data

MDPI

Data Descriptor

An Annotated Corpus of Crime-Related Portuguese Documents for NLP and Machine Learning Processing

Gonçalo Carnaz [1,*], Mário Antunes [2,3] and Vitor Beires Nogueira [1]

[1] Informatics Departament, University of Évora, 7002-554 Évora, Portugal; vbn@uevora.pt
[2] Computer Science and Communication Research Centre (CIIC), School of Technology and Management, Polytechnic of Leiria, 2411-901 Leiria, Portugal; mario.antunes@ipleiria.pt
[3] INESC TEC, CRACS, 4200-465 Porto, Portugal
* Correspondence: d34707@alunos.uevora.pt

Abstract: Criminal investigations collect and analyze the facts related to a crime, from which the investigators can deduce evidence to be used in court. It is a multidisciplinary and applied science, which includes interviews, interrogations, evidence collection, preservation of the chain of custody, and other methods and techniques of investigation. These techniques produce both digital and paper documents that have to be carefully analyzed to identify correlations and interactions among suspects, places, license plates, and other entities that are mentioned in the investigation. The computerized processing of these documents is a helping hand to the criminal investigation, as it allows the automatic identification of entities and their relations, being some of which difficult to identify manually. There exists a wide set of dedicated tools, but they have a major limitation: they are unable to process criminal reports in the Portuguese language, as an annotated corpus for that purpose does not exist. This paper presents an annotated corpus, composed of a collection of anonymized crime-related documents, which were extracted from official and open sources. The dataset was produced as the result of an exploratory initiative to collect crime-related data from websites and conditioned-access police reports. The dataset was evaluated and a mean precision of 0.808, recall of 0.722, and F1-score of 0.733 were obtained with the classification of the annotated named-entities present in the crime-related documents. This corpus can be employed to benchmark Machine Learning (ML) and Natural Language Processing (NLP) methods and tools to detect and correlate entities in the documents. Some examples are sentence detection, named-entity recognition, and identification of terms related to the criminal domain.

Dataset: https://github.com/goncalofcarnaz/Annotated-Corpus-of-Criminal-Related-Portuguese-Documents

Keywords: crime-related documents; cybersecurity; criminal investigation; Portuguese language corpus; natural language processing; 5W1H

Citation: Carnaz, G.; Antunes, M.; Nogueira, V. An Annotated Corpus of Crime-Related Portuguese Documents for NLP and Machine Learning Processing. *Data* **2021**, *6*, 71. https://doi.org/10.3390/data6070071

Academic Editor: Erik Cambria

Received: 2 June 2021
Accepted: 24 June 2021
Published: 26 June 2021

1. Background and Summary

Criminal activity is present daily, in a multiplicity of illegal actions in several domains, such as drug trafficking, computer crime, and theft, just to mention a few examples. Upon the occurrence of a crime, criminal police investigators are in charge and start a set of actions to enable the construction of the so-called chain of custody [1] to identify the alleged culprits and to present them in court.

These actions are multidisciplinary and, depending on the crime, they may involve distinct tasks, such as digital and biological forensics analysis, interviews with witnesses, and interrogations with suspects and other individuals that may be potentially implicated. All of these actions produce textual faithful reports throughout the investigation, where all the

facts are extensively described, together with an exhaustive identification of the individuals, places, and other entities that could be relevant for the course of the investigation.

During the investigation, these reports are carefully analyzed, searching for relations between names, places, license plates, and other entities. This is a manual and time-consuming task for the investigators, which usually concentrate the information gathered on a wall dashboard, where many pieces of papers with names and entities are posted and, in a certain way, visually interconnected. Some tools are being used to help investigators' work, such as customized Microsoft™ Excel spreadsheets or the widely used criminal investigation tool IBM™i2 Analyst's Notebook (see https://www.ibm.com/products/i2-analysts-notebook (accessed on 1 June 2021)).

