Could you tell me the process ID? Structuring Text Documents from the Brazilian Electronic Information System Using a Named Entity Recognition Approach

Mayara C. Marinho, Ana Clara B. Borges, Laryssa O. Ferreira, Vanessa P. Costa, J. L. Bordim, Vinicius R. P. Borges
mayarachewm@gmail.com,borges.anacb@gmail.com,{laryssa.ferreira,costa.vanessa}@aluno.unb.br,
{bordim,viniciusrpb}@unb.br
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Brazil

Abstract

Context: In the context of valuing open data, transparency, and efficiency in public services, there is a growing demand for studies to improve government systems. Some public agencies use the Brazilian Electronic Information System (*Sistema Eletrônico de Informações* - SEI), a procedural management system that centralizes electronic processes and promotes administrative efficiency. **Problem:** Although SEI has contributed to advancements in public administration, significant challenges remain in information search and retrieval due to the inefficient keyword-based approach currently available. These difficulties are enhanced by the high amount of documents generated daily, which are written in natural language and present variability in categories, writing styles, and structures. As a result, searching for relevant documents in SEI is time-consuming, leading users to create unnecessary processes and inconsistent resolutions when compared to previously completed processes. **Solution:** A Natural Language Processing (NLP) pipeline was proposed to extract information from SEI documents using Named Entity Recognition (NER) models. **IS Theory:** Organizational Information Processing. **Method:** This research adopts a descriptive approach. Public SEI documents were collected using a web crawler, and trained annotators built a corpus to enable the training of state-of-the-art NER models. The models' performances were compared and quantitatively analyzed. **Summary of Results:** A Brazilian Portuguese labeled corpus of SEI for NER was curated and validated, leading to an NLP pipeline for information extraction. **Contributions and Impact in the IS area:** This research provides a baseline for structuring data from Electronic Information Systems, enabling more effective strategies for search and retrieval tasks.

CCS Concepts

• **Information systems** → **Information retrieval**; *Retrieval models and ranking*; Language models; • **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; Information extraction.

Keywords

Natural Language Processing, Named Entity Recognition, Portuguese Processing, Electronic Information System

1 Introduction

In the past decades, the government has invested heavily in the improvement of public service performance, aiming for agility, productivity, transparency, user satisfaction, and cost reduction. In this context, an initiative called National Electronic Process (PEN) was formulated to create a public infrastructure for electronic administrative processes and documents, and one of the main solutions proposed was the Brazilian Electronic Information System (in Portuguese, *Sistema Eletrônico de Informações* - SEI).

Decree No. 36.756, September 16, 2015, established SEI as the official system for managing documents and administrative processes of the Public Administration of the Federal District, Brazil. According to the mentioned decree, SEI was created with the purpose of: (1) increasing productivity and speed in processing documents; (2) improving data security and reliability; (3) creating more appropriate conditions for the production and use of information; (4) facilitating access to information; and (5) reducing the use of paper, operational and documentation storage costs. Studies also concluded that SEI implementation successfully achieved most of its main purposes [4][23].

Nowadays, SEI is a fundamental system that supports the daily activities of 115 government entities and public bodies [1]. The administrative staff and civil servants - the target users - use SEI to create, store, and organize official processes to enable processing through different departments and divisions. An electronic process receives a unique identifier (process ID) after its creation, allowing users to track its progress until resolution. Moreover, SEI supports various types of documents, which users can create (or upload in the case of external documents) and store in chronological order of their inclusion in an electronic process. Another important feature of SEI is the electronic signature of documents, which is legally required to validate the documents.

Traditionally, users retrieve processes and documents in SEI by querying the search engine with process IDs or document keywords. Furthermore, predefined filters can be applied to select specific documents based on their categories, such as forms, opinions, orders, and acts. After the search results are returned, users must manually verify the relevance of the retrieved documents. This task is often tedious and time-consuming due to the presence of various document types, the large volume of documents, typographical errors,

[1] https://portalsei.df.gov.br/category/seigdf/atendimento/, accessed on February 28, 2025.

and noticeable variability in writing styles within the same type of document. As SEI documents are written in natural language and authored by diverse users, the current exact-match search approach applied to its unstructured database significantly affects the effectiveness of information retrieval.

The Access to Information Law [2] (LAI), which is similar to the Freedom of Information Act (FOIA) in the US, establishes the right of the general public to access information, ensuring transparency in the acts of the Union, States, Federal District, and Municipalities, as mandated by the Federal Constitution. Within this framework, SEI plays a fundamental role, allowing the public to search for and retrieve public documents. This contributes to increased transparency and accountability of public acts. However, in its current structure, SEI provides limited tools to enable users to obtain accurate and precise information on specific topics.

As SEI documents are presented in an unstructured format, search efficiency is further compromised. The lack of structured or semi-structured text makes it difficult to extract relevant information, especially in cases where exact keywords are unavailable, or documents include varied terminology and phrasing. This limitation affects not only external users seeking transparency but also the efficiