Natural Language Processing Applied to Forensics Information Extraction With Transformers and Graph Visualization

Fillipe Barros Rodrigues, William Ferreira Giozza, *Senior Member, IEEE*, Robson de Oliveira Albuquerque, and Luis Javier García Villalba, *Senior Member, IEEE*

***Abstract*— Digital forensics analysis is a slow process mainly due to the large amount and variety of data. Some forensic tools help categorize files by type and allow automatization of tasks, like named entity recognition (NER). NER is a key component in many natural language processing (NLP) applications, such as relation extraction (RE) and information retrieval. The introduction of neural networks and transformer architectures in the last few years made it possible to develop more accurate models in different languages. This work proposes a reproducible setup to build a forensic pipeline for information extraction using NLP of texts. Our results show that it is possible to develop both NER and RE models in any language and tune its hyper-parameters to achieve state-of-art performance and build comprehensive knowledge graphs, decreasing the amount of time required for human supervision and review. We also find that solving this task in phases can further improve the performance, not only for digital investigation applications, but also for general-purpose information extraction and analysis.**

***Index Terms*— Digital forensics, named entity recognition (NER), natural language processing (NLP), relation extraction (RE), transformers.**

I. Introduction

TEXT data play an important role in every forensic analysis process. Most of the data in the cyber world is unstructured, consisting of texts, photos and videos. Even though the digital forensic process model is not standardized, there is an abstract-level consensus surrounding it. Kohn *et. al* [1] proposed an overview of the most significant models described over the years, consisting of six processes: *documentation, preparation, incident, incident response, digital forensic investigation*, and *presentation*. This article focuses on improving the efficiency of the digital forensic investigation process (as shown in Fig. 1), which is based on several examination and analysis subprocesses. These processes require a great amount of time and effort to be completed and the process as a whole could benefit from a structured computational solution.

In this context, natural language processing (NLP) algorithms became a relevant approach to deal with such huge and diverse volume of data and to extract useful insights [2]. More specifically, named entity recognition (NER) systems have been adopted by several languages to extract entities such as locations, organizations and people. These entities alone can lead to the identification of key elements of an investigation, for example. However, a more refined analysis could benefit from the extraction of relations between entities, like the member of an organization, a list of relatives, or the location a crime took place.

The recent advances in deep neural networks have enabled researchers to develop more powerful NLP models based on BERT [3], which can then be fine-tuned for better performance in specific tasks. Multiple tasks can benefit from this, including NER and RE. With a couple of thousand sentences, a new NLP model can be finetuned from a pretrained BERT-based model in any language, overcoming the performance of traditional approaches that do not use neural networks.

Training a new NER or RE model usually requires many annotated data, following either a supervised learning paradigm or a distantly supervised learning paradigm.

Manuscript received October 9, 2021; revised December 12, 2021 and February 8, 2022; accepted February 12, 2022. This work was supported in part by the Professional Postgraduate Program in Electrical Engineering of University of Brasília, in part by the HEROES Project, and in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant 101021801. The work of William Ferreira Giozza and Robson de Oliveira Albuquerque was supported in part by the FAP-DF—Brazilian Federal District Research Support Foundation under Grant 00193-00000853/2021-28, in part by the Brazilian National Council for Scientific and Technological Development—CNPq under Grant 312180/2019-5 PQ-2, in part by the National Institute of Science and Technology (INCT) on Cybersecurity under Grant 465741/2014-2, in part by the Brazilian Ministry of the Economy under Grant 005/2016 and Grant 083/2016, in part by the Administrative Council for Economic Defense—CADE under Grant 08700.000047/2019-14, in part by the General Attorney of the Union—AGU under Grant 697.935/2019, in part by the National Auditing Department of the Brazilian Health System—DENASUS under Grant 23106.118410/2020-85, and in part by the General Attorney's Office for the National Treasure—PGFN under Grant 23106.148934/2019-67. *(Corresponding author: Luis Javier García Villalba.)*

Fillipe Barros Rodrigues and William Ferreira Giozza are with the Professional Post-Graduate Program in Electrical Engineering, Department of Electrical Engineering, University of Brasília, Brasília 70.910-900, Brazil (e-mail: fillipefbr@gmail.com; giozza@unb.br; robson@redes.unb.br).

Robson de Oliveira Albuquerque is with the Professional Post-Graduate Program in Electrical Engineering, Department of Electrical Engineering, University of Brasília, Brasília 70.910-900, Brazil, and also with the Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Universidad Complutense de Madrid (UCM), 28040 Madrid, Spain (e-mail: robson@fdi.ucm.es).

Luis Javier García Villalba is with the Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Universidad Complutense de Madrid (UCM), 28040 Madrid, Spain (e-mail: javiergv@fdi.ucm.es).

Digital Object Identifier 10.1109/TCSS.2022.3159677

Collection
Examination
Analysis
Reporting

Fig. 1. Digital forensics investigation process.

Some preprocessing techniques are often necessary to prepare the data and achieve better results. Nevertheless, there are tools that already offer functional models for several languages and use cases, like SpaCy [4]. To build a customized workflow for specific applications or goals, these tools may not be sufficient nor achieve the desired results in practice. Therefore, a more personalized approach is useful in such cases.

In this work, it is proposed a knowledge-based information retrieval system that combines NER and RE models enhanced by neural networks and transformers architectures. These models with a baseline setup were compared with new ones obtained from hyper-parameters tuning for English and Portuguese. Preliminary results showed that it was possible to improve a Portuguese NER model on the Paramopama [5] corpus by 2%. By outputting entities and their relations in the form of an interactive graph, it was demonstrated how the pipeline can help to automate the information extraction analysis, not only in the fields of digital forensics and digital investigation, but also for big data analytics and business intelligence applications.

Despite the good preliminary results achieved, the performance of the models proposed in this work is deeply dependent on the quality of the datasets used for training. Additional cleaning and other preprocessing steps are usually required to start developing a good NLP model, and these steps may demand great effort and time. Besides, some scenarios and domain-specific data are harder to process, like social media content and chat conversations. The results presented in this article for different application scenarios considered structured texts, without any typography or grammar errors, which may not be sufficient for some other applications. Long texts tend to present complex relations between named entities, which are not always in the same sentence, making it difficult to predict and extract semantic relationships. Extracting information for multilingual texts is also a great challenge. In this work, English and Portuguese NLP models and applications were prioritized over other languages, mainly due to the relevance of the English language worldwide and the lack of support for Portuguese in many NLP applications.

The rest of this article is structured as follows. In Section II, it is introduced the most relevant related work regarding NER, relation extraction (RE) and information extraction systems. Then, in Section III, we discuss the steps that make up our proposed IE pipeline. Preprocessing steps used for training and NLP models are described in Section IV, including the parameters used for training both NER and RE models. Section V shows the relation schemas adopted in this work, including node and relationship types. A baseline setup, alongside some experiments with transformers for NER and RE, is presented in Section VI. Results are presented and discussed in Section VII. Finally, Section VIII presents the conclusions of this research.

II. Related Work

Currently, BERT-based architectures have become a standard for several natural language processing (NLP) applications, including NER and RE, for instance.

A. Digital Forensics

Nowadays, huge amounts of heterogeneous data have become the new normal all over the cyberworld. Due to this fact, investigating cybercrimes is a task undoubtedly difficult and time consuming. Caviglione *et al.* [6] explore the future of digital forensics and discuss how to maintain societies secure and pursue criminals effectively. Similar aspects are discussed by Ukwen and Karabatak [7], addressing a review of NLP-based systems in digital forensics and cybersecurity, to serve as a basis for researchers and practitioners in these fields and also to provide a roadmap for the future.

In this context, technology has been an important ally in the fight against crime. In particular, some softwares help investigators and security experts save time and be more productive, like the IPED (digital evidence processor and indexer) tool [8]. IPED is an open-source software developed by digital forensic experts from Brazilian Federal Police that can be used to process and analyze digital evidence, often seized at crime scenes by law enforcement or in a corporate investigation by private examiners. It offers several functionalities for digital forensics, like language detection, signature analysis, audio transcription, and NER, for example. However, some modules do not yet support other languages or use cases, like the NER module, which does not offer support for Portuguese out of the box. We hope that the results of our work will serve as a basis for the integration of new NLP models in the tool, in particular to meet the demand for text analysis and information extraction in Portuguese.

Van Baar *et al.* [9] explore different implementations of the digital forensic process and analyze factors that impact the efficiency of this process. They point out that digital investigators should not be tasked with system administrative tasks. In the traditional process, they are responsible for the entire investigation environment, which leads to a lot of administrative overhead. Besides, digital investigators are either underqualified or overqualified for many of the tasks they often perform. That is why it was proposed a Digital Forensics as a Service (DFaaS) model. In the DFaaS setup, digital investigators focus on the forensics tasks (seizing material and extracting data from it), whereas the extracted data is sent to a centralized system that automatically extracts informations from the data and give this information back to investigators and analysts (which also proves to be effective especially in the era of big data [10]). Nevertheless, their proposed systems lack descriptions on specific tasks that could be automated by the centralized system. For instance, no details of machine learning solutions or NLP models were mentioned. DFaaS is a new way of working and cooperating, particularly for governmental organizations, that can be adapted to support artificial intelligence, dynamic reporting, and other solutions [11].

B. Named Entity Recognition

Junior *et al.* [5] presented a new NE-tagged corpora for Brazilian Portuguese named Paramopama, due to the lack of corpus for this language. They also evaluated the quality of the dataset by measuring precision, recall, and F-measure of a NER classifier trained on this corpus. Their results show that their dataset has yielded better results than other well-known Portuguese NE-tagged corpora. However, by the time of their publication, there was not a BERT transformer that could be used to boost performance.

Schmitt *et al.* [12] compared the performance of some well-known NER softwares on the market, such as Stanford NLP, NLTK, Spacy, and OpenNLP. Their goal was to address the difficulty encountered by NLP practitioners to clearly and objectively identify which software performs the best, due to the lack of transparency preventing the reproducibility of experiments. The comparison was limited to fewer entity types and languages, mainly English.

Considering the advent of such architectures, in [13] it is proposed a methodology to improve the performance of NER systems in every language. They made experiments over five different datasets, using two high-resourced languages (English and Spanish) and three low-resourced languages (Croatian, Slovene, and Finnish), managing to improve the state-of-the-art F-Score for them. However, this work did not contemplate how to tune hyper-parameters for the models nor how to preprocess the datasets.

In [14], it is presented a new NER dataset for Brazilian Portuguese legal texts, consisting of six entity types: person, legal cases, time, location, legislation, and organization. The authors also presented a model for NER trained over a long short-term memory and conditional random fields (LSTM-CRF) architecture with the LeNER-Br dataset, which achieved an average F-Score of 86%. However, their corpus is domain-specific and does not work very well for general-purpose NER applications.

New transformer-based state-of-the-art NER models have emerged in recent years. FLERT [15] introduces a way to capture document-level features to enhance performance by modeling document context. Another approach is proposed by [16], which uses an Automated Concatenation of Embeddings (ACE) mechanism to extract better pretrained contextualized embeddings of word representations for structured prediction tasks. FLERT is a strong baseline for NER, while the ACE model is the current state-of-the-art for NER in the CoNLL03 benchmark [17], with a F-Score of 94.6%.

C. Relation Extraction

Early efforts in RE had focus on predicting relationships between entities within a given sentence by modeling the possible interactions [18]. This approach does not consider interactions between sentences or distant entities in the text. Extracting relations on document-level is a considerably more difficult task, since several sentences and their relations must be considered [19]. Nevertheless, both approaches are useful in different scenarios and sentence-level relation extraction is usually easier and faster to implement as part of an NLP system [20].

In [20] it is proposed BERT-based models for RE, demonstrating how these models can achieve the state-of-the-art performance without external features, serving as the basis for future work on other downstream tasks. However, they did not explore further improvements for the models on the experimental setup, as such for example feature engineering techniques and hyper-parameters tuning.

Han *et al.* [21] created an open-source and extensible framework to implement neural models for RE. Thanks to their design pattern, they showed it is easy to extend the production-ready available models that come with the framework and to train custom models based on other datasets in any language. Since the focus of their work was on the RE part, they did not release details on how to extract the entities in the text or to build a customized NER system, leaving that part for the users.

Lately, different models have achieved state-of-the-art performance on relation classification benchmarks. In particular, the work of [22] introduces an improved baseline for relation classification by adopting new ways of representing entities. More specifically, the authors show that the typed entity marker (TEM) entity representation technique yields better results when compared with the entity marker (EM) technique. This approach outperforms another strong baseline proposed by [23], that introduced a new training setup called *matching the blanks* (MTB), which relies on entity resolution annotations.

A new paradigm for relation classification is proposed by Lyu and Chen [24], who argue that the existing methods regard all relations as the candidate relations for a pair of entities, neglecting the restrictions on candidate relations by entity types. Thus, they propose a new model, called RECENT, which exploits entity types to restrict candidate relations. This is the current state-of-the-art RE model evaluated on the TACRED benchmark [25], with a F-Score of 75.2%.