Several comprehensive research works have also produced dedicated tools and frameworks to automatically extract entities and their relationships, from a set of documents. Some of these tools and frameworks are indicated as follows: the Jigsaw [2], the Police Intelligence Analysis Framework [3], and the Combined Websites and Textual Document Framework (CWTDF) [4].

There are also crime-related ontologies to interpret terms and relations in this context, and they are further used for knowledge representation in some existing frameworks. Some examples are the Project Multi-Modal Situation Assessment and Analytics Platform (MOSAIC) [5]; the CAPER [6], which uses simultaneously the *European LEAs Interoperability Ontology* and the *Multi-Lingual Crime Ontology*; and the ePOOLICE [7] project. However, these tools usually have two main flaws. Firstly, they are not multi-lingual, and the existing annotated corpus for criminal domain are mostly available for the English language. Secondly, these tools possess limited visualization features, namely on representing graphically the recognized named-entity relationships.

Portuguese has around 230 million native speakers. It is the ninth most spoken Indo-European language (see https://www.visualcapitalist.com/100-most-spoken-languages/ (accessed on 1 June 2021)) and the sixth most spoken by number of native speakers (see https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world (accessed on 1 June 2021)). To the best of the authors' knowledge, an annotated set of crime-related documents written in the Portuguese language, composed of a training and testing corpus, which can be widely used to evaluate information retrieval system performance has not yet been developed. Bearing that in mind, the construction of an annotated corpus for crime-related documents is of crucial importance.

In the criminal domain, if the corpus content represents the specifications of linguistic phenomena, and if an extrapolation to a more significant population from which it is taken is possible, then it is possible to say that it *"represents that language variety"*. In [8], the authors proposed to extract attributes that can be used to define the different types of texts and contribute to creating a balanced corpus. The criminal domain has its own vocabulary and narrative and assimilates the writing style. Consequently, criminal domain experts advise the inclusion of criminal news and official websites, arguing that they follow the same narrative form and similar requirements of the criminal investigation reports.

This paper presents an annotated corpus for the Portuguese language, which can be applied to information retrieval from crime-related documents. The corpus was evaluated in [9] where a framework was deployed to apply Natural Language Processing (NLP) and Machine Learning (ML) methods, with the aim to extract and classify named-entities and relations extracted from Portuguese criminal reports and documents. A 5WH1 (Who, What, Why, Where, When, and How) information extraction method was also applied, and the relations extracted were stored and represented in a graph database. The corpus was evaluated by a developed prototype, composed of the following components and technologies: Apache Tika toolkit for detection and extraction of metadata and text from files, Newspaper3k (https://newspaper.readthedocs.io/ (accessed on 9 June 2021)) for article scraping and curation; NLPNET (https://www.github.com/erickrf/nlpnet/ (accessed on 9 June 2021)), a Python library for Natural Language Processing (NLP) tasks based on neural networks; Apache OpenNLP toolkit (http://opennlp.apache.org/ (accessed on 9

June 2021)); and the NLPPort (https://www.github.com/rikarudo/NLPPORT/ (accessed on 9 June 2021)) toolkit. For Named-Entities Recognition (NER) evaluation, the documents were manually annotated and processed by the framework. The NER achieved an F1-score of 0.73, while 5W1H (Who, What, Whom, When, Where, How) information extraction performance attained an F1-score of 0.65.

The proposed crime-related documents dataset for the Portuguese language has the following benefits for researchers and practitioners: (1) a clean and organized set of Portuguese crime-related documents in XML format; (2) a corpus with annotated named-entities extracted from the available documents; (3) an initial approach of annotated documents to answer the 5W1H questions set; and (4) an annotated corpus for the narcotics type of crime.

The remainder of this paper is organized as follows. Section 2 describes the dataset collection and processing, namely the criminal news articles, PGdLisboa News, and Criminal Investigation Reports. It also details the anonymization applied to these documents and the dataset that was built to extract the semantic information from sentences, by using the 5W1H approach. Section 3 details the methods applied to process and use the data. Finally, Section 4 describes the technical validation of the dataset.

