or unstructured data (textual data) in several file formats;

- A set of approaches resorted to relational databases or ontologies for knowledge representation. These approaches introduced two different issues: relational databases for the criminal domain fail to represent its unstructured data; ontologies are suitable for the domain but are time-consuming and difficult to build from scratch;
- The use of external knowledge bases for data enrichment, like the *GeoNames* (www.geonames.org/) (Accessed on: 1 June 2021) geographical database;
- The analyzed frameworks were applied mostly to the English language, and are focused on unstructured data (textual data), such as police reports, and have used NLP tasks for different approaches.
- It is possible to confirm the lack of frameworks for criminal domain documents written in Portuguese language and variants (e.g., Brazilian-Portuguese).

3. SEMCrime Framework

This section presents the *SEMCrime* framework, based on a *Neo4j* graph database, and divided into modules that enables the lexical, syntactic, and semantic analysis of criminal-related documents. *SEMCrime* aims to mitigate a gap identified in implementations applied to the criminal domain, by analyzing and processing criminal-related documents written in the Portuguese language. A set of methods have been proposed to identify and classify NEs (Named-Entities) or terms related to the criminal domain, such as narcotics names, and an information extraction method based on *5W1H* information was proposed to better understand the documents. The information extracted was populated into a graph database following a set of rules, and a data enrichment method using *GeoNames* geographical database was proposed in the framework.

An excerpt from a criminal investigation report (persons and locations were changed to maintain anonymity) is analyzed below, to demonstrate the overall framework functioning, namely the need to have an end-to-end pipeline to represent 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 José 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”. (In Portuguese)

“In 18 of April 2008, during a home search in Pedro Silva home, that lives on José Leite Street, in Santa Apolónia District, in Coimbra, was found and seized inside the defendant’s wallet: a small portion of cannabis, with a weight of 1.8 grams.” (In English)

The sentence above outlines some issues that we need to understand to fulfill the framework requirements:

- The documents are in their original file formats, such as Microsoft™ Word, Portable Document Format (PDF), and HTML (for websites);
- After being retrieved, text may contain errors, which can cause problems with later tasks, such as tokenization or stemming;
- The text must be understood from a lexical and syntactical perspective, like tokenization or sentence splitting;
- Several entities were identified, such as persons, locations or references to dates. Entities related to the criminal domain, such as narcotics, have also been identified. The extraction of these entities is useful for the semantic analysis;
- 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 and visualization.

Figure 1 depicts the the *SEMCrime* framework, which is composed of the following main modules:

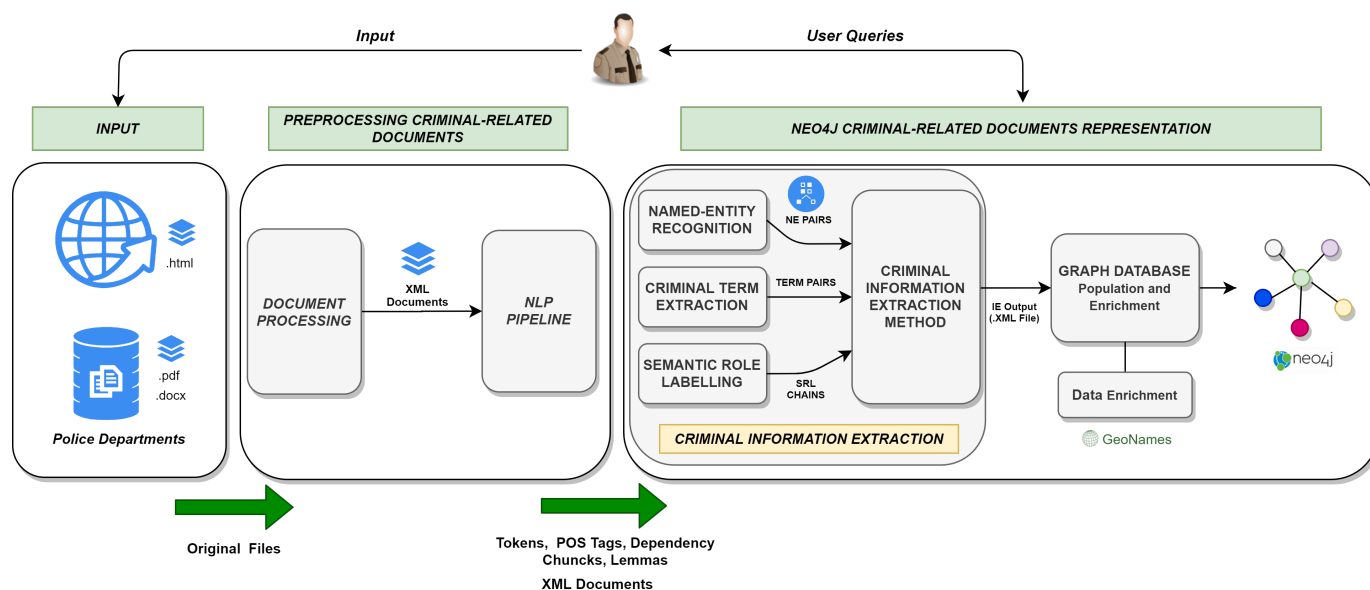


Figure 1. *SEMCrime* Framework Architecture.