D. NLP With BERT

The era of social networks and big data made room for several NLP applications in the last years. More specifically, automatic text summarization became a fundamental step to extract significant information from different sources. Ma *et al.* [26] created a topic-aware extractive summarization model based on BERT, which is able to gather contextual representations to explore topic inference, generating consistent topics with the state-of-the-art results.

In [27] it is presented a BERT-based model that aims to identify speech acts as means to comprehend the communicative intention of a speaker. The proposed model achieved an accuracy of 77.52%, outperforming other baselines approaches. Nevertheless, this work focused on developing a speech act for Twitter, assessing the content and intent of tweets, and not extending their analysis for other domains.

Polignano *et al.* [28] trained a BERT language understanding model for the Italian language (AlBERTo), focused on the language used in social networks, specifically on Twitter. The model is able to evaluate three sentiment analysis sub-tasks for the Italian language:

1) *Subjectivity Classification*: To decide whether a given message is subjective or objective;
2) *Polarity Classification*: To decide whether a given message is of positive, negative, neutral, or mixed sentiment;
3) *Irony Detection*: To decide whether a given message is ironic or not.

Despite achieving good results, AlBERTo was designed with a very specific choice: tweets in the Italian language. Therefore, this model does not perform well on other NLP tasks, nor on other languages.

The work of [29] introduces BioBERT, which is a domain-specific language representation model pre-trained on large scale biomedical corpora. With almost the same architecture across tasks, the authors show that BioBERT outperforms BERT in a variety of biomedical text mining tasks, including NER (0.62% F1 score improvement), biomedical RE (2.80% F1 score improvement), and biomedical question answering (12.24% mean reciprocal rank improvement.) Similar to other domain-specific NLP models, BioBERT is restricted to some applications, evidently, in this case, to the biomedical field and related areas.

The work of [30] demonstrates the application of BERT to coreference resolution, achieving strong improvements on this task for the English language. According to their experiments, the authors state that BERT-large is particularly better than BERT-base at distinguishing between related but distinct entities, even though there is still room for improvement in modeling document-level context to deal with spread-out clusters. Another improvement for the coreference resolution task is proposed by [31], via an "entity equalization" mechanism. The key element of their approach is to capture properties of entity clusters and use those in the resolution process using BERT embeddings. Future work for their approach also includes the plan to further enrich these representations by considering information from across the document.

SpanBERT [32] is a pretrained language model based on the transformer. It extends BERT by incorporating a training objective of span prediction and achieves improved performance on coreference resolution (CR) and RE. Another model, LUKE [33], pretrains the language model on both large text corpora and knowledge graphs (KGs). It adds frequent entities into the vocabulary and proposes an entity-aware self-attention mechanism. LUKE achieves the state-of-the-art performance on several entity-related tasks, such as NER and RE. A similar language representation model, ERNIE, was proposed by [34]. It incorporates KGs to enhance language representation with external knowledge, establishing a good baseline for knowledge-driven tasks. However, it still underperforms LUKE and other state-of-the-art models. The work of [35] presented the first public large-scale pretrained language model for English Tweets, named BERTweet. It achieves the state-of-the-art performance on Tweet NLP tasks, including NER and text classification.

III. Methodology

We propose in this work a pipeline for information extraction consisting of several components. The architecture of the proposed pipeline is shown in Fig. 2. Each one of the components is described as follows.

A. Step 1: Preprocessing

The first step of the information extraction system is called *preprocessing* and it refers to the tasks made prior to processing the input data through the pipeline. In our work, this step corresponds to *data collection* and *model training and fine-tuning*.

The data collection task refers to the acquisition of the text data that will be processed by the information extraction pipeline in the next steps. The data may come from a variety of sources, including databases, websites, chats, or documents, to name a few. A forensics approach can be used here to collect this data, and it is recommended to clean the data as much as possible, removing unnecessary information and reviewing the texts to generate good results by the end of the pipeline IE process.

The model training and fine-tuning task refers to the training and fine-tuning of the NLP models used in the IE pipeline. The main NLP models proposed in the next steps are task-specific, including models for coreference resolution, NER, and RE. It is recommended to train and fine-tune NLP models for a specific domain or application to achieve the best possible results. The models' language is also a pertinent factor that should be considered during this task.

B. Step 2: Text Input Data

The second component of the IE system is characterized by text input data (Fig. 2). This data can be anything, from news articles to social media chats, for example. There is not a limitation for the language type present in the input texts. However, some NER and RE models were trained on specific languages only, and therefore may not perform as well in other languages. Some multilingual models can be used to overcome this limitation, as we will discuss later. Ideally, the input data must be free of typographical and grammatical errors to output better results in the subsequent steps.

C. Step 3: Named Entity Input and Coreference Resolution

Step 2 of the processing pipeline as shown in Fig. 2 consists of two optional components: named entity input (NEI) and CR. NEI can help improve the model output by forcing it to take into account some entities considered relevant for the task at hand. Since the NER models have limited and frequently different labels for the entities they recognize, it may be important to include some entities manually. Moreover, CR may help improve the accuracy of the RE model in Step 5, since the identification of antecedent chains can output cleaner relation pairs and less ambiguous text [36].

D. Step 4: Named Entity Recognition Model Selection

In Step 4 (Fig. 2), we select one of our trained NER models to be used for Entity Extraction in Step 5. The models differ from each other in the language they were trained on, in the types of entities they are able to recognize and in

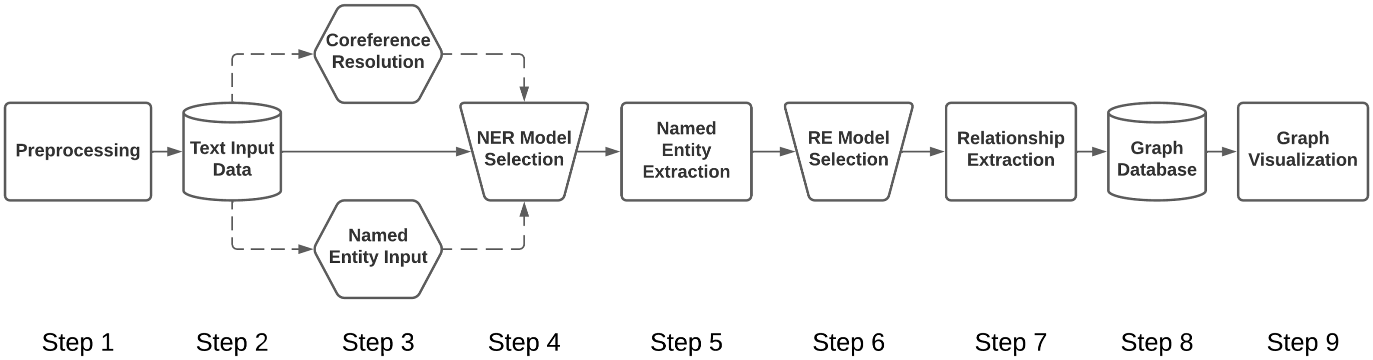


Fig. 2. Information extraction pipeline architecture.

their performance. All NER models were fine-tuned on BERT and its derivatives, like ELECTRA [37] and RoBERTa [38], for example. The model selection is strictly related to the data's input language from Step 1. If the input language is known beforehand, it is recommended to select a NER model fine-tuned for that particular language, given the fact that non-multilingual models may not perform well on different languages. However, if the input language is unknown, or if it is composed of multilingual documents, a multilingual model may perform well for most languages.

E. Step 5: Named Entity Extraction

Once the underlying NER model is selected, entities in the input data are recognized and extracted in Step 5 (Fig. 2), generating a list of all entities that will later be used to feed the RE models. Usually, some cleaning process is applied in this step with the intent of removing duplicated entity mentions from the final entities list. This is desirable to generate a more concise and efficient graph visualization. However, it also may be desirable to keep a record of how many times the same entity was mentioned in a document or across documents, for example.

F. Step 6: Relation Extraction Model Selection

Step 6 (Fig. 2) consists of choosing an appropriate RE model to be applied over the named entities extracted from the previous step. Three RE models were used, including two for English texts and one for Portuguese texts. Similar to NER models, they differ from each other in the types of relations they can recognize. Therefore, it is crucial to select a RE model for the target language, since for this step, no multilingual RE models were conceived. Besides, different RE models have different sets of predefined relations. This means that applying different RE models over the same input text may output different relations between entities in the text, some of which may be detected by one of them and not by the other, or they both may detect the same relation but with different names or confidence scores. It also means that this is an interactive process and most of the time it may be desirable to apply more than one RE model over the same text input in order to combine outputs and extract more relevant informations.

G. Step 7: Relationship Extraction

After choosing an appropriate RE model, the relations between entity pairs are extracted in Step 7 (Fig. 2), generating a list of entities and their corresponding relations, if any. The RE output for a pair of entities is usually represented in the form of a triplet: `(e1, rel, e2)`, where `e1` is the source Entity, `e2` is the target entity, and `rel` is the predicted relation between them. This output format facilitates the construction of a directed graph; however, it is necessary to note that some relations only exist in one direction. For example, in the sentence "John is the father of Daniel," there are two possible outcomes: `(John, father, Daniel)` and `(Daniel, parent, John)`. A RE model may have these two predefined relations, whereas others may have only one of them. Therefore, it is important to consider both directions `(source entity <-> target entity)` and decide whether the output is relevant for both of them, only one of them or none of them. Some RE models are able to detect when there is no relation between a pair of entities at all, while others will output some predefined relation with a low confidence score. This characteristic is related to how the model was trained and with which data and predefined relations set.

H. Step 8: Graph Database

Step 8 (Fig. 2) corresponds to the storage process for the entities and relations extracted from the previous steps. This data is usually stored in a graph database, but it can be as easily stored in traditional relational and nonrelational databases. The data format is flexible and it may be defined by the interested user. A JavaScript Object Notation (JSON) is usually a good option format since it can be exported for several different applications and it is extensively used in web applications.

I. Step 9: Graph Visualization

Step 9 (Fig. 2) is the output of the proposed system, in the form of an interactive graph, in which nodes represent named entities and edges represent the relations between them. A graph visualization helps analysts and investigators with an overview of all relevant insights obtained from the input data, making it easy to create filters, plots, and detailed reports.

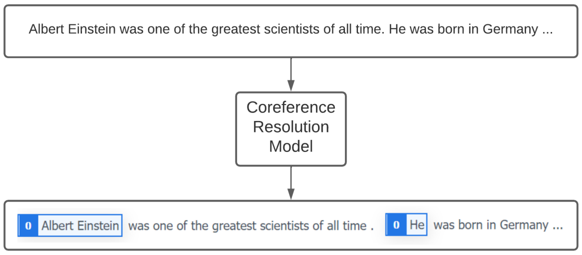

Albert Einstein was one of the greatest scientists of all time. He was born in Germany ...
Coreference Resolution Model
0 Albert Einstein was one of the greatest scientists of all time .
0 He was born in Germany ...

Fig. 3. CR example.

For consistency purposes, it is important to have a set of predefined relations between named entity types. In this way, it is easier to search similar relationships across different documents mentioning related entities and plot them together in the same graph, for instance.

IV. Preprocessing

To achieve the output shown in Fig. 2 as a graph visualization, several NLP steps had to be implemented. Fig. 3 shows an example of how the coreference model works with a text input.

A. Coreference Resolution

The model essentially searches the text based on parsing and morphological dependencies to replace proper nouns and possessive pronouns with their respective terms. In the example of the Fig. 3, “He,” in the second phrase, was replaced by its original reference, “Einstein.” For this part, we used Neuralcoref [39], which uses the corpora of the CoNLL-2012 shared task [40] and a neural net scoring model described in [41]. Since the original repository only includes production-ready support for the English language, applications of CR were not considered in this work for other languages.

However, it is possible to train a CR model based on [39] for other languages, provided an annotated dataset in the target language and some modifications on the loading and parsing scripts. For Portuguese, corpora [42] and [43] are good options to train a new coreference model. Further experiments to expand the CR model to support Portuguese were not considered in this work due to the fact that it would require a significant understanding of the language parse trees to reach an acceptable identification of mentions, which is out of our scope.

B. Named Entity Input

NEI can help improve the model output by forcing it to take into account some entities considered relevant for the task at hand. Since the NER models have limited and frequently different labels for the entities they recognize, it may be important to include some entities manually. Spacy, for example, enables the use of a custom Entity Ruler based on a token-level or document-level matcher to identify and label entities in the input text. Traditional regular expressions matching is also supported, making it easy and often useful to highlight words of interest within the text.