2. Data Description

The dataset described in this paper corresponds to an annotated corpus derived from Portuguese crime-related investigation reports and criminal news and is available at the following GitHub repository: https://github.com/goncalofcarnaz/Annotated-Corpus-of-Criminal-Related-Portuguese-Documents (accessed on 25 June 2021). It was tested and evaluated in the SEMantic Crime framework, developed by the authors and recently published in [9,10].

The dataset is composed of a set of XML files, each one corresponding to an annotated document. The crime-related documents are from three distinct types and were originally retrieved from the following sources:

- Criminal Investigation Reports (CIR): These reports synthesize, in one or multiple documents, the information collected during a criminal investigation, namely witnesses, suspects, police investigators, or fact descriptions. These documents are related with interviews, forensics analysis, interrogations, and other reports produced during the investigation. CIR were manually anonymized (detailed in Section 3.3) and their original context preserved.
- Criminal News (CN): The documents were published in online newspapers [11], usually written by investigative journalists. Two examples can be pointed out: the crime section of the online version of the Portuguese newspaper *"Jornal de Noticias"* (https://www.jn.pt/tag/crime.html (accessed on 1 June 2021)) and, in English language, the crime section of *"CBS News"* (https://www.cbsnews.com/crime/ (accessed on 1 June 2021)). Both examples describe victims, suspects, and facts.
- Procuradoria-Geral Distrital de Lisboa (PGdLisboa) News: Another source for criminal reports is the PGdLisboa website (https://www.pgdlisboa.pt/ (accessed on 1 June 2021)). The news are about cases with a final decision and are no longer subject to appeal.

Table 1 enumerates the amount of crime-related documents and syntactic components that were used to build the corpus.

Table 1. Crime-related document corpora.

Source	Words	Characters	Sentences	Texts
Criminal News	18,192	111,637	667	80
PGdLisboa News	16,020	100,720	533	80
Criminal Investigation Reports	4781	24,567	380	3
Total	**38,993**	**236,924**	**1580**	**163**

The components retrieved from the documents (described in Table 1) have produced a set of named-entities from different types, which are indicated in Table 2.

Table 2. Total of annotated named-entities in crime-related documents.

Named-Entities by Type	Count
Numeric	557
Person	401
Location	161
Organization	817
Time/Date	417
Crime Type	307
Narcotics	27
License Plates	7

The dataset construction was made by a crawler software (described below), which processes the available documents that have the following general features:

- are written in a free text form, whether in unstructured or semi-structured format; and
- can be available online or offline.

The XML files were created from files with different formats, namely Microsoft™ Word, PDF, and HTML. The following applications were applied to process each file to the XML formats:

- Apache Tika toolkit to process Microsoft™ Word and PDF files; and
- Newspaper3k toolkit, for online article scraping and curation, to process HTML files.

A cleaning method was developed, which ensures the following set of rules:

- remove spaces, line breaks, duplicate white-spaces, and tabs;
- consider commas followed by a space;
- each sentence contains a single end-mark;
- remove all characters that are not in the ASCII character set; and
- split attached words, such as "RuaPrincipal" should be replaced by ''Rua Principal".

A method to process abbreviations and acronyms was proposed and validated by Carnaz et al. [9]. In general terms, a database of acronyms and abbreviations was setup and is being fed with new coming and confirmed entries. A pattern-based rule set is used to search for terms that are candidates to be considered as abbreviations or acronyms. If these new terms already exist on the list, they are expanded. Otherwise, they are annotated and added to the list.

Figure 1 depicts the overall process to collect and process the data. The documents are processed and converted to an XML format. After that, each document undergoes an "Extract, Transform, Load" process, to be subsequently annotated.

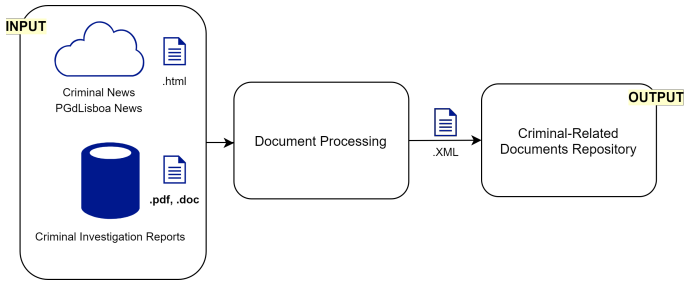

Figure 1. Data collection and processing.