- **Input**: takes as input a set of criminal-related documents, obtained from online newspapers and police departments, in its original formats (Microsoft™ Word, Portable Document Format (PDF) or HTML file formats);
- **Preprocessing Criminal-Related Documents**, which is formed by the following pipeline modules:
 - **Document Processing**: this module permits the extraction, transformation and loading of criminal-related documents. Different tasks were applied, for example, a cleaning task to extract words or symbols that may cause “noise” in data. The output is a in Extensible Markup Language (XML) file, with tags to identify the documents content;
 - **NLP Pipeline**: to enable the syntactic analysis of documents. The output is a group of *Tokens*, *POS Tags*, *Dependency Chunks* and *Lemmas* identified in each sentence that belongs to documents.
- **Neo4j Representation**: this module was proposed to semantically understand the documents and the representation of the retrieved data into *Neo4j* graph database. This module is divided into the following blocks:
 - **Criminal Information Extraction**: uses a *Named-Entity Recognition* module to identify the named-entities relevant to the domain, a *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 *Criminal 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 block enables the population of the Neo4j graph database, and the data enrichment using the *GeoNames* geographical database.

3.1. Criminal-Related Documents

The criminal domain has its own vocabulary and narrative, depending on the language in which documents are written. Criminal domain experts advise the inclusion of *criminal*

news and official websites like *PGdLisboa News* documents, arguing that they follow the same narrative form and the same requirements of the criminal investigation reports. These documents intend to answer the 5W1H approach with the following questions: the “Why?”, “Where?”, “When?”, “What?” and “Who” and “How?” [31], which are intrinsically related to the investigation process. In the scope of this research work, a corpus was built, which combines documents from the following sources: *Criminal Investigation Reports*, *Criminal News*, and *PGdLisboa News* retrieved from police departments and online newspapers, described below:

- Criminal Investigation Reports: these reports synthesize in one or multiple documents the information collected during a criminal investigation by grouping the contents of an investigation, such as witnesses, suspects, police investigators, or fact descriptions;
- Criminal News: documents that are published in online newspapers [19] during criminal investigations performed by police departments, written by investigative journalists;
- PGdLisboa News: another source for criminal reports is the *Procuradoria-Geral Distrital de Lisboa* (<https://www.pgdlisboa.pt/> (Accessed on: 1 June 2021)) website. The news are about cases in which there has been a final decision and are no longer subject to appeal.

The analyzed and processed corpus has an amount of 163 documents, which correspond to 1580 sentences and 38,993 words.

3.2. Preprocessing Criminal-Related Documents

As we have already discussed, the amount of data from different sources is an issue for Police departments, such as the Law Courts, criminal news from online newspapers, and the number of criminal investigation reports produced by criminal investigations upon police departments. This fact raises a challenge regarding preprocessing tasks, namely an Extract, Transform, and Load (ETL) task and Natural Language Processing pipeline. 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. Nevertheless, the document’s content followed a template (whether police reports or articles from newspapers) that allowed the annotation through XML tags, such as Document Name, Title, Author(s), Publication Date, and News Text. Listing 1 details the XML Schema for Criminal and PGdLisboa news.

Listing 1. Criminal News and PGdLisboa news—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>
```

To allow documents preprocessing of documents, we propose the following modules:

- *Document Processing* module focused on the adaptation of an ETL approach to identify the required mappings and transformations needed to be done automatically and perform operations that lead to a transformation of unstructured into semi-structured data (represented in an XML file);

- *NLP Pipeline* module enables the NLP tasks regarding lexical and syntax analysis of each document in the Portuguese language, such as sentence splitting, tokenization, or lemmatization.

The analyzed documents reveal acronyms and abbreviations related to criminal domain, such as “PJ” that means “Policia Judiciaria” (in English “Criminal Police”). To treat this, we have added tasks that normalize the acronyms and abbreviations using a pattern matching approach supported by a list of acronyms and abbreviations.

The introduction of this module had a significant impact on the following processing steps, essentially because extracts, transforms, normalizes and loads the documents into a semi-structure form, i.e., XML format tagged with meta-information related to documents structure. Listing 2 shows an example of the XML output file, written in Portuguese language.

Listing 2. Criminal News XML Source Format (extract).

```
<?xml version="1.0" encoding="UTF-8"?>
<News1>
  <documentname>News1</documentname>
  <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>
```

The corresponding English translation is described in Listing 3:

Listing 3. Criminal News XML Source Format (extract).