TABLE I
Datasets Used for NER

Dataset	Language	Domain	Sentences
First HAREM [46]	Portuguese	General	5,000
Mini HAREM [46]	Portuguese	General	1,000
Second HAREM [47]	Portuguese	General	3,500
Paramopama [5]	Portuguese	Wikipedia	12,500
LeNER-Br [14]	Portuguese	Legal	10,392
WikiNER [48]	Portuguese	Wikipedia	125,821
CoNLL03 [17]	English	News	22,137
WNUT17 [49]	English	Social Media	5,690

TABLE II
RE Models

Name	Language	Sentences	Entities	Relations
DBPedia [51]	Portuguese	91,914	3	9
Wiki [52]	English	56,000	4	25
TACRED [25]	English	106,264	5	42

C. Named Entity Recognition Corpora

Regarding the NER model training, a couple of pretrained BERT derivative transformers were used, to fine-tune them with custom datasets for this matter. Table I shows some information about the datasets used for NER fine-tuning [44], [45].

As shown in Table I, the datasets used for NER were in two languages: Portuguese [44] and English [45]. We chose to focus on these languages because there are not many production-level solutions for Portuguese NER yet, we also added English because it is widely used worldwide and because of its relevance in texts over the Internet and social media.

For First HAREM, Mini HAREM, Second HAREM, and Paramopama corpora, the training, development, and test splits were 60%, 20%, and 20%, respectively. For the others (LeNER-Br, CoNLL03 and WNUT17) we adopted the splits from their original sources.

The tagging scheme used for NER corpora was the IOB scheme [50], in which all data files contain one word per line with empty lines representing sentence boundaries. At the end of the line, there is a tag that expresses whether the current word is the beginning of an entity (B), inside an entity (I), or not an entity at all (O). Here it is an example sentence

Albert	B-PER
Einstein	I-PER
was	O
in	O
Germany	B-LOC
.	O.

D. Relation Extraction Corpora

For the RE stage, we used three different models, as shown in Table II.

The framework behind RE operates over four different possible approaches: sentence-level RE, bag-level RE, document-level RE, and few-shot RE. Han *et al.* [21] explain the fundamental differences between each approach. In our work,

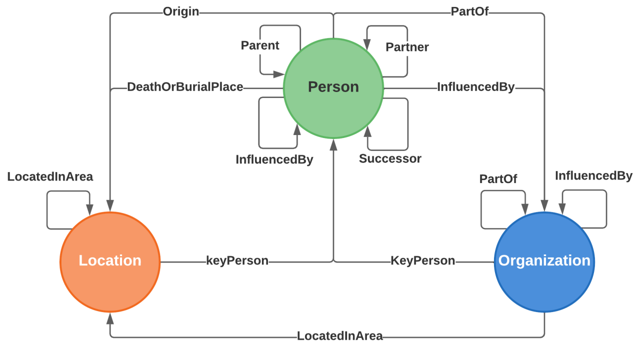


Fig. 4. DBPedia relations schema (adapted from [51]).

we prioritized sentence-level RE since our training dataset was based on sentences, besides this being the primary approach adopted by [21]. However, it is simple to select another option based on the characteristics of the data being analyzed.

For the DBPedia corpus [53], the training, Validation, and test split were 60%, 20%, and 20%, respectively. For the Wiki corpus [52], we used the splits provided by [54], and the splits for TACRED are described in [25].

The data format required to train a RE model with [54] is a text file in which each line is a JSON (JavaScript Object Notation) object containing a list of tokens, two entities with their respective position indexes in the sentence and the relation label between them. Both Wiki and TACRED models were trained on English data, whereas the DBPedia model was developed over a Portuguese corpus. Figs. 4–6 show the possible relationships between entity types for DBPedia, Wiki, and TACRED, respectively.

V. Relations Schemas

In this section, we present the relations schemas used in our work. We created three schemas, each of which is represented in the form of a graph, and developed a model based on each one (Table II).

A. DBPedia Relations Schema

For our DBPedia schema, as shown in Fig. 4, there are nine different types of predefined relationships (including a *other* class) between three node classes: person, organization, and location.

B. Wiki Relations Schema

Our Wiki schema (Fig. 5) presents 25 relationship classes and four node classes (person, organization, location, and miscellaneous).

C. TACRED Relations Schema

Our TACRED schema consists of 42 relation classes (including a *no_relation* class) and five entity classes as nodes (person, organization, location, date, and miscellaneous), as shown in Fig. 6. The original dataset [25] contains other entity classes (`CITY`, `COUNTRY`, and `NUMBER`, for example) that were converted to one of our proposed entity classes for practical applications. In Section VI-B, we evaluate the model's performance based on this schema. However, in Section VII-E, we present results based on the original version of TACRED for a fair comparison with other models.

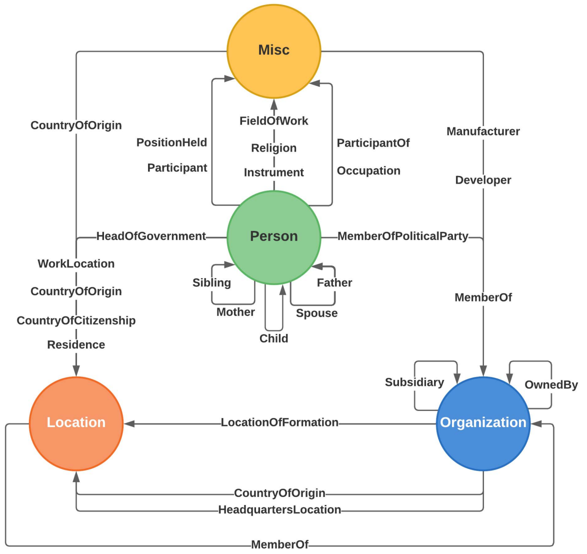


Fig. 5. Wiki relations schema (adapted from [52]).

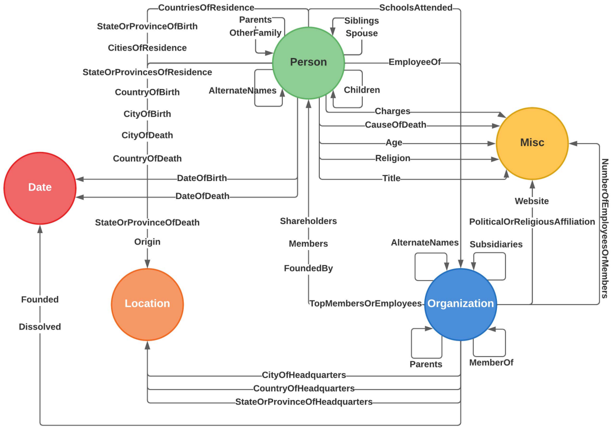


Fig. 6. TACRED relations schema (adapted from [25]).

VI. Baseline Setup

This section presents the main results for NER and RE experiments using common hyper-parameters for training.

A. NER Results

To establish a performance comparison between the datasets used for NER fine-tuning and different transformer models, we made an experiment with some common hyper-parameters. Table III shows the hyper-parameters used for training the models, Table IV shows the overall results for Portuguese models, and Table V shows the overall results for English models, whereas Fig. 7 compares Precision, Recall, and F-Score values for Portuguese NER systems and Fig. 8 compares Precision, Recall, and F-Score values for English NER systems, respectively.

TABLE III
HYPER-PARAMETERS VALUES USED TO TRAIN THE NER MODELS

Hyper-parameter	Value
Number of epochs	20
Early stop patience	1600
Scheduler	Linear with warm-up
Dropout rate	0.1
Batch size	128
Optimizer	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Learning rate	2×10^{-5}
Warmup steps	250
Total steps	20000
Training Eval frequency	200
Clipping gradient norm	1.0

TABLE IV
OVERALL RESULTS FOR NER SYSTEMS FOR PORTUGUESE

System	Corpus	Transformer	P[1]	R[1]	F[1]
FHBPT	First H.[2]	BERT-PT[3]	77.9	82.0	79.9
MHBPT	Mini H.[2]	BERT-PT[3]	80.8	78.1	79.4
MHBM	Mini H.[2]	BERT-M[3]	75.6	74.9	75.2
MHDBM	Mini H.[2]	DistilBERT-M[3]	73.3	69.6	71.4
SHBPT	Second H.[2]	BERT-PT[3]	84.5	86.9	85.7
PBPT	Paramopama	BERT-PT[3]	88.9	89.2	89.0
LBPT	LeNER-Br	BERT-PT[3]	90.3	88.2	89.2
WBPT	WikiNER	BERT-PT[3]	90.5	90.9	90.7

[1] P = Precision; R = Recall; F = F-Score.
[2] First H. = First HAREM; Mini H. = Mini HAREM;
Second H. = Second HAREM;
[3] BERT-PT = BERT Portuguese [55]; BERT-M = BERT Multilingual [3];
DistilBERT-M = DistilBERT Multilingual [56].

TABLE V
OVERALL RESULTS FOR NER SYSTEMS FOR ENGLISH

System	Corpus	Transformer	P[1]	R[1]	F[1]
CB	CoNLL03	BERT [3]	90.9	90.7	90.8
CDB	CoNLL03	DistilBERT [56]	89.0	89.5	89.3
CR	CoNLL03	RoBERTa [38]	91.8	92.4	92.1
CDR	CoNLL03	DistilRoBERTa [56]	91.2	91.6	91.4
CE	CoNLL03	ELECTRA [37]	90.5	91.4	91.0
WB	WNUT17	BERT [3]	58.7	32.9	42.0
WR	WNUT17	RoBERTa [38]	56.8	42.0	48.3
WE	WNUT17	ELECTRA [37]	57.6	38.0	45.8

[1] P = Precision; R = Recall; F = F-Score.

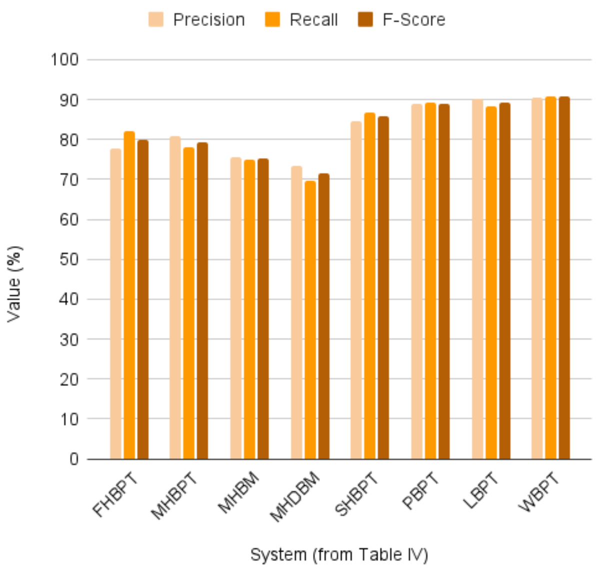


Fig. 7. Precision, recall, and F-Score output for Portuguese NER systems.

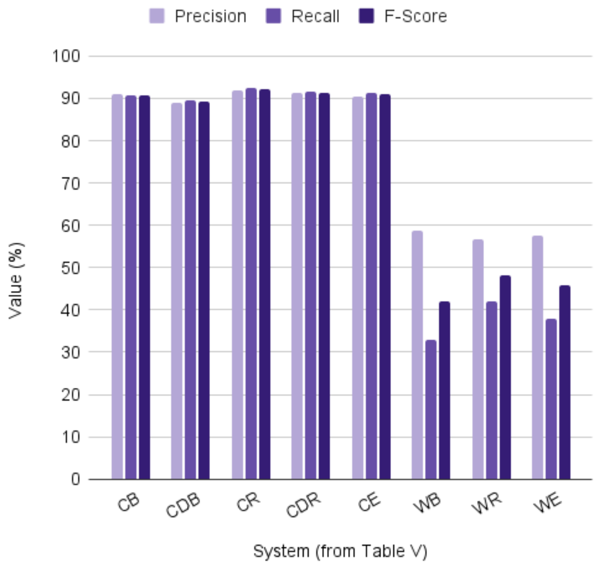


Fig. 8. Precision, recall, and F-Score output for English NER systems.

As we can see from the results shown in Tables IV and V, the best models for Portuguese were the ones trained on WikiNER and Paramopama datasets, with F-Scores of 90% and 89%, respectively. For English, the best ones were based on RoBERTa and DistilRoBERTa transformers over the CoNLL03 corpus, with F-Scores of 92% and 91%, respectively.

The results may vary depending mainly on the corpus used for training and the base transformer adopted. For example, we chose to present the WNUT17 [49] corpus since it is composed of social media texts, containing many words and expressions that are not present on the other datasets. This allows us to obtain a model that performs better on NER for social media texts, even though it may not perform as well in other domains. The same occurs with the LeNER-Br corpus, which is based on brazilian legal documents, containing entities that only make sense in such texts.

In practice, apart from the final scores, choosing the best model requires knowledge about the nature of the input text. Some corpus allow flexibility for several domains, while others achieve the state-of-the-art results for a specific domain.