Several tasks were applied to extract data from police reports and open sources:

- data were extracted from websites and files in Microsoft™ Word, PDF, or HTML formats;

- words or symbols that may cause *"noise"* or are not relevant were removed by the cleaning tasks;
- transformations were applied to expand acronyms and abbreviations; and
- an XML schema and the corresponding XML files were created (see Sections 3.1 and 3.2) for each crime-related documents types.

A Java class was developed to convert the documents to the XML formats detailed in Section 3. The class and the corresponding methods are available in the GitHub repository.

3. Methods

This section details the methods applied to collect, process, and use the data. More specifically, it describes the processing of PGdLisboa and criminal news, as well as criminal investigation reports.

3.1. Online Criminal News and PGdLisboa Articles

Online newspapers are a privileged medium to spread crime-related news, where actors and facts are identified and described. It is an open source of knowledge available in a wide set of languages and an interesting way for dataset enrichment. Despite the restrictions imposed by the criminal domain, namely the issues related to data and investigation confidentiality, these documents are worth collecting and including in the dataset, due to the following main reasons:

- the narrative is similar to the one observed in police investigation reports;
- the use of entities to describe the crime, such as individuals' names, is also part of criminal news;
- the use of terms that obfuscate the entities, such as personal names being replaced by "suspect", is also identified in these documents.

Listing 1 details the XML schema that was used for online criminal news and for the PGdLisboa articles. The content of the online criminal news and PGdLisboa online articles follows a well-known and easily recognizable template. The ''`element name`'' XML tag was used to annotate the most relevant data, namely document name, title, author(s), publication date, and the text itself. The XML files related to the documents extracted from the online news, and PGdLisboa articles, are available at the GitHub repository (folder `/Data Set/Data Collection/`).

Listing 1: Crime 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>
```

Table 3 describes the different types of the ''`element name`'' XML tag that were processed in the criminal news and PGdLisboa articles and depicted above in Listing 1.

Table 3. Crime and PGdLisboa news elements description.

Element Name (Tag)	Name	Description
documentname	Document Name	name of the document
authors	Authors	document authors
publicationdate	Publication Date	publication date
title	Title	title in document
newstext	News Text	body of the document

3.2. Criminal Investigation Reports

The criminal investigation reports detail the information collected during an investigation, each one possessing one or more documents. These documents are usually closed and with restricted access (classified), which brings additional challenges to the documents' analysis. Listing 2 depicts the XML schema applied to the criminal reports, where it is possible to identify the ''`element name`'' XML tags used to extract the most relevant content. The layout is similar to the one used by the criminal and PGdLisboa news processing (Section 3.1).

Listing 2: 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 = ''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''/>
 <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>
```

The layout structure was extracted by manually analyzing the reports and the set of tags that were used for text labeling. The document name, author(s), publication date, process identifier number (internal police number to identify each report), title, and document body are defined as annotation sections in the documents. The documents were anonymized to omit persons, phone numbers, and other confidential data (Section 3.3). Images were also disregarded from the document's preprocessing and analysis. Table 4 describes the distinct ''`element name`'' tags that were identified in the XML schema.

Table 4. Criminal investigation reports elements description.

Element Name (tag)	Name	Description
documentname	Document Name	name of the document
authors	Authors	document authors
publicationdate	Publication Date	publication date
cpn	CPN	Criminal Process Number
title	Title	title of the document
documentbody	Document Body	body of the document

3.3. Anonymization

In order to address the privacy and data protection concerns, the documents related with the criminal investigation reports were manually anonymized to remove all Personally Identifiable Information (PII) such as name, address, phone number, license plate, and other personal information. The following tags were defined to identify entities that have to be anonymized in the criminal investigation reports: PERSON, LOCATION, NUMERIC, ORGANIZATION, TIME/DATE, LICENSEPLATES, and PHONENUMBER. For each predefined tag, a sequential number was added at the end in each occurrence in the text. Below, we present a sentence that illustrates the output provided by the anonymization task in a criminal investigation report:

In Portuguese:

"Na sequência das detenções efectuadas, foram o PERSON01 e a PERSON02 presentes à Justiça".