```
<?xml version="1.0" encoding="UTF-8"?>
<News1>
  <documentname>News1</documentname>
  <authors>['Global_Media_Group']</authors>
  <publicationdate>2019-07-17 20:34:00+00:00</publicationdate>
  <title>Woman killed by her nephew with a white weapon in Sintra. </
  ↪ title>
  <newstext>The friend was stabbed and the other shot (...) </newstext>
</News1>
```

3.3. Neo4j Criminal-Related Documents Representation

This section illustrates the steps followed to represent the data retrieved from criminal-related documents into a *Neo4j* graph database. The *Neo4j Criminal-Related Documents Representation* module is divided into the following components:

- **Criminal Information Extraction:** allows the identification and classification of named-entities (see Section 3.3.1), criminal terms (see section 3.3.2), and semantic roles (see Section 3.3.3);
- **Criminal Information Extraction Method:** to enable the semantic understanding of documents, we have introduced the *5W1H Information Extraction Method* (see Section 3.3.4) that identifies and classifies the 5W's questions based on the 5W1H information, the crime type, and criminal terms;
- **Graph Database Population and Enrichment:** the results of *5W1H Information Extraction Method* needed to be represented in the *Neo4j* graph database, to enable this, a *Graph Database Population and Enrichment* module has been introduced to populate and enrich the *Neo4j* graph database (see Section 3.3.5).

To carried out the viability of our proposal, we proposed several pieces of software that support the work based on rule-sets or training sets (used by supervised approaches) that were manually generated. There are some reasons for this decision:

- this approach allow us to obtain initial results, and decrease the time-consuming associated with supervised methods. Therefore, this approach enables us to capture the language meaning by using linguistic rules, such as relationships between words;
- the built of an annotated corpus for a criminal domain is a time-consuming task and requires efforts from domain experts to produce it, therefore, approaches that used rules can help to validate applied NLP methods;
- our framework used a combined set of approaches, namely supervised (e.g., NER Classifier for Narcotics) and unsupervised (e.g., NER Classifier for Crime Types) thus allowing a combined approach to obtain results.

However, there are some limitations, that are detailed as follows:

- restricted to the terms defined in gazetteers, for example, in the NER module the use of gazetteers to identify and classify role types could have a lack of terms, and then an identification and classification issue;
- the manual maintenance of rules can be a time-consuming task, when rules increase or need an updating task;
- the rules defined must be well-defined, because a misspelled rule will lead to an error;
- the rules are limited to a certain domain, limiting the portability of such approximations.

3.3.1. NER Module

The introduction of a NER module enables us to identify and classify the named-entities relevant to our domain, such as persons, locations, organizations, narcotics, or crime types. An unified approach was proposed, named as *NER-SEMCrime*, applied to the criminal domain using the Portuguese language. To enable this, a group of classifiers was added with specific training, using an annotated dataset to train supervised approaches. Along with this work, 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 considering the restrictions already mentioned), we have studied the following learning methods: pattern rules, gazetteer-based, and supervised learning methods. Thus, having an approach that integrates NEs mentions from domain-dependent/independent to resolve a NER problem and add value to the domain. We have moved 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.

To enable an annotated output of named-entities, a collection of classifiers was proposed using the learning methods enumerated before, namely:

- COMMON Classifier: for persons, locations, time/date and organizations;

- *PATTERNS Classifier* for mobile phone numbers, email addresses, license plates, and zip codes;
- *NARCOTICS Classifier*: for narcotics names: the narcotics names are mentioned in their current and street name (same as slang) across documents. Another motivation is that drug trafficking is one of the most reported (<https://www.pordata.pt/Europa/Crimes+por+categoria-3285/>) (Accessed on: 1 June 2021)) and typified crimes investigated by the Portuguese criminal police;
- *CRIME TYPE Classifier*: for crime type names: identified words or compound words in criminal-related documents that indicates crime names. For instance, in the context of road accidents investigation we can find crimes such as homicide, drug dealing or assault;
- *ROLE TYPE Classifier*: for role type: the use of specific terms used to identify the persons and organizations by its roles. For example, the use of "suspect" to identify a person that is a suspect of a crime, not providing the real proper noun;