B. Relation Extraction Results

The hyper-parameters used for training the DBPedia and Wiki RE models, presented in Table VI, yielded the results shown in Table VII, with a micro F1 score of 84%, 88%, and 79% on the test sets of the DBPedia RE model, the Wiki RE model and the TACRED RE model, respectively. All models were trained based on Sentence-Level RE, with a BERT-based transformer for the target language.

TABLE VI
HYPER-PARAMETERS VALUES USED TO TRAIN THE RE MODELS

Hyper-parameter	Value
Number of epochs	3
Batch size	64
Max sentence length	128
Optimizer	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Learning rate	2×10^{-5}
Metric	Micro F-Score
Relation representation	Entity Marker

TABLE VII
OVERALL RESULTS FOR RE SYSTEMS

Dataset	Transformer	Precision	Recall	F-Score
DBPedia	BERT-Base-PT	87.3	82.1	84.6
Wiki	BERT-Base	91.6	84.8	88.1
TACRED	BERT-Base	81.5	76.6	79.0

TABLE VIII
TOOLS USED FOR EACH NLP TASK

Task	Tool	Version
Digital evidence collection	IPED [8]	3.18.9
Visualization of coreference chains	AllenNLP [57]	2.1.0
Coreference resolution	Neuralcoref [39]	4.0.0
Named entity recognition	SpaCy [4]	3.2.0
Relation extraction	OpenNRE [54]	0.1
Graph database and graph visualization	Neo4j Desktop [58]	1.4.8
Models training and fine-tuning	Google Colab [59]	Free [1]
Pipeline application	Avell computer [2]	Ubuntu 20.04.2 LTS

[1] Intel Xeon CPU @2.30GHz, 13GB DDR4 RAM and a 12GB GDDR5 NVIDIA Tesla K80 GPU.

[2] Intel Core i7-7700HQ CPU @2.80GHz, 32GB DDR4 RAM and a 4GB GDDR5 NVIDIA GeForce GTX 1050 Ti GPU.

VII. RESULTS AND DISCUSSION

This section summarizes the main topics of our work, including the results of hyper-parameters tuning for Portuguese named entity recognition and examples of the pipeline application for a complete flow of information extraction for English and Portuguese. Table VIII shows the tools and versions used for each NLP task.

A. Hyper-Parameters Tuning

To achieve better results and an optimal model for NER in Portuguese, we ran some tests with different hyper-parameters. English models were not considered for this task since they already show the state-of-the-art results for many transformer architectures and different annotated corpora. In this sense, our focus was to choose a Portuguese corpus that allows us to run multiple tests and yet achieve good performance with a reasonable amount of time required for training.

TABLE IX
TUNED HYPER-PARAMETERS FOR NER IN PORTUGUESE

Hyper-parameter	Value
Transformer	BERT-Base-PT
Number of epochs	30
Training max steps	30000
Dropout rate	0.5
Batch size	128
Learning rate	2×10^{-5}
Warmup steps	150
Total steps	20000
Training Eval frequency	200

TABLE X
HYPER-PARAMETERS TUNING RESULTS

Entity	Precision	Recall	F-Score
Location	94.82	94.28	94.55
Organization	68.57	80.53	74.07
Time	86.89	90.64	88.73
Person	88.09	96.10	91.92
Overall	**90.60**	**92.59**	**91.59**

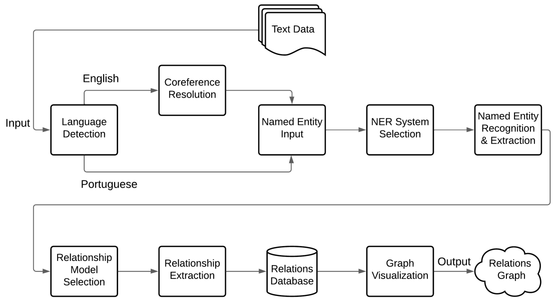


Fig. 9. Pipeline application example flow.

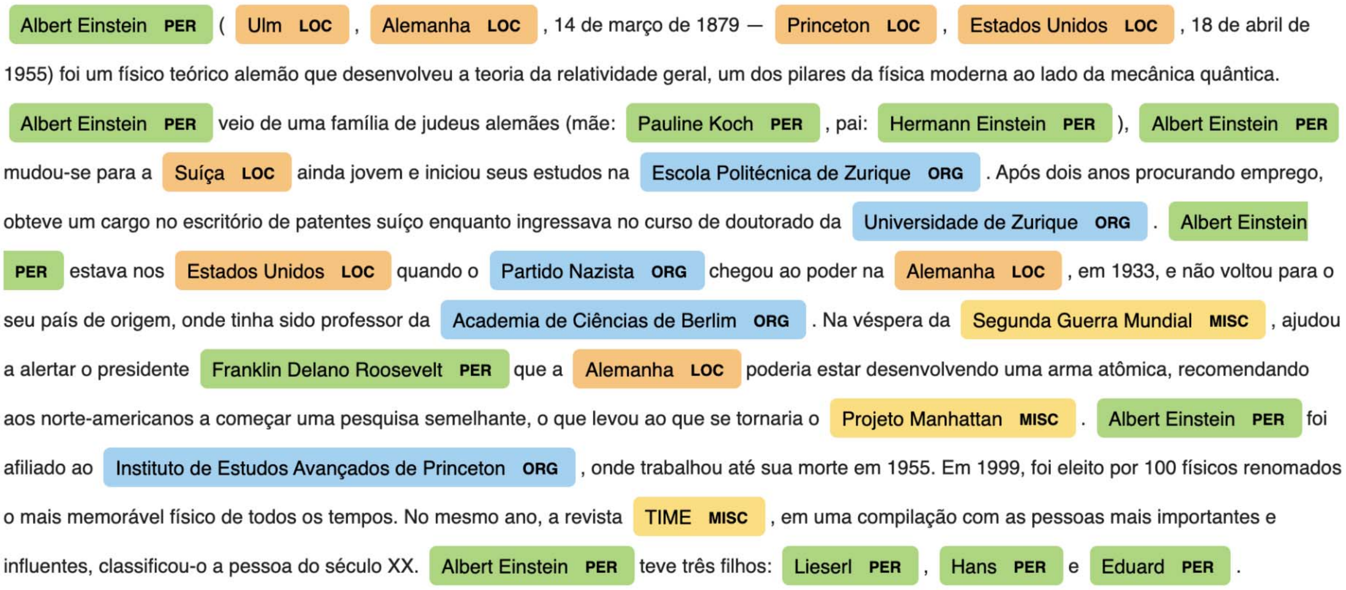
We chose the Paramopama corpus as the input data for training, since it was one of the best performing corpus for Portuguese in our tests, and it is significantly smaller than the original WikiNER corpus. Table IX shows the tuned hyper-parameters for this corpus and Table X shows the NER model training output using these hyper-parameters for Portuguese.

We were able to achieve an improvement of about 2% for the F-Score metric by just analyzing the best hyper-parameters for the model. This setup is not guaranteed to be the best, since we did not test all possible combinations of parameters for training. However, it is a considerable improvement for a state-of-the-art NER model for the Portuguese language.

B. Pipeline Application

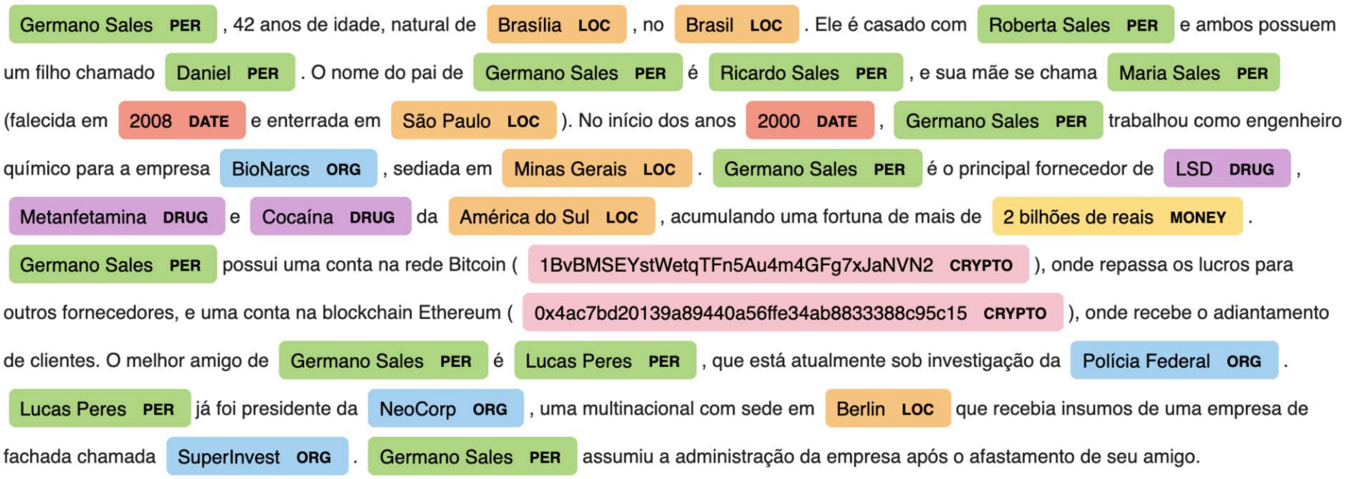
Once the models were properly fine-tuned, we were able to apply the full pipeline architecture to process text data and store them in a graph database. To better understand the processing flow, we will present four examples, two for Portuguese (described in Section VII-C) and two for English (described in Section VII-D).

Fig. 9 shows the complete flow of NLP for these examples. Firstly, the input text data are analyzed to determine in which language it was written. If English is the language of the data, a CR model is applied to enhance entity recognition. The next step is to input named entities manually or by following a



Albert Einstein PER (Ulm LOC , Alemanha LOC , 14 de março de 1879 — Princeton LOC , Estados Unidos LOC , 18 de abril de 1955) foi um físico teórico alemão que desenvolveu a teoria da relatividade geral, um dos pilares da física moderna ao lado da mecânica quântica. Albert Einstein PER veio de uma família de judeus alemães (mãe: Pauline Koch PER , pai: Hermann Einstein PER), Albert Einstein PER mudou-se para a Suíça LOC ainda jovem e iniciou seus estudos na Escola Politécnica de Zurique ORG . Após dois anos procurando emprego, obteve um cargo no escritório de patentes suíço enquanto ingressava no curso de doutorado da Universidade de Zurique ORG . Albert Einstein PER estava nos Estados Unidos LOC quando o Partido Nazista ORG chegou ao poder na Alemanha LOC , em 1933, e não voltou para o seu país de origem, onde tinha sido professor da Academia de Ciências de Berlim ORG . Na véspera da Segunda Guerra Mundial MISC , ajudou a alertar o presidente Franklin Delano Roosevelt PER que a Alemanha LOC poderia estar desenvolvendo uma arma atômica, recomendando aos norte-americanos a começar uma pesquisa semelhante, o que levou ao que se tornaria o Projeto Manhattan MISC . Albert Einstein PER foi afiliado ao Instituto de Estudos Avançados de Princeton ORG , onde trabalhou até sua morte em 1955. Em 1999, foi eleito por 100 físicos renomados o mais memorável físico de todos os tempos. No mesmo ano, a revista TIME MISC , em uma compilação com as pessoas mais importantes e influentes, classificou-o a pessoa do século XX. Albert Einstein PER teve três filhos: Lieserl PER , Hans PER e Eduard PER .

Fig. 10. NER for Scenario 1 (*einstein.txt*).



Germano Sales PER , 42 anos de idade, natural de Brasília LOC , no Brasil LOC . Ele é casado com Roberta Sales PER e ambos possuem um filho chamado Daniel PER . O nome do pai de Germano Sales PER é Ricardo Sales PER , e sua mãe se chama Maria Sales PER (falecida em 2008 DATE e enterrada em São Paulo LOC). No início dos anos 2000 DATE , Germano Sales PER trabalhou como engenheiro químico para a empresa BioNarcs ORG , sediada em Minas Gerais LOC . Germano Sales PER é o principal fornecedor de LSD DRUG , Metanfetamina DRUG e Cocaína DRUG da América do Sul LOC , acumulando uma fortuna de mais de 2 bilhões de reais MONEY . Germano Sales PER possui uma conta na rede Bitcoin (1BvBMSEYstWetqTFn5Au4m4GFg7xJaNVN2 CRYPTO), onde repassa os lucros para outros fornecedores, e uma conta na blockchain Ethereum (0x4ac7bd20139a89440a56ffe34ab8833388c95c15 CRYPTO), onde recebe o adiantamento de clientes. O melhor amigo de Germano Sales PER é Lucas Peres PER , que está atualmente sob investigação da Polícia Federal ORG . Lucas Peres PER já foi presidente da NeoCorp ORG , uma multinacional com sede em Berlin LOC que recebia insumos de uma empresa de fachada chamada SuperInvest ORG . Germano Sales PER assumiu a administração da empresa após o afastamento de seu amigo.