In English:

"Following the arrests made, PERSON01 and PERSON02 were brought to justice".

This way, the official documents were de-identified, by changing names, places, and other PII. Notwithstanding, the criminal investigation reports have already become res judicata, being publicly available, after proper request for full-access to the authorities. It is worth noting that, despite the documents' anonymization, they kept the original context.

3.4. Named-Entities Annotation

NLP tools and frameworks, such as those that use Named-Entities Recognition (NER) processing, take advantage of Named-Entities (NE) that have been identified and annotated, such as persons, locations, and license plates.

The crime-related documents were manually annotated by applying the XML template illustrated in Listing 3. The ''`element name`'' XML tags were used to extract the most relevant content of the documents. Some examples are `documentname`, `authors`, and `publicationdate`. The documents are available at GitHub repository, in the folder `/Data Set/NER/Criminal-Related Documents NE Annotated`.

Listing 3: Criminal-Related Documents Named-Entities Annotation - XML Schema.

```
<?xml version = ‘‘1.0’’ encoding = ‘‘UTF-8’’?>
<xs:schema xmlns:xs = ‘‘http://www.w3.org/2001/XMLSchema’’>
 <xs:element name = ‘‘DocumentID’’>
 <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:element name = ‘‘sentences’’ num = ‘‘xs:string’’ />
 <xs:element name = ‘‘SentenceID’’ desc = ‘‘xs:string’’ />
 <xs:complexType>
 <xs:element name = ‘‘entity type’’ type = ‘‘xs:string’’/>
 </xs:complexType>
 </xs:sequence>
 </xs:complexType>
 </xs:element>
</xs:schema>
```

Each sentence has an identification and a list of names and entities that are eligible to be annotated. Listing 4 depicts the sentences analyzed by the XML processing. In this example, the sentence 1 (identified by the tag `<sent1>`) has two annotated entities: a person name and a number.

Listing 4: Sentence XML Template.

```
<sent1 desc=‘‘Pedro Henriques tinha um plano e cumpriu-o em 24 horas
    ↪ .’’>
       <NE>
              <Person>Pedro Henriques</Person>
              <Numeric>24 horas</Numeric>
       </NE>
</sent1>
```

3.5. Narcotics Corpus

A specific corpus was built to accommodate the terms intrinsically related to narcotics in the Portuguese language. The following presumptions were made on the extraction of terms related with this type of crime:

- the narcotics are mentioned in their official designation as well as the one used on the *street*, through slang; and
- drug trafficking is one of the most reported and typified crimes investigated by the criminal police [12].

To the best of the authors' knowledge, there are no annotated texts related to the narcotics' crime domain in the Portuguese language. To overcome this limitation, a manual annotation was made, by labeling the correct entities using a narcotics list, with current official and street names. The corpus is available in the GitHub repository (folder `Data Set/NER/Narcotics`). It was built by extracting texts from daily newspapers and blogs that mention narcotics' terms. The sentence below illustrates how the documents have been annotated using the Apache OpenNLP (https://opennlp.apache.org/ (accessed on 25 June 2021)) tool notation.

In Portuguese:

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

In English:

There were also seized 5500 doses of <START: narcotics> liamba <END>, 323 plants of <START: narcotics> cannabis <END>, 16 doses of <START: narcotics> hashish <END> and 12 doses of <START: narcotics> MDMA <END> (better known as <START: narcotics> ecstasy <END>) and 930 euros.