Figure 2 shows the *NER-SEMCrime* module that has the following processing chain: sentence detection and tokenization to perform a preprocessing task; followed by a parallel-group of NER classifiers; and finally, the *ENSEMBLE Method Named-Entities* that joins the *NE Pairs* from each classifier.

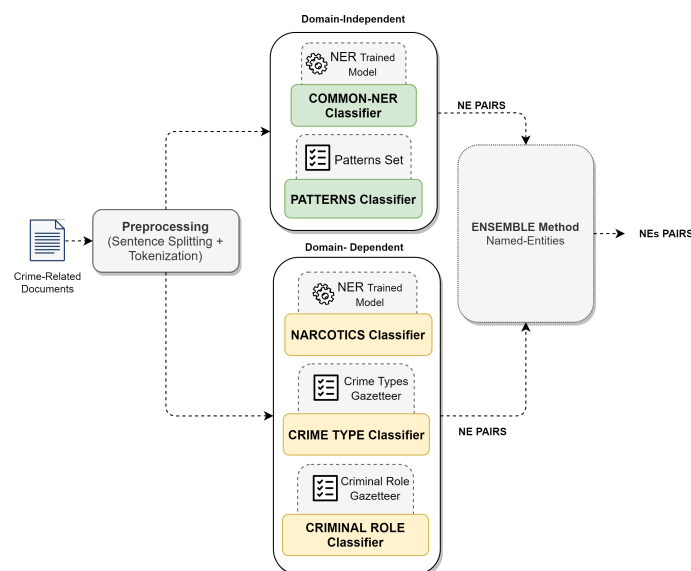


Figure 2. *NER-SEMCrime* Module Architecture.

The sentence below illustrates the (Portuguese) output of this module:

Após investigação da <organization> Policia Judiciaria </organization>, o suspeito <person> Luis Silva </person> foi indiciado pelos crimes de <crimetype> roubo </crimetype>. Os crimes foram cometidos em <location> Coimbra </location>, durante <date> Setembro </date>, com auxilio do veiculo de matricula <licenseplates> XX-XX-11 </licenseplates>. O suspeito era consumidor de <narcotics> cocaina </narcotics>.

The corresponding English translation:

After investigating of the <organization> Policia Judiciaria </organization>, the suspect <person> Luis Silva </person> was indicted for the crimes of <crimetype> theft </crimetype>. The crimes were committed in <location> Coimbra </location>, during <date> September </date>, with the aid of the registration vehicle <licenseplates> XX-XX-11 </licenseplates>. The suspect was a consumer of <narcotics> cocaine </narcotics>.

3.3.2. Criminal Term Extraction Module

The documents contain domain-specific terms identified as being part of the criminal context, such as “*buscas domiciliárias*” (in English: “*home searches*”). We have proposed a *Criminal Term Extraction* module to extract those domain-specific terms. Moreover, we have extracted a list of terms from IATE (<https://iate.europa.eu/> (Accessed on: 1 June 2021)) terminological database related to Portuguese criminal law. We have then obtained a list of around 400 criminal terms and have created a gazetteer. Figure 3 shows a representation of our approach.

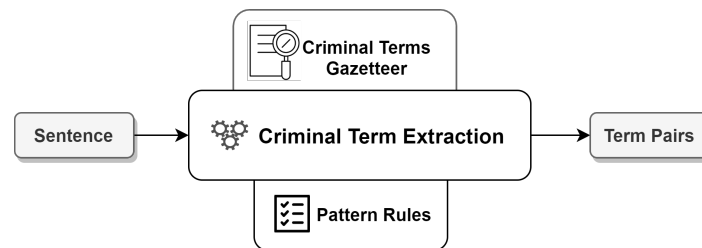


Figure 3. Criminal Term Extraction Scheme.

The sentence below illustrates the (Portuguese) output of this module:

Após investigação, <criminalterm> buscas domiciliárias </criminalterm> foram realizadas na casa do suspeito.

The corresponding English translation:

After investigation, <criminalterm> home searches </criminalterm> were made in the suspect house.

3.3.3. Semantic Role Labeling Module

The introduction of a Semantic Role Labeling (SRL) task in our *Criminal Information Extraction Module* aims to identify and classify the verb-argument structure of each sentence in criminal-related documents. Using this task, we can semantically understand sentences. For example, given a sentence “*O Rui Silva assaltou o Banco de Portugal, pelas 14 horas*”, the (Portuguese) output is:

O Rui Silva Arg0 assaltou v o Banco de Portugal Arg1, pelas 14 horas ArgM-TMP.

The corresponding English translation:

Rui Silva Arg0 robbed v the Bank of Portugal Arg1, by the 2 pm ArgM-TMP.

The detection of arguments associated with the predicate (or verb) of a sentence and how they are classified into their specific semantic roles helped in answering the 5Ws questions.

The purpose of this task was to split sentences into their semantic components, thus allowing analysis by the 5W1H information extraction method (presented below).

3.3.4. 5W1H Information Extraction Method

The introduction of this method relies on the knowledge about criminal investigation, where crime is normally described as a sequence of events, spread among multiple documents. These events represent several situations and actions that can lead to a crime, based on transitional or permanent events that link entities, like a person or organization [31], representing several situations and actions that can lead to a main criminal event.

To understand the sentences that populate each document, we have proposed a method that relies upon the SRL, NER, and *Criminal Term Extraction* modules. This established the answer of the 5Ws questions (Who, What, Where, When, and Why). This module explores the 5Ws questions, crime type and criminal terms identification. Therefore, the contributions for the framework are:

- extracting the event type and elements to answer the 5W1H information in the Portuguese language applied to the criminal domain, permitting the construction of triples that can be used in several tools or knowledge bases, such as graph databases;
- to extract the crime type and criminal terms to enable domain comprehension by adding information that is connected to the criminal domain.

We established a set of hand-crafted rules to extract the information, and fill the Criminal Information Extraction Method file (XML format) with the following output (see Listing 4) for the (Portuguese) sentence below:

O Rui Silva e o Pedro Silva WHO assaltaram WHAT o Banco de Portugal WHOM em Coimbra WHERE pelas 14 horas WHEN.

The corresponding English translation:

Rui Silva and Pedro Silva WHO robbed WHAT the Bank of Portugal WHOM in Coimbra WHERE by 2 pm WHEN.

The Semantic Role Labelling identifies the semantic roles associated with the annotated named-entities, like the persons or organizations, as showed in Listing 4 related to the example sentence.

Listing 4. Extraction Method file example.