Fig. 11. NER for Scenario 1 (*germano.txt*).

pattern or rule to obtain more entities, although being an optional step. Then, a specific NER model is selected for better performance based on the language determined in the previous step. Once the named entities are recognized and extracted, a relationship model is applied between them, generating an output in the form of a JSON file that can then be easily imported into a graph knowledge application or database.

For better post-processing information retrieval, each node (entity) and edge (relationship) contain some metadata. Regarding the nodes, their entity types and the documents or texts they originated from are stored. For edges, the relationship type and its confidence level are stored for further analysis. Hence, with the aid of a query language, it is possible to filter relationships and extract information in a more granular way.

C. Scenario 1: Information Extraction for Portuguese

For this example, suppose a forensics analysis was made over a suspect's computer. Thanks to IPED, all digital assets found were categorized and labeled according to its content type. Consequently, the text files (which are the target of our NLP work) could be easily accessed and processed to begin the information extraction. Two files were considered for this task: *einstein.txt*, which contains an overview of Albert Einstein's life, and *germano.txt*, containing descriptions about the suspect's rivals in the crime world.

1) Language Detection: IPED's language detection module stated that the language of both files' contents was Portuguese; therefore, to extract the named entities from the text, a Portuguese NER system was chosen; in this case, a NER system obtained from the Paramopama corpus using the BERTimbau [55] transformer.

2) Named Entity Input: Since we suspect that there may be mentions of drugs in the text, especially because drug dealing is a relevant concern for law enforcement agencies, we used a list containing the names of several drugs for the NEI step. Besides, IPED is also able to detect mentions of money quantities and cryptocurrencies addresses in the text, so we also included these entity types in our analysis.

3) Coreference Resolution: As shown in Fig. 9, this step was skipped for the Portuguese scenario for two reasons: it is an optional step and to the best of our knowledge, currently, there are no publicly available tools for CR for Portuguese. We leave that part for future work, as discussed in Sections IV and VIII.

4) Named Entity Recognition and Extraction: Figs. 10 and 11 show the output of our NER system for both files (*einstein.txt* and *germano.txt*). We can see that the NER system was able to correctly identify several entities in the text, including the names of people, organizations, locations, drugs, cryptocurrencies' addresses, money values, and dates. The system attributed different colors based on the entities' types, making it easier to differentiate them.

5) Relationship Model Selection and RE: After the entities were extracted, it is time to configure a RE model and apply it. Since it is a Portuguese text, the DBPedia model was used, following the schema presented in Fig. 4, to detect relationships between entity pairs. Since the DBPedia RE model is based on sentence-level, the document was split into sentences for better performance while deciding which pairs of entities should yield better relationships.

6) Graph Visualization: Once the valid relationships are extracted, a JSON file is generated containing all entities and relationships from the text. This file can then be imported in a graph visualization application, like Neo4j [58]. Figs. 12 and 13 show a graph visualization based on the output of our information extraction system. Since the DBPedia schema only contains relationships for three entity types (person, organization, and location), we filtered the results to show only nodes with relationships between them, leaving other entities (like Drug, for example) out.

D. Scenario 2: Information Extraction for English

This second example is similar to the previous one, with two peculiarities: the addition of a CR step and two RE models for English. Suppose, now, that two other key files for the investigation were found on the suspect's computer, named *corona.txt* and *heroes.txt*.

1) Language Detection: IPED has identified that, unlike the previous files, these ones were written in English.

2) Named Entity Input: The first file's content (*corona.txt*) is about the Coronavirus, so a regular expression was used in the NEI step to detect all mentions of words that contain the term "virus," creating, thus, a new entity category for these words. For the second file (*heroes.txt*), there were no NEIs.

3) Coreference Resolution: Before feeding the NER system, the texts were submitted to a CR model, which intends to remove duplicate entity mentions and improve the overall results. Fig. 14 shows the coreference chains for *corona.txt* identified by the CR system: chain 0, which refers to the word "Coronavirus," chain 1, which refers to David Tyrrel, and chain 2, which refers to the virologist June Almeida. Since chain 0 refers to the term and not the virus itself, we chose not to resolve it in this example. The other chains, however, were resolved to the root term. For *heroes.txt* a bigger number of coreference chains were identified, as shown in Fig. 15. The main references that were resolved for this file include mentions of "Wayne" and "Bruce" (resolved to "Bruce Wayne"), as well as "Clark" (resolved to "Clark Kent"), "Barry" (resolved to "Barry Allen"), "Oliver" (resolved to "Oliver Queen") and "Iris" (resolved to "Iris West").

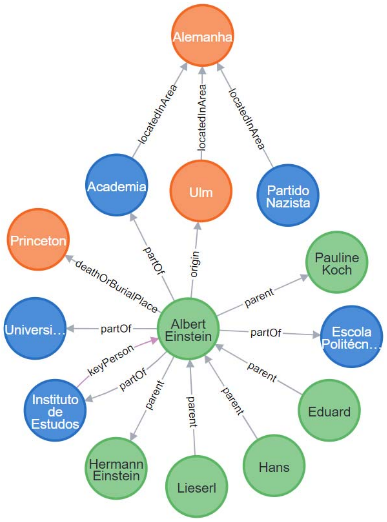


Fig. 12. Scenario 1 (*einstein.txt*) graph output for DBPedia RE model.

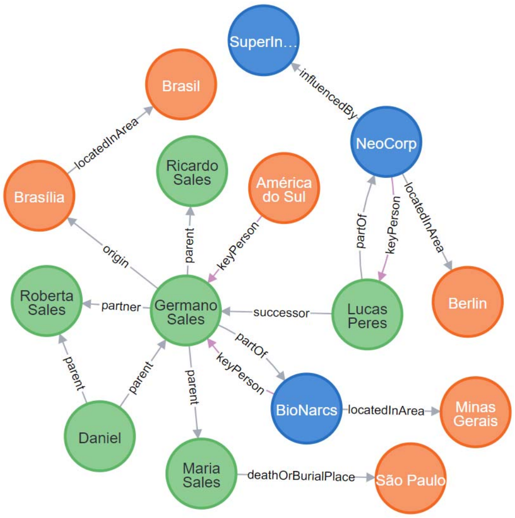
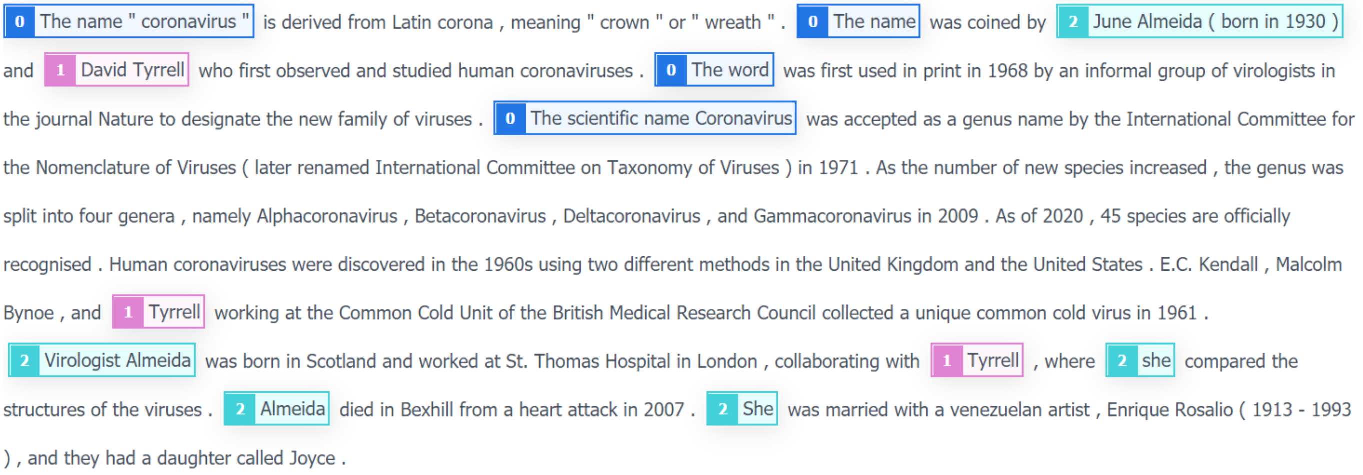


Fig. 13. Scenario 1 (*germano.txt*) graph output for DBPedia RE model.

4) NER Model Selection and Named Entity Extraction: The next step in our NLP flow is to select an appropriate NER model for the input text. For these examples, we used System 1 from Table V. Figs. 16 and 17 show the output of the NER system for both of them. Again, we can see that all relevant entities were correctly identified and classified according to their type, including the new Virus category.

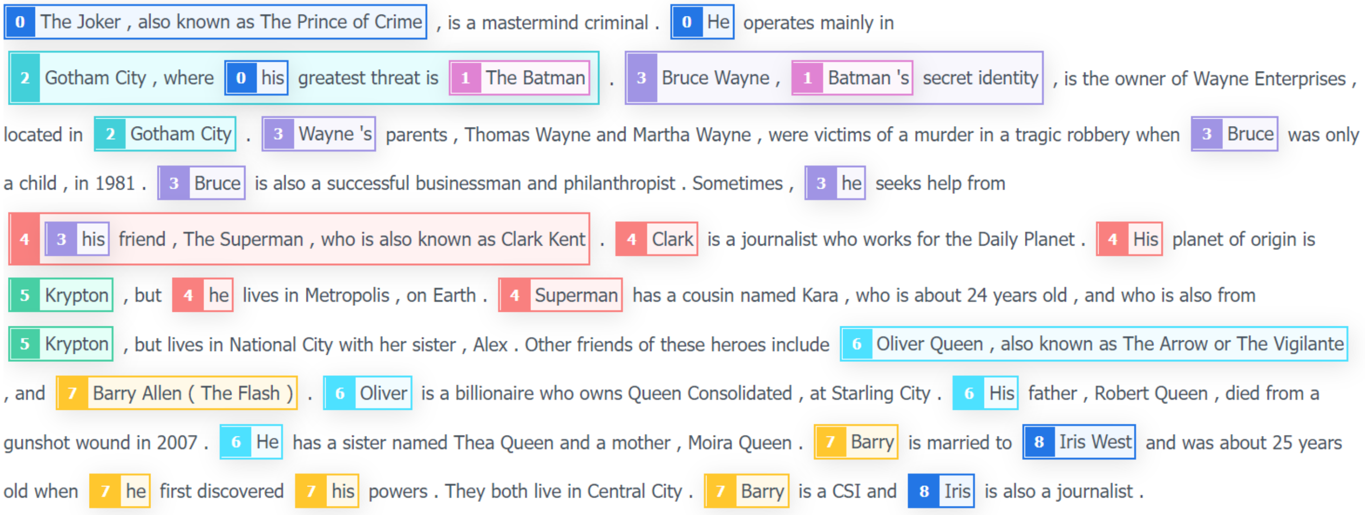
5) Relationship Model Selection and RE: Now it is time to extract the relationships between the entities found. For this part, we used two different models for English (Wiki and TACRED) and outputed their results separately, for each example (*corona.txt* and *heroes.txt*).