3.6. 5W1H Annotation

This section introduces a dataset to help researchers that intend to use the 5W1H approach (Who, What, Whom, When, Where, How) to extract the semantic information from sentences [13]. This approach was introduced by Griffin [14] and is widely used in journalism. However, in criminal investigation, the same methodology is applied by the investigators, as they seek to answer the 5W1H questions to analyze the facts and further identify the criminals [15]. The 5W1H methodology provides facts about a criminal document, by answering the following questions:

- Who?: Who was involved?
- What?: What happened?
- When?: When did it happen?
- Where?: Where did it happen?
- Why?: Why did it happen?
- How?: How did it happen?

A set of documents was annotated by Portuguese domain experts, who used the corresponding five annotator tags in the XML template depicted in Listing 5.

Listing 5: 5W1H Annotation 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 = ‘‘EventID’’ type = ‘‘xs:string’’ />
 <xs:complexType>
 <xs:element name = ‘‘WHO’’ type = ‘‘xs:string’’/>
 <xs:element name = ‘‘WHAT’’ type = ‘‘xs:string’’/>
 <xs:element name = ‘‘WHERE’’ type = ‘‘xs:string’’/>
 <xs:element name = ‘‘WHEN’’ type = ‘‘xs:string’’/>
 <xs:element name = ‘‘WHY’’ type = ‘‘xs:string’’/>
 <xs:element name = ‘‘HOW’’ type = ‘‘xs:string’’/>
 </xs:complexType>
 </xs:sequence>
 </xs:complexType>
 </xs:element>
</xs:schema>
```

An annotation scheme was defined to extract useful 5W1H questions from the documents. This annotation scheme can be used by supervised learning algorithms. Several research works apply the 5W1H approach with an annotated corpus in the English language [13,16,17], reinforcing the need to have a similar corpus for the Portuguese language. The annotated documents are available in the GitHub repository, in the folder `Data Set/5W1H`.

4. Technical Validation

The corpora presented in this paper was evaluated by processing the machine learning models implemented on the Apache OpenNLP platform and a prototype developed in Java [9]. The perceptron algorithm was used to evaluate the named-entities recognition, the 5W1H extraction model, and the narcotics terms extraction. The results obtained with NE recognition are summarized in Table 5.

Table 5. Crime-related documents evaluation for named-entity recognition.

	Precision	**Recall**	**F1-Score**
Criminal News	0.846	0.659	0.712
PGdLisboa News	0.850	0.679	0.716
Criminal Investigation Report	0.728	0.829	0.771
Average	0.808	0.722	0.733

The experiments were conducted over the crime-related documents dataset and were supported by a prototype developed in Java. The results identify with an average precision of 0.808, recall of 0.722, and F1-score of 0.733, obtained with the processing of criminal news, PGdLisboa articles, and criminal investigation reports. These results illustrate the correctness of the classifier to identify the named-entities that are annotated in the dataset [9].

The *5W1H Information Extraction Method* was evaluated using a set of 20 crime-related documents, annotated by external contributors. Table 6 summarizes the performance evaluation obtained with the proposed set, namely precision, recall, and F1-score.

Table 6. 5W1H information extraction method evaluation.

Precision	Recall	F1-Score
0.732	0.634	0.653

The corpus for Narcotics was also evaluated with the ML perceptron algorithm, and the results are summarized in Table 7. The corpus was evaluated by applying a 10-fold stratified cross-validation method. The dataset was divided into 10 equal parts and, for each run, nine parts were used to train the model and the remaining one to test. The average results obtained for precision, recall, and F1-score are 0.784, 0.768, and 0.771, respectively. The experiments were made with the Apache OpenNLP platform, and the Narcotics corpus and scripts were uploaded into a GitHub repository (folder `Data Set/NER/Narcotics/Narcotics Classifier`) [9].

Table 7. Narcotics dataset evaluation with 10-fold cross-validation.

Precision	Recall	F1-Score
0.784	0.768	0.771

To the best of authors' knowledge, the exploratory dataset that was delivered and made available , is the first comprehensive approach to have a dataset in the Portuguese language related to the criminal domain. It should benefit the ML and NLP practitioners, on benchmarking models and frameworks on Portuguese language processing.