```
<?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>
```

The corresponding English translation in Listing 5:

Listing 5. Extraction Method file example.

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

3.3.5. Graph Database Population and Enrichment

The data retrieved from criminal-related documents is converted into a *Neo4j* graph database. To enable this, the Graph Database Population and Enrichment module was developed, to process the XML input file that contains the 5W1H information extracted. The main reasons for using graph databases in this context are described below:

- the graph databases are specially adapted to deal with unstructured data. They represent the entities and relations into nodes, edges, and properties, without requiring a database schema. This approach is ideal for representing data that cannot be easily organized or interpreted by relational databases;
- they are suitable to calculate paths between entities, which can be useful in criminal-related applications. For example, it is possible to obtain relationships between entities not explained in the documents, and apply semantic queries adapted to the linguistic context.

This module was inspired on the Simple Event Model [32] 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 the criminal-related documents. We have defined the nodes by determining whose nodes 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, as depicted in Figure 4: Event, Event Type, Actor, Time, Location, and Crime. Relations were added to link each nodes, such as HAS_ACTOR, HAS_EVENTTYPE, HAS_TIME, HAS_LOCATION, HAS_WHY, HAS_CRIMINALTERM, and HAS_CRIMETYPE. Figure 4 shows the graph representation using the modelling decisions.

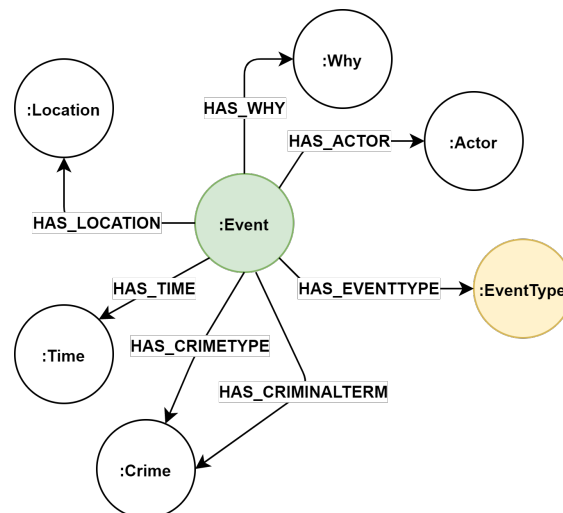


Figure 4. Modelling decisions for graph representation.

To illustrate the proposal, we used the following two (Portuguese) sentences to exemplify the graph database representation.

“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 989999000”.

The corresponding English translation:

“Rui Silva and Pedro Silva robbed the Bank of Portugal in Coimbra, by 2 pm. Rui Silva called Pedro Silva, using a Nokia mobile phone, with the number 989999000”.

Figure 5 shows the graph database after been populated, obtaining 10 nodes and relations.

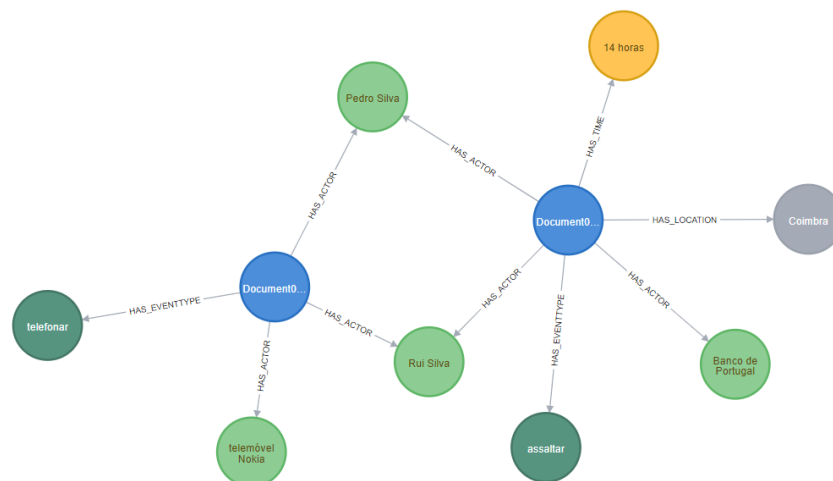


Figure 5. Neo4j representation of sentences.

This module adds a data enrichment feature to the framework, which uses the *GeoNames* geographical database to retrieve latitude and longitude coordinates. This enables us to get and pinpoint the *:Location* node on a map. Figure 6 depicts the enrichment data scheme used on our proposal.

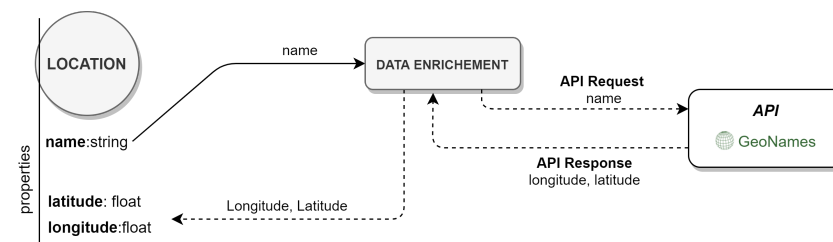


Figure 6. Enrichment Data Scheme.

The task workflow starts with the *GeoNames* API request using the property name (a street or city name) from Location node, that returns an *GeoNames* API response with latitude and longitude.

4. Implementation and Results

For framework evaluation, we have proposed a prototype that enabled us to explore the design issues, functionality, and evaluation measures. Our prototype was developed using Java and Python programming language, *Apache Tika* toolkit for Java, *Newspaper3k* (<https://newspaper.readthedocs.io/en/latest/> (Accessed on: 1 June 2021)) for article scraping and curation, *NLPNET* (<https://www.github.com/erickrf/nlpnet/> (Accessed on: 1 June 2021)) a Python library for NLP tasks based on neural networks, *Apache OpenNLP* toolkit (<http://opennlp.apache.org/> (Accessed on: 1 June 2021)), and the *NLPPort* (<https://www.github.com/rikarudo/NLPPORT/> (Accessed on: 1 June 2021)) toolkit.

For NER evaluation, manual annotation was performed against a set of criminal-related documents (due to the lack of a gold standard for the criminal domain in the Portuguese language) after annotating the documents by identifying and classifying each sentence named-entity and entity types. The evaluation measures were extracted for Precision (P), Recall (R), and F-Measure (F1).

Table 1 summarises the evaluation measures related to NER task.

Table 1. Criminal-Related Documents Evaluation.

	P	R	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

The obtained results described in Table 1 identify the average for the three groups of documents, a Precision above 80%, and the Recall and F-Measure above 70%. We have obtained best results regarding Precision measures in Criminal and PGdLisboa News compared to Criminal Investigation Reports results. However, in terms of F-Measure and Recall results, the Criminal Investigation Reports achieved the best results. These results were obtained due to the fact of some features such as capitalization, enables the correctness of identification and classification of named-entities.

The *5W1H Information Extraction Method* was evaluated using an annotated set of 20 criminal-related documents. Table 2 summarises the evaluation measures for the proposed set.

Table 2. 5W1H Information Extraction Method Evaluation.

P	R	F1
0.732	0.634	0.653

The obtained results aggregate the average for Precision, Recall, and F1-Measure for the 5Ws (Who, Where, Why, What, and When) questions. These promising results encourages us to continue the investigation. The *5W1H* method reached 73% for Precision, however, with a Recall under 63%. Globally, giving us an auspicious path regarding applying this framework to the criminal domain.