6) Graph Visualization: For the *corona.txt* file, Fig. 18 shows the graph visualization for the Wiki RE model, and



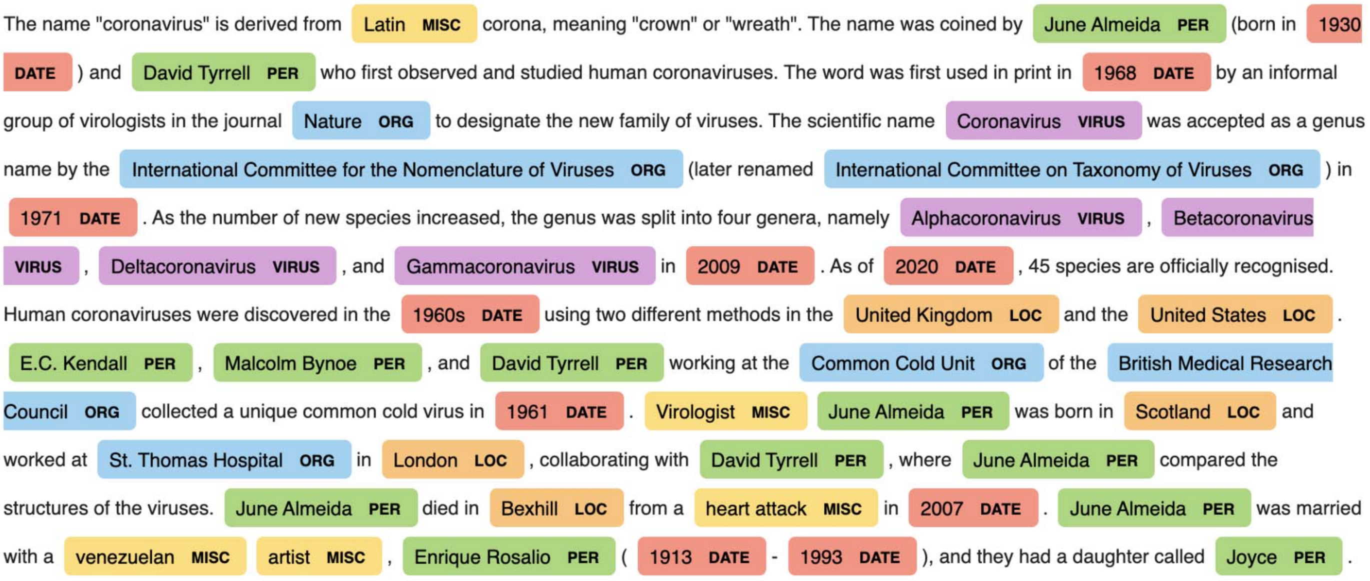
0 The name " coronavirus " is derived from Latin corona , meaning " crown " or " wreath " . 0 The name was coined by 2 June Almeida (born in 1930) and 1 David Tyrrell who first observed and studied human coronaviruses . 0 The word was first used in print in 1968 by an informal group of virologists in the journal Nature to designate the new family of viruses . 0 The scientific name Coronavirus was accepted as a genus name by the International Committee for the Nomenclature of Viruses (later renamed International Committee on Taxonomy of Viruses) in 1971 . As the number of new species increased , the genus was split into four genera , namely Alphacoronavirus , Betacoronavirus , Deltacoronavirus , and Gammacoronavirus in 2009 . As of 2020 , 45 species are officially recognised . Human coronaviruses were discovered in the 1960s using two different methods in the United Kingdom and the United States . E.C. Kendall , Malcolm Bynoe , and 1 Tyrrell working at the Common Cold Unit of the British Medical Research Council collected a unique common cold virus in 1961 . 2 Virologist Almeida was born in Scotland and worked at St. Thomas Hospital in London , collaborating with 1 Tyrrell , where 2 she compared the structures of the viruses . 2 Almeida died in Bexhill from a heart attack in 2007 . 2 She was married with a venezuelan artist , Enrique Rosalio (1913 - 1993) , and they had a daughter called Joyce .

Fig. 14. Scenario 2 (*corona.txt*) coreference resolution.



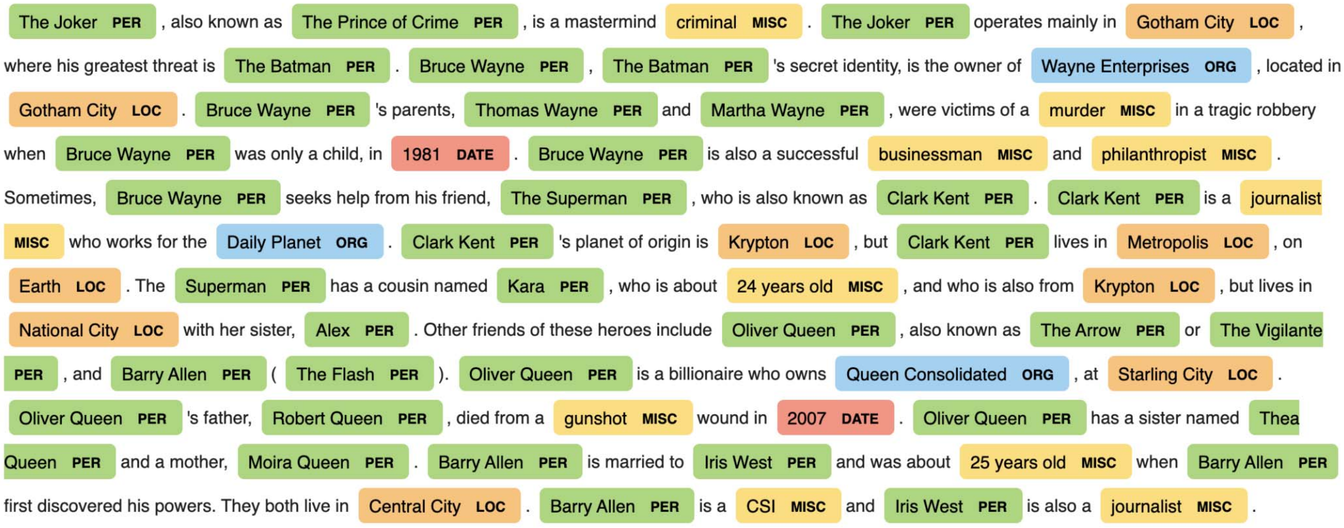
0 The Joker , also known as The Prince of Crime , is a mastermind criminal . 0 He operates mainly in 2 Gotham City , where 0 his greatest threat is 1 The Batman . 3 Bruce Wayne , 1 Batman 's secret identity , is the owner of Wayne Enterprises , located in 2 Gotham City . 3 Wayne 's parents , Thomas Wayne and Martha Wayne , were victims of a murder in a tragic robbery when 3 Bruce was only a child , in 1981 . 3 Bruce is also a successful businessman and philanthropist . Sometimes , 3 he seeks help from 4 3 his friend , The Superman , who is also known as Clark Kent . 4 Clark is a journalist who works for the Daily Planet . 4 His planet of origin is 5 Krypton , but 4 he lives in Metropolis , on Earth . 4 Superman has a cousin named Kara , who is about 24 years old , and who is also from 5 Krypton , but lives in National City with her sister , Alex . Other friends of these heroes include 6 Oliver Queen , also known as The Arrow or The Vigilante , and 7 Barry Allen (The Flash) . 6 Oliver is a billionaire who owns Queen Consolidated , at Starling City . 6 His father , Robert Queen , died from a gunshot wound in 2007 . 6 He has a sister named Thea Queen and a mother , Moira Queen . 7 Barry is married to 8 Iris West and was about 25 years old when 7 he first discovered 7 his powers . They both live in Central City . 7 Barry is a CSI and 8 Iris is also a journalist .

Fig. 15. Scenario 2 (*heroes.txt*) coreference resolution.



The name "coronavirus" is derived from Latin MISC corona, meaning "crown" or "wreath". The name was coined by June Almeida PER (born in 1930 DATE) and David Tyrrell PER who first observed and studied human coronaviruses. The word was first used in print in 1968 DATE by an informal group of virologists in the journal Nature ORG to designate the new family of viruses. The scientific name Coronavirus VIRUS was accepted as a genus name by the International Committee for the Nomenclature of Viruses ORG (later renamed International Committee on Taxonomy of Viruses ORG) in 1971 DATE . As the number of new species increased, the genus was split into four genera, namely Alphacoronavirus VIRUS , Betacoronavirus VIRUS , Deltacoronavirus VIRUS , and Gammacoronavirus VIRUS in 2009 DATE . As of 2020 DATE , 45 species are officially recognised. Human coronaviruses were discovered in the 1960s DATE using two different methods in the United Kingdom LOC and the United States LOC . E.C. Kendall PER , Malcolm Bynoe PER , and David Tyrrell PER working at the Common Cold Unit ORG of the British Medical Research Council ORG collected a unique common cold virus in 1961 DATE . Virologist MISC June Almeida PER was born in Scotland LOC and worked at St. Thomas Hospital ORG in London LOC , collaborating with David Tyrrell PER , where June Almeida PER compared the structures of the viruses. June Almeida PER died in Bexhill LOC from a heart attack MISC in 2007 DATE . June Almeida PER was married with a venezuelan MISC artist MISC , Enrique Rosalio PER (1913 DATE - 1993 DATE), and they had a daughter called Joyce PER .

Fig. 16. NER for Scenario 2 (*corona.txt*).



The Joker PER , also known as The Prince of Crime PER , is a mastermind criminal MISC . The Joker PER operates mainly in Gotham City LOC , where his greatest threat is The Batman PER . Bruce Wayne PER , The Batman PER 's secret identity, is the owner of Wayne Enterprises ORG , located in Gotham City LOC . Bruce Wayne PER 's parents, Thomas Wayne PER and Martha Wayne PER , were victims of a murder MISC in a tragic robbery when Bruce Wayne PER was only a child, in 1981 DATE . Bruce Wayne PER is also a successful businessman MISC and philanthropist MISC . Sometimes, Bruce Wayne PER seeks help from his friend, The Superman PER , who is also known as Clark Kent PER . Clark Kent PER is a journalist MISC who works for the Daily Planet ORG . Clark Kent PER 's planet of origin is Krypton LOC , but Clark Kent PER lives in Metropolis LOC , on Earth LOC . The Superman PER has a cousin named Kara PER , who is about 24 years old MISC , and who is also from Krypton LOC , but lives in National City LOC with her sister, Alex PER . Other friends of these heroes include Oliver Queen PER , also known as The Arrow PER or The Vigilante PER , and Barry Allen PER (The Flash PER). Oliver Queen PER is a billionaire who owns Queen Consolidated ORG , at Starling City LOC . Oliver Queen PER 's father, Robert Queen PER , died from a gunshot MISC wound in 2007 DATE . Oliver Queen PER has a sister named Thea Queen PER and a mother, Moira Queen PER . Barry Allen PER is married to Iris West PER and was about 25 years old MISC when Barry Allen PER first discovered his powers. They both live in Central City LOC . Barry Allen PER is a CSI MISC and Iris West PER is also a journalist MISC .

Fig. 17. NER for scenario 2 (*heroes.txt*).

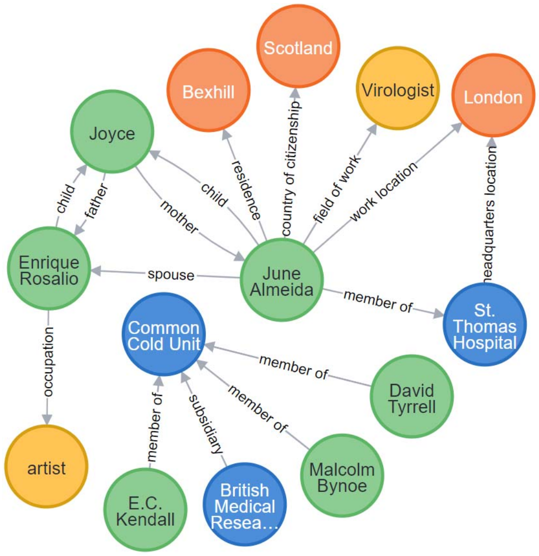


Fig. 18. Scenario 2 (*corona.txt*) graph output for Wiki RE model.

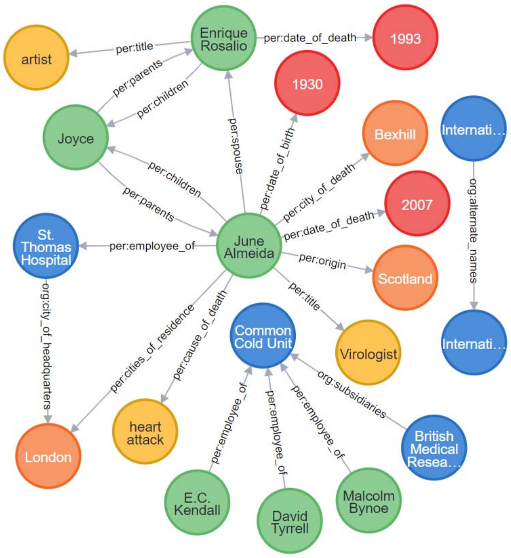


Fig. 19. Scenario 2 (*corona.txt*) graph output for TACRED RE model.

Fig. 19 shows the graph visualization for the TACRED RE model, respectively. We can notice that the Wiki RE model was able to detect several relevant relationships, like the relatives of June Almeida, her field of work, and work location, for example. The TACRED model also achieved good results, with the benefit of being able to detect relationships for Date entity types, like June's date of birth and date of death. This can be a key information when dealing with forensics and when trying to understand the chronology of facts under investigation, for example. For the *heroes.txt* file, Fig. 20 shows the graph visualization for the Wiki RE model, and Fig. 21 shows the graph visualization for the TACRED RE model. Both outputs achieved good results, with the detection of relevant relations, like the occupation, residence, work location, and family members of the persons in the file's content. The TACRED RE model was able to extract the ages mentioned in the text for Barry Allen and Kara, the cause of death and date of death for Robert Queen and dates of death for Bruce's parents (not detected by the Wiki model). However, the Wiki model correctly detected Kara's residence in Krypton and Barry Allen's occupation as a CSI (both missed by the TACRED model).

Often, in practice, it may be relevant to switch between different NER and RE systems to detect and analyze as many relations as possible, specially because longer texts tend to contain complex relations between entities. Even though the outputs showed in Figs. 18 and 19 and in Figs. 20 and 21 are similar, they can complement each other and provide better results.