For future work, the dataset will be continuously updated with more anonymized criminal reports and online news articles. Curation tasks will be continuously applied to enrich the quality of the dataset and enhance the learning methods performance. Data curation encloses several challenging sub-problems to maintain a dataset available and with high quality to be used by data science researchers. These tasks are intrinsically related to: (1) the increasing volume of the dataset, which need to have a track changes log; (2) anomalous data detection, such as the personal identification number in a format $111-111-1111$ is anomalous when the other values are in the format $111111-1111$; and (3) the document update with the introduction of new named-entities relevant to domain.

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

Funding: The APC was financed by Polytechnic of Leiria.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are publicly available under a Creative Commons Attribution 4.0 International License, in the following GitHub repository: https://github.com/goncalofcarnaz/Annotated-Corpus-of-Criminal-Related-Portuguese-Documents (accessed on 25 June 2021) .

Acknowledgments: The authors acknowledge the facilities provided by Polytechnic of Leiria and University of Évora, for the support to this research.

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
CAPER	Collaborative information Acquisition Processing Exploitation and Reporting
CIR	Criminal Investigation Reports
CN	Criminal News
CWTDF	Combined Websites and Textual Document Framework
HTML	Hyper-Text Markup Language
LEA	Law Enforcement Agencies
ML	Machine Learning
MOSAIC	Multi-Modal Situation Assessment and Analytics Platform
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
PDF	Portable Document Format
PII	Personally Identifiable Information
XML	Extensible Markup Language

References

1. Prayudi, Y.; Sn, A. Digital chain of custody: State of the art. *Int. J. Comput. Appl.* **2015**, *114*, 975–8887. [CrossRef]
2. 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, 30 October–1 November 2007; Volume 1, pp. 131–138. [CrossRef]
3. 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.
4. 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.
5. 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.
6. 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. [CrossRef]
7. 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. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2014**, *8577 LNAI*, 275–280. [CrossRef]
8. Atkins, S.; Clear, J.; Ostler, N. Corpus design criteria. *Lit. Linguist. Comput.* **1992**, *7*, 1–16. [CrossRef]
9. Carnaz, G.; Nogueira, V.B.; Antunes, M. A Graph Database Representation of Portuguese Criminal-Related Documents. *Informatics* **2021**, *8*, 37. [CrossRef]
10. Carnaz, G.; Nogueira, V.B.; Antunes, M. Knowledge Representation of Crime-Related Events: A Preliminary Approach. In *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*; OpenAccess Series in Informatics (OASIcs); Rodrigues, R., Janousek, J., Ferreira, L., Coheur, L., Batista, F., Oliveira, H.G., Eds.; Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany, 2019; Volume 74, pp. 13:1–13:8. [CrossRef]
11. Wiedemann, G.; Yimam, S.M.; Biemann, C. A Multilingual Information Extraction Pipeline for Investigative Journalism. *arXiv* **2018**, arXiv:1809.00221.
12. Biabani, G. The Explanation Related to the Relationship between Drug Abuse and Crime. *Q. J. Soc. Dev. (Previously Human Dev.)* **2020**, *14*, 199–200.
13. Chakma, K.; Das, A. A 5w1h based annotation scheme for semantic role labeling of English tweets. *Comput. Sist.* **2018**, *22*, 747–755. [CrossRef]
14. Griffin, P.F. The correlation of english and journalism. *Engl. J.* **1949**, *38*, 189–194. [CrossRef]
15. Braz, J. *Investigaç ao Criminal*; Almedina: Coimbra, Portugal, 2013.
16. Das, A.; Ghosh, A.; Bandyopadhyay, S. Semantic role labeling for Bengali using 5Ws. In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010), Beijing, China, 21–23 August 2010; pp. 1–8.
17. Hamborg, F.; Lachnit, S.; Schubotz, M.; Hepp, T.; Gipp, B. Giveme5W: Main event retrieval from news articles by extraction of the five journalistic w questions. In *International Conference on Information*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 356–366.