The overall perspective of the framework evaluation process can be improved by undergoing another experiments, such as increasing the training datasets or adding more training rules (when needed), which calls for future improvements for each module.

Case Study

The objective of this section is to evaluate our proposal against a real criminal investigation report. The framework was evaluated by a police investigator, and a comparison with IBM™ i2 Analyst's Notebook was made. That is, the same police report was evaluated by the proposed framework and the IBM™ i2 application, and the conclusions were registered. The following actions were taken:

- the analysis of the criminal investigation report, made by a domain expert, each person and location names, phone numbers, or license plates were changed or masked;
- the report was submitted into our framework;
- the domain expert analyzed the criminal investigation report using the IBM™ i2 Analyst's Notebook tool.

Figures 7 and 8 show the results obtained with our proposal and with IBM™ i2 Analyst's Notebook, respectively. 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 on the name (depends on the domain experts perspective and could differ from each other). From the obtained results, we can infer the viability of using SEMCrime framework to extract and represent the data retrieved from criminal-related documents.

After populating the graph database with data retrieved from documents, we query the Neo4J database using Cypher Query Language statements. Two questions were made against the database, and the results can be seen in Figures 9 and 10.

Question 1: *Who is/are the actor(s) that are related to a consumption of narcotic drugs (in Portuguese “consumo de estupefacientes”)?*

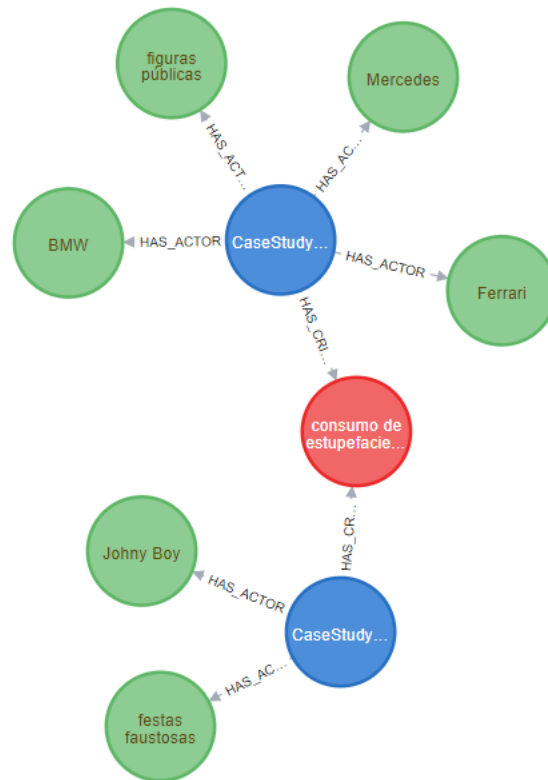


Figure 9. Cypher Query Output (Question 1).

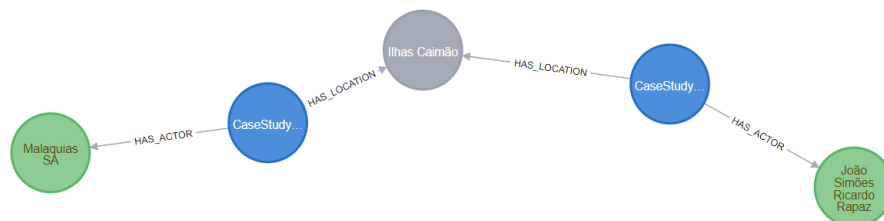


Figure 10. Cypher Query Output (Question 2).

Question 2: *What is the shortest path between “João Simões Ricardo Rapaz” and “Malaquias SA”?*

To achieve the shortest path between the suspect “João Simões Ricardo Rapaz”, and an off-shore organization named “Malaquias SA”, a Neo4j inbuilt shortest-path search algorithm was used, like the single shortest path algorithm. The algorithm parameters were set up with start and end nodes, without specifying the direction of the relations.

Summing up, the case study evaluation obtained similar results with both manual analysis performed by a police investigator and the proposed framework, on identifying and classifying entities and detecting events. Overall, our approach performs well to represent a cluster of relations and entities that populate the criminal-related documents, and end-used queries.

5. Conclusions and Future Work

This paper addressed the challenge that emerges from the need for automatic processing and representation of unstructured data from criminal-related documents. As a result, the *SEMCrime* framework deals with the enumerated challenge. For that to be possible, we have defined the following assumptions:

- The focus is on the Portuguese language, without discarding what has been done in other languages;
- The approaches applied to the criminal domain and related works were studied and analyzed;
- A survey of existing ETL, NLP, Graph Database approaches was made and, for each one, a list was presented, with the features that can be proposed, used or adapted;

The *SEMCrime* framework solves an emerging and ambitious challenge regarding the processing of Portuguese unstructured criminal reports files, mainly because it is applied to a domain without a solid background and relevant work-related to the Portuguese language, despite the works already published and applied to other cases such as the English language. During the implementation, we have faced the following main challenges:

- the criminal-related documents have different content structures and file formats; For instance, investigation reports of road accidents with fatal victims have a particular template;
- the extracted plain text may contain errors or noise identified during the extraction phase, such as double space or extra symbols;
- the existence of abbreviations and acronyms related to the domain;
- the existence of entities related to the domain that are not identified and classified by the NER approaches, such as narcotics or crime types;
- the use of domain-specific terms related to the criminal domain, such as “Pulseira Eletrónica” (in English: “ankle bracelet”);
- as in other written text, the criminal-related documents need to be semantically understood;

From the study developed, several approaches identified methods, tools, or techniques that could be useful for our approach, and whenever is possible and viable, these methods, tools, or techniques were applied to retrieve, transform, syntax and semantic analysis, and represent data into a graph database. As a result, a set of critical tasks were identified, like the analysis of the criminal-related documents and their content, 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 acronym “PSP” refers to “Policia de Segurança Publica”;