E. Comparison With the State-of-the-Art

In this section, we compare our implementations of NER and RE models with some state-of-the-art models from the literature. Table XI and Fig. 22 show the comparisons of NER

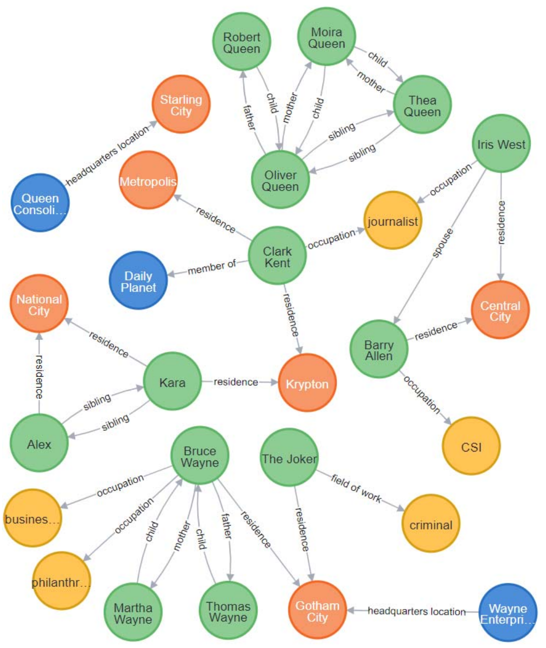


Fig. 20. Scenario 2 (*heroes.txt*) graph output for Wiki RE model.

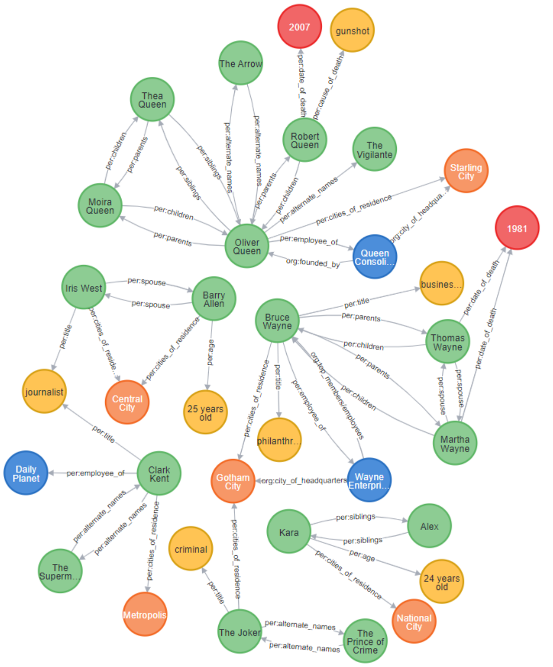


Fig. 21. Scenario 2 (*heroes.txt*) graph output for TACRED RE model.

TABLE XI
COMPARISON OF OUR RESULTS FOR NER WITH THE STATE-OF-THE-ART

Corpus	Model	Size	F-Score	Reference
Mini HAREM	BERT-PT **(ours)**	Base	**79.4**	–
	BERT-PT	Large	83.7	[55]
Paramopama	BERT-PT **(ours)**	Base	**89.0**	–
	LSTM-CRF	–	90.5	[14]
	BERT-PT$_{\text{HT}}$ **(ours)**	Base	**91.2**	–
LeNER-Br	LSTM-CRF	–	86.6	[14]
	BERT-PT **(ours)**	Base	**89.2**	–
CoNLL03	RoBERTa **(ours)**	Base	**92.1**	–
	BERT	Large	92.8	[3]
	FLERT	Large	94.0	[15]
	LUKE	Large	94.3	[33]
	ACE	Large	94.6	[16]
WNUT17	RoBERTa **(ours)**	Base	**48.3**	–
	BERTweet	Base	56.5	[35]

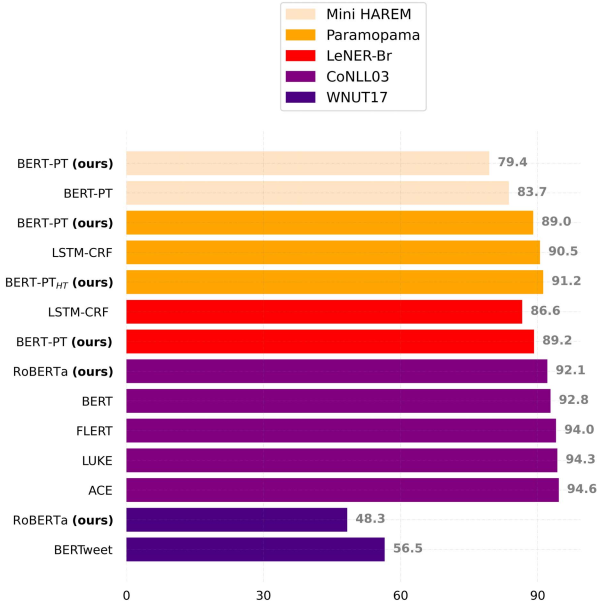


Fig. 22. Comparison between our NER models and the state-of-the-art. The y-axis show different state-of-the-art models and their corresponding F-Scores are shown in the x-axis. Different datasets are represented by different colors.

models, whereas Table XII and Fig. 23 show the comparisons of RE models.

For the NER task, we compare our best models for Mini HAREM, Paramopama, LeNER-Br, CoNLL03, and WNUT17 with the current state-of-the-art. We did not find current relevant results for the First and Second HAREM collections, nor for the Portuguese WikiNER corpus, so we did not include these collections in our comparisons. For the Paramopama dataset, we compare the previous state-of-the-art (LSTM-CRF model) with our model before hyper-parameters tuning (BERT-PT) and after hyper-parameters tuning (BERT-PT$_{\text{HT}}$). Our fine-tuned model outperforms the previous state-of-the-art, achieving a new standard for Portuguese NER in this dataset.

For the CoNLL03 dataset, we compare our model with three others. The ACE model [16] is the current state-of-the-art for NER on CoNLL03. Nevertheless, we would like to point out that these models were fine-tuned for a longer period of time if compared with our solution, since we limited the number of epochs to 20 in our experiments for practical reasons, while most of the state-of-the-art NER models usually adopt 50 epochs for training. In addition, the codes for FLERT [15], LUKE [33], and ACE [16] were released to the community, which means that their results can be easily replicated and applied to different datasets with some modifications. Regarding the WNUT17 benchmark, the BERTweet model significantly outperforms ours, achieving a F-Score of 56.5%.

TABLE XII
COMPARISON OF OUR RESULTS FOR RE WITH THE STATE-OF-THE-ART

Corpus	Model	Size	F-Score	Reference
DBPedia	KNN	–	55.6	[51]
	BERT-PT$_{\text{EM}}$ **(ours)**	Base	**84.6**	–
	BERT-PT$_{\text{TEM}}$ **(ours)**	Base	**86.6**	–
TACRED	BERT	Base	66.0	[34]
	ERNIE	Base	67.9	[34]
	BERT$_{\text{EM}}$ **(ours)**	Base	**69.9**	–
	BERT$_{\text{EM}}$	Large	70.1	[23]
	SpanBERT	Large	70.8	[32]
	BERT$_{\text{EM}}$+MTB	Large	71.5	[23]
	BERT$_{\text{TEM}}$ **(ours)**	Base	**71.8**	–
	LUKE	Large	72.7	[33]
	RoBERTa$_{\text{TEM}}$	Large	74.6	[22]
	RECENT	Large	75.2	[24]
M-TACRED	BERT$_{\text{EM}}$ **(ours)**	Base	**79.0**	–
	BERT$_{\text{TEM}}$ **(ours)**	Base	**80.5**	–

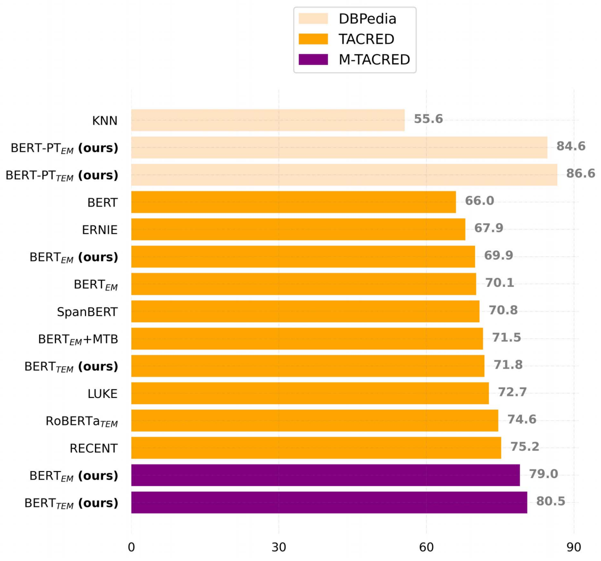


Fig. 23. Comparison between our RE models and the state-of-the-art. The y-axis show the different state-of-the-art models and their corresponding F-Scores are shown in the x-axis. Different datasets are represented by different colors.

Nevertheless, BERTweet was pretrained on a large domain-specific corpus, whereas we used a general-purpose model for the same task.

For the task of RE, since the Wiki dataset used in our work is not an official benchmark, we did not use it for comparison. In addition, because our implementation of the TACRED dataset required some modifications in the original dataset for practical purposes, as explained in Section V, we decided to train two new models with the original dataset to establish a fair comparison with the current state-of-the-art. The first TACRED model, BERT$_{\text{EM}}$, uses an EM described by [23], which is the same method we used previously with all the RE models. The second TACRED model, BERT$_{\text{TEM}}$, uses a TEM described in [22]. We also included the results on our modified version of TACRED (M-TACRED) at the bottom of Table XII for comparison. For M-TACRED, we included the result of our BERT model with EM from Section VI and the result for the same base model trained with TEM, which achieved an improvement of 1.5% in the F-Measure.

For the DBPedia RE models' comparison, we were not able to find other works that used the same dataset and the same group of relation classes. Because of that, we chose to compare it only with the original implementation [51]. Our BERT-PT$_{\text{TEM}}$ model significantly outperforms the previous state-of-the-art for the original DBPedia model, which was based on a K-Nearest-Neighbors (KNN) model. Both of our TACRED RE models also achieved good performances, outperforming other approaches, but still a little behind the current state-of-the-art. However, our models are based on BERT-Base, which has significantly less parameters than BERT-Large and other large transformers.

VIII. Conclusion

This article proposed a systematic solution for information extraction through natural language processing of text data. We showed that named entity recognition (NER) is a common task to gather specific token types like persons, organizations, and locations. The advent of transformers made it easier to create state-of-the-art performing models fine-tuned for multiple tasks, including NER and RE.

Nevertheless, these cutting-edge models are not yet implemented for some applications, as of some forensic tools. In addition, as far as we are aware, there are not many solutions available for Portuguese yet, with most of the efforts being concentrated in the English language. We also demonstrated, however, that the availability of new transformer models for Portuguese, like BERTimbau, with the appropriate corpora, made it possible to fine-tune new models and achieve the state-of-the-art performance in any language.

Finally, we were able to train several NER models for both Portuguese and English with great performances, with the two best ones being the Portuguese model trained on Paramopama (F-Score of 91%) and the English model trained on CoNLL03 (F-Score of 92%). For the RE part, we trained a new model for Portuguese based on the DBPedia corpus, achieving a F-Score of 86%.

Future work for further improvements with information retrieval using NLP systems is still in progress. In [60], it is shown how to create a massive NER corpus for Portuguese using open source datasets. Low amounts of labeled data are usually the bottleneck of many NLP downstream tasks, thus, a massive corpus could help improve the results. Feature engineering, alongside with hyper-parameters tuning, can also provide better models and scores, and there are other approaches for relation extraction that could be considered as well, like open information extraction, to extend the relations detected and build a more comprehensive graph database.

More efforts are also desirable to fine-tune NER and RE models for other languages. The CoNLL 2003 [17] task also provided a NER dataset for German, and CoNLL 2012 shared task [40] provided datasets for Chinese and Arabic that could be used to develop and test NER systems for these languages. For coreference resolution, to the best of our knowledge, there are no available solutions for Portuguese, and we believe that other Latin-derived languages, like Spanish and Italian, are equally challenging for this task.

References

[1] M. D. Kohn, M. M. Eloff, and J. H. P. Eloff, "Integrated digital forensic process model," *Comput. Secur.*, vol. 38, pp. 103–115, Oct. 2013, doi: 10.1016/j.cose.2013.05.001.

[2] V. J. Alles, W. F. Giozza, and R. de Oliveira Alburquerque, "Natural language processing to classify named entities of the Brazilian Union Official Diary," in *Proc. 13th Iberian Conf. Inf. Syst. Technol. (CISTI)*, 2018, pp. 1–6, doi: 10.23919/CISTI.2018.8399215.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[4] *Spacy*. Accessed: Sep. 20, 2021. [Online]. Available: https://spacy.io

[5] C. M. Júnior, H. Macedo, T. Bispo, F. Santos, N. Silva, and L. Barbosa, "ParamoPama: A Brazilian-Portuguese corpus for named entity recognition," in *Proc. Encontro Nat. Intell. Artif. Comput. (ENIAC)*, 2015, pp. 1–5.