- a NLP pipeline was introduced, performing tasks like tokenization or sentence splitting;
- an NER module was used to identify and classify the NEs relevant to the domain. In this phase, we have proposed classifiers that identify NEs 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 enable the identification of the semantic roles;
- identifies and classifies the *Who, What, Where, When, Why and How*; using this method, we tried to find the answers to the 5Ws in each sentence, which outputs a network of entities and relations.

After identifying the tasks that allow us to proposed an framework that approaches the challenge, retrieving unstructured data from criminal-related documents and populate it into a *Neo4j* graph database. For evaluation, an prototype was developed to obtain evaluation measures, such as in the NER module and the 5H1W method results.

Retrospectively, our proposal was ambitious, mainly since the framework is applied to a domain without a solid background and relevant work-related to the Portuguese reality, despite the works already published and applied to other cases such as the English

language. The results obtained with the prototype reveal the viability of our proposal to support police investigators in their daily activities when they need to process large amounts of data in documents, reducing manual work and time-consuming.

For future work, in order to improve the proposed work some paths may be taken, such as the introduction of a coreference resolution module, to determine whether two expressions to the same entity in the real world, in order to improve the identification of entities and their relationships; the performance of the SRL module could be improved; add the *1H* (How) question; improve the NER module with new NE related to the domain; train the *Criminal Role* and *Crime Type* NEs with a supervised approach; propose an annotated corpus related to the criminal domain for a NER task; understand slang written in criminal-related documents, and finally extract biographical information about persons and organizations (this includes the affiliations, such as gang affiliations).

Author Contributions: Conceptualization, G.C., V.B.N., M.A.; Data curation, G.C.; Formal analysis, G.C., V.B.N., M.A.; Funding acquisition, V.B.N.; Investigation, G.C., V.B.N., M.A.; Methodology, G.C., V.B.N., M.A.; Software, G.C.; Supervision, V.B.N., M.A.; Validation, V.B.N., M.A.; Visualization, G.C.; Writing - original draft: G.C., V.B.N., M.A.; Writing - Review & Editing; G.C., V.B.N., M.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT-Fundação para a Ciência e a Tecnologia, under the project with reference FCT DSAIPA/DS/0090/2018, “MOPREVIS - Modelação e Predição de Acidentes de Viação no Distrito de Setúbal”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is available under a Creative Commons Attribution 4.0 International License, at the following GitHub repository: <https://github.com/goncalofcarnaz/Annotated-Corpus-of-Criminal-Related-Portuguese-Documents>.

Acknowledgments: The authors would like to thank project “MOPREVIS - Modelação e Predição de Acidentes de Viação no Distrito de Setúbal”, with reference FCT DSAIPA/DS/0090/2018, financed by the Foundation for Science and Technology (FCT) within the scope of the National Initiative on Digital Skills e.2030, Portugal INCoDe.2030

Conflicts of Interest: The authors declare no conflict of interest

Abbreviations

The following abbreviations are used in this manuscript:

5W1H	Who, What, Where, When, Why and How
CRF	Conditional Random Field
CWTDF	Combined Websites and Textual Document Framework
ETL	Extract, Transform, Load
GNR	Republican National Guard
NE	Named Entity
NER	Named Entity Recognition
NICAV	Traffic Accident Criminal Investigation Nucleus
NLP	Natural Language Processing
OSINT	Open Source Intelligence
PIAF	Police Intelligence Analysis Framework
POS	Part-Of-Speech
SRL	Semantic Role Labelling
SVM	Support Vector Machine

References

1. Gleick, J.; Calil, A. *A Informação: Uma História, Uma Teoria, Uma Enxurrada*; Companhia das Letras: Sao Paulo, Brazil, 2013.
2. Oussous, A.; Benjelloun, F.Z.; Lahcen, A.A.; Belfkih, S. Big Data technologies: A survey. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *30*, 431–448.

3. Cavanillas, J.M.; Curry, E.; Wahlster, W. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*; Springer: Berlin/Heidelberg, Germany, 2016.
4. Chen, H.; Zeng, D.; Atabakhsh, H.; Wyzga, W.; Schroeder, J. COPLINK: Managing law enforcement data and knowledge. *Commun. ACM* **2003**, *46*, 28–34.
5. Stasko, J.; Görg, C.; Liu, Z.; Singhal, K. Jigsaw: Supporting investigative analysis through interactive visualization. In Proceedings of the VAST IEEE Symposium on Visual Analytics Science and Technology, Sacramento, CA, USA, 28 October–1 November 2007; Volume 1, pp. 131–138, doi:10.1109/VAST.2007.4389006.
6. Stampouli, D.; Roberts, M.; Powell, G.; Lopez, T.S. Implementation of a police intelligence analysis framework. *Int. J. Secur. Its Appl.* **2011**, *5*, 13–22.
7. Köpcke, H.; Rahm, E. Frameworks for entity matching: A comparison. *Data Knowl. Eng.* **2010**, *69*, 197–210.
8. Albertetti, F.; Stoffel, K. From police reports to data marts: A step towards a crime analysis framework. In Proceedings of the 5th International Workshop on Computational Forensics, Tsukuba, Japan, 11 November 2012.
9. Poelmans, J.; Elzinga, P.; Neznanov, A.A.; Dedene, G.; Viaene, S.; Kuznetsov, S.O. Human-centered text mining: A new software system. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **2012**, *7377*, 258–272, doi:10.1007/978-3-642-31488-9_21.
10. Hosseinkhani, J.; Chaprut, S.; Taherdoost, H. 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), Prague, Czech Republic, 24–26 September 2012; pp. 210–215.
11. Hossain, M.S.; Butler, P.; Boedihardjo, A.P.; Ramakrishnan, N.; Tech, V. Storytelling in Entity Networks to Support Intelligence Analysts. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012.
12. Adderley, R.; Seidler, P.; Badii, A.; Tiemann, M.; Neri, F.; Raffaelli, M. Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights. *Fourth Int. Conf. Adv. Inf. Min. Manag.* **2014**, *1*, 36–40.
13. Casanovas, P.; Arraiza, J.; Melero, F.; González-Conejero, J.; Molcho, G.; Cuadros, M. Fighting Organized Crime Through Open Source Intelligence: Regulatory Strategies of the CAPER Project. *Front. Artif. Intell. Appl.* **2014**, *271*, 189–198, doi:10.3233/978-1-61499-468-8-189.
14. Brewster, B.; Andrews, S.; Polovina, S.; Hirsch, L.; Akhgar, B. Environmental scanning and knowledge representation for the detection of organised crime threats. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **2014**, *8577*, 275–280, doi:10.1007/978-3-319-08389-6_22.
15. Sowa, J.F. Chapter 5 Conceptual Graphs. In *Handbook of Knowledge Representation; Foundations of Artificial Intelligence*; van Harmelen, F., Lifschitz, V., Porter, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Volume 3, pp. 213–237, doi:10.1016/S1574-6526(07)03005-2.
16. Wijeratne, S.; Doran, D.; Sheth, A.; Dustin, J.L. Analyzing the social media footprint of street gangs. In Proceedings of the 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), Baltimore, MD, USA, 27–29 May 2015; pp. 91–96, doi:10.1109/ISI.2015.7165945.
17. Elyezjy, N.T.; Elhalees, A.M. Investigating Crimes using Text Mining and Network Analysis. *Int. J. Comput. Appl.* **2015**, *126*, 19–25.
18. Mata, F.; Torres-ruiz, M.; Guzmán, G.; Quintero, R.; Zagal-flores, R.; Moreno-ibarra, M.; Loza, E. A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes : A Case Study of Mexico City. *Mob. Inf. Syst.* **2016**, *2016*, 11.
19. Wiedemann, G.; Yimam, S.M.; Biemann, C. A Multilingual Information Extraction Pipeline for Investigative Journalism. *arXiv* **2018**, arXiv:1809.00221.
20. Al-Zaidy, R.; Fung, B.C.M.; Youssef, A.M. Towards Discovering Criminal Communities from Textual Data. In *Proceedings of the 2011 ACM Symposium on Applied Computing, TaiChung, Taiwan, 1 January 2011*; ACM: New York, NY, USA, 2011; pp. 172–177, doi:10.1145/1982185.1982225.
21. Pinheiro, V.; Furtado, V.; Pequeno, T.; Nogueira, D.; Aplicada, I. Natural Language Processing Based on Semantic Inferentialism for Extracting Crime Information from Text. In Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, BC, Canada, 23–26 May 2010.
22. Pinheiro, V.; Pequeno, T.; Furtado, V.; Assunção, T.; Freitas, E. SIM: Um modelo semântico-inferencialista para sistemas de linguagem natural. In Proceedings of the Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, Vila Velha, Brazil, 26–29 October 2018; pp. 353–358.
23. Furtado, V.; Ayres, L.; Oliveira, M.D.; Vasconcelos, E.; Caminha, C.; Orleans, J.D.; Belchior, M. Collective intelligence in law enforcement—The WikiCrimes system. *Inf. Sci.* **2010**, *180*, 4–17, doi:10.1016/j.ins.2009.08.004.
24. Camara Junior, A.T.D. Processamento de linguagem natural para indexação automática semântico-ontológica. *Rev. Ibero Am. Ciência Informação* **2013**, *9*, 569.
25. Arulanandam, R.; Savarimuthu, B.T.R.; Purvis, M.A. Extracting Crime Information from Online Newspaper Articles. In *Proceedings of the Second Australasian Web Conference, Auckland, New Zealand, 20–23 January 2014*; Australian Computer Society, Inc.: Darlinghurst, Australia, 2014; Volume 155, pp. 31–38.
26. Shabat, H.A.; Omar, N. Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East J. Sci. Res.* **2015**, *23*, 1215–1221, doi:10.5829/idosi.mejsr.2015.23.06.22271.

27. Ejem, R. *Relation Extraction in Police Records*; Univerzita Karlova, Matematicko-Fyzikální Fakulta: Chéquia, Czech Republic, 2017.
28. Martin-Rodilla, P.; Hattori, M.L.; Gonzalez-Perez, C. Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship. *Information* **2019**, *10*, 231.
29. Sarmiento, L. SIEMÊS—A named-entity recognizer for portuguese relying on similarity rules. In *Proceedings of the International Workshop on Computational Processing of the Portuguese Language, Itatiaia, Brazil, 13–17 May 2006*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 90–99.
30. Gianola, L. Aspects Textuels de la Procédure Judiciaire Exploitée en Analyse Criminelle et Perspectives Pour son Traitement Automatique. Ph.D. Thesis, Université de Cergy-Pontoise, Cergy, France, 2020.
31. Braz, J. *Investigacao Criminal*; Almedina: Coimbra, Portugal, 2019.
32. Van Hage, W.R.; Malaisé, V.; Segers, R.; Hollink, L.; Schreiber, G. Design and use of the Simple Event Model (SEM). *J. Web Semant.* **2011**, *9*, 128–136.