[6] L. Caviglione, S. Wendzel, and W. Mazurczyk, "The future of digital forensics: Challenges and the road ahead," *IEEE Security Privacy*, vol. 15, no. 6, pp. 12–17, Nov./Dec. 2017, doi: 10.1109/MSP.2017.4251117.

[7] D. O. Ukwen and M. Karabatak, "Review of NLP-based systems in digital forensics and cybersecurity," in *Proc. 9th Int. Symp. Digit. Forensics Secur. (ISDFS)*, Jun. 2021, pp. 1–9, doi: 10.1109/ISDFS52919.2021.9486354.

[8] *IPED Tool*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/sepinf-inc/IPED

[9] R. B. van Baar, H. M. A. van Beek, and E. J. van Eijk, "Digital forensics as a service: A game changer," *Digit. Invest.*, vol. 11, pp. S54–S62, May 2014, doi: 10.1016/j.diin.2014.03.007.

[10] H. M. A. van Beek, E. J. van Eijk, R. B. van Baar, M. Ugen, J. N. C. Bodde, and A. J. Siemelink, "Digital forensics as a service: Game on," *Digit. Invest.*, vol. 15, pp. 20–38, Dec. 2015, doi: 10.1016/j.diin.2015.07.004.

[11] H. M. A. van Beek, J. van den Bos, A. Boztas, E. J. van Eijk, R. Schramp, and M. Ugen, "Digital forensics as a service: Stepping up the game," *Forensic Sci. Int., Digit. Invest.*, vol. 35, Dec. 2020, Art. no. 301021, doi: 10.1016/j.fsidi.2020.301021.

[12] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, gate," in *Proc. 6th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, 2019, pp. 338–343, doi: 10.1109/SNAMS.2019.8931850.

[13] L. A. Cabrera-Diego, J. G. Moreno, and A. Doucet, "Simple ways to improve NER in every language using markup," in *Proc. WWW*, 2021, pp. 17–31.

[14] P. H. L. de Araujo, T. E. de Campos, R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "Lener-Br: A dataset for named entity recognition in Brazilian legal text," in *Int. Conf. Comput. Process. Portuguese Lang.* Springer, 2018, pp. 313–323, doi: 10.1007/978-3-319-99722-3_32.

[15] S. Schweter and A. Akbik, "FLERT: Document-level features for named entity recognition," 2020, *arXiv:2011.06993*.

[16] X. Wang *et al.*, "Automated concatenation of embeddings for structured prediction," 2020, *arXiv:2010.05006*.

[17] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," 2003, *arXiv:cs/0306050*.

[18] J. Lee, S. Seo, and Y. S. Choi, "Semantic relation classification via bidirectional LSTM networks with entity-aware attention using latent entity typing," *Symmetry*, vol. 11, no. 6, p. 785, Jun. 2019, doi: 10.3390/sym11060785.

[19] G. Nan, Z. Guo, I. Sekuliá, and W. Lu, "Reasoning with latent structure refinement for document-level relation extraction," 2020, *arXiv:2005.06312*.

[20] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*.

[21] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, and M. Sun, "OpenNRE: An open and extensible toolkit for neural relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 169–174, doi: 10.18653/v1/D19-3029.

[22] W. Zhou and M. Chen, "An improved baseline for sentence-level relation extraction," 2021, *arXiv:2102.01373*.

[23] L. Baldini Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," 2019, *arXiv:1906.03158*.

[24] S. Lyu and H. Chen, "Relation classification with entity type restriction," 2021, *arXiv:2105.08393*.

[25] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 35–45.

[26] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "T-BERTSum: Topic-aware text summarization based on BERT," *IEEE Trans. Computat. Social Syst.*, early access, Jun. 24, 2021, doi: 10.1109/TCSS.2021.3088506.

[27] T. Saha, S. R. Jayashree, S. Saha, and P. Bhattacharyya, "BERT-caps: A transformer-based capsule network for tweet act classification," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 5, pp. 1168–1179, Oct. 2020, doi: 10.1109/TCSS.2020.3014128.

[28] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile, "AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets," in *Proc. 6th Italian Conf. Comput. Linguistics (CLiC-it)*, vol. 2481, 2019, pp. 1–6.

[29] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[30] M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, "BERT for coreference resolution: Baselines and analysis," 2019, *arXiv:1908.09091*.

[31] B. Kantor and A. Globerson, "Coreference resolution with entity equalization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 673–677.

[32] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, Oct. 2020, doi: 10.1162/tacl_a_00300.

[33] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," 2020, *arXiv:2010.01057*.

[34] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," 2019, *arXiv:1905.07129*.

[35] D. Quoc Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," 2020, *arXiv:2005.10200*.

[36] R. Gabbard, M. Freedman, and R. Weischedel, "Coreference for learning to extract relations: Yes virginia, coreference matters," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 288–293.

[37] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.

[38] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[39] *Neuralcoref*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/huggingface/neuralcoref

[40] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," in *Proc. Joint Conf. EMNLP CoNLL-Shared Task*, 2012, pp. 1–40.

[41] K. Clark and C. D. Manning, "Deep reinforcement learning for mention-ranking coreference models," 2016, *arXiv:1609.08667*.

[42] A. Antonitisch, A. Figueira, D. Amaral, E. B. Fonseca, S. C. de Abreu, and R. Vieira, "Summ-it++: An enriched version of the summ-it corpus," in *Proc. LREC*, 2016, pp. 2047–2051.

[43] R. Vieira, A. Mendes, P. Quaresma, E. Fonseca, S. Collovini, and S. Antunes, "Corref-PT: A semi-automatic annotated Portuguese coreference corpus," *Computacióny Sistemas*, vol. 22, no. 4, pp. 1259–1267, Dec. 2018.

[44] *NER Datasets for Portuguese*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/davidsbatista/NER-datasets/blob/master/Portuguese/README.MD

[45] *NER Datasets for English*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/juand-r/entity-recognition-datasets

[46] D. Santos, N. Seco, N. Cardoso, and R. Vilela, "Harem: An advanced ner evaluation contest for Portuguese," in *Proc. 5th Int. Conf. Lang. Resour. Eval.*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, and D. Tapias, Eds., Genoa, Italy, May 2006, pp. 1–8.

[47] C. Freitas, P. Carvalho, H. Gonçalo Oliveira, C. Mota, and D. Santos, "Second harem: Advancing the state of the art of named entity recognition in Portuguese," in *Proc. Eur. Lang. Resour. Assoc.*, 2010, pp. 1–6.

[48] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," *Artif. Intell.*, vol. 194, pp. 151–175, Jan. 2013, doi: 10.1016/j.artint.2012.03.006.

[49] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proc. 3rd Workshop Noisy User-generated Text*, 2017, pp. 140–147.

[50] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural Language Processing Using Very Large Corpora*. Dordrecht, The Netherlands: Springer, 1999, pp. 157–176, doi: 10.1007/978-94-017-2390-9_10.

[51] D. S. Batista, D. Forte, R. Silva, B. Martins, and M. Silva, "Exploring DBpedia and Wikipedia for Portuguese Semantic Relationship Extraction," *Linguamatica*, vol. 5, no. 1, pp. 41–57, 2013. [Online]. Available: https://www.linguamatica.com/index.php/linguamatica/article/view/157

[52] X. Han *et al.*, "FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," 2018, *arXiv:1810.10147*.

[53] *DBPedia Semantic Relationship Dataset for Portuguese*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/davidsbatista/Annotated-Semantic-Relationships-Datasets/blob/master/datasets/DBpediaRelations-PT-0.2.txt.bz2

[54] *OpenNRE Framework*. Accessed: Sep. 20, 2021. [Online]. Available: https://github.com/thunlp/OpenNRE

[55] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Brazilian Conference on Intelligent Systems*. Springer, 2020, pp. 403–417, doi: 10.1007/978-3-030-61377-8_28.

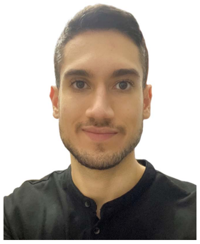
[56] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[57] *AllenNLP Natural Language Processing Platform*. Accessed: Sep. 20, 2021. [Online]. Available: https://allennlp.org/

[58] *Neo4J Graph Database Platform*. Accessed: Sep. 20, 2021. [Online]. Available: https://neo4j.com/

[59] *Google Colab*. Accessed: Sep. 20, 2021 [Online]. Available: https://colab.research.google.com/

[60] D. Menezes, R. Milidiu, and P. Savarese, "Building a massive corpus for named entity recognition using free open data sources," in *Proc. 8th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2019, pp. 6–11, doi: 10.1109/BRACIS.2019.00011.



Fillipe Barros Rodrigues was born in Brasília, Brazil, in 1996. He received the bachelor's degree in network engineering from the University of Brasília (UnB), Brasília, Brazil, in 2019, where he is currently pursuing the master's degree in electrical engineering.

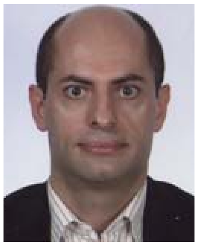
His research interests include natural language processing, cybersecurity, and blockchain. In addition, he has experience in the field of data science, with an emphasis on data visualization and network analysis, as well as software development for web and mobile applications.

William Ferreira Giozza (Senior Member, IEEE) was born in Jaguarão, RS, Brazil, in 1953. He received the bachelor's degree in electronics engineering from the Aeronautical Technological Institute (ITA), São José dos Campos, SP, Brazil, in 1976, the master's degree in electrical engineering from the Federal University of Paraíba (UFPb), Campina Grande, PB, Brazil, in 1979, and the Ph.D. degree in computer science from the University of Paris 6, Paris, France, in 1982.

From 1982 to 1998, he was an Associate Professor with UFPb. From 1998 to 2009, he was a Full Professor with Salvador University, Salvador, BA, Brazil. Since 2009, he has been with the Decision Technologies Laboratory-LATITUDE, Electrical Engineering Department, University of Brasília (UnB), Brasília, Brazil, where he has been In-Charge of optical communication, high-speed networking, and cybersecurity education and research.

Robson de Oliveira Albuquerque received the degree in computer science from the University Católica of Brasília (UCB), Brasília, Brazil, in 1999, the master's degree in electrical engineering from the University of Brasília (UNB), Brasília, in 2003, the DEA degree from the University Complutense of Madrid (UCM), Madrid, Spain, in 2007, the Doctorate Degree from UNB in 2008, and the Ph.D. degree from UCM in 2016.

In 2001, he finished a specialization course in computer networks with Brasília Educacional Union (UNEB). In 2020, he finished his postdoc in cybersecurity with UNB in association with the professional postgraduate program in electrical engineering. He has more than 25 years of experience in computer networks, information systems and network security. His field of study and research includes information systems, computer networks, network security, information security, and cyber security. His professional skills include IT consulting for private organizations and the Brazilian Federal Government. He is a member of the Professional Post-Graduate Program in Electrical Engineering (PPEE) in the Electrical Engineering Department, University of Brasília. He contributes as a Researcher and Professor with the Brazilian National Science and Technology Institute on Cybersecurity (CyberSecurity INCT)—-LATITUDE Laboratory. He is member of AQUARELA Research Group with the University of Brasilia and also he is a member of GASS Research Group with the University Complutense of Madrid. He has several research works published in journals and conferences around the world.



Luis Javier García Villalba (Senior Member, IEEE) received the bachelor's degree in telecommunications engineering from the Universidad de Málaga, Málaga, Spain, and the master's degree in computer networks and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid, Madrid, Spain.

He was a Visiting Scholar with COSIC (Computer Security and Industrial Cryptography, Department of Electrical Engineering, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium) in 2000 and a Visiting Scientist with IBM Research Division (IBM Almaden Research Center, San Jose, CA, USA) in 2001 and 2002. He is currently an Associate Professor with the Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid (UCM) and the Head of the Complutense Research Group GASS (Group of Analysis, Security and Systems) which is located in the Faculty of Computer Science and Engineering at the UCM Campus. His professional experience includes the management of both national and international research projects and both public (Spanish Ministry of R&D, Spanish Ministry of Defense, Horizon 2020 — European Commission, etc.) and private research projects (Hitachi, IBM, Nokia, Safelayer Secure Communications, TB Solutions Security, etc.). He is the author or coauthor of numerous international publications.

Dr. García Villalba is an Editor or Guest Editor of numerous journals such as *Entropy MPDI*, *Future Generation Computer Systems*, *Future Internet MDPI*, the IEEE LATIN AMERICA TRANSACTIONS, *IET Communications*, *IET Networks*, *IET Wireless Sensor Systems*, *International Journal of Ad Hoc and Ubiquitous Computing*, *International Journal of Multimedia and Ubiquitous Engineering (IJMUE)*, *Journal of Supercomputing*, and *Sensors MDPI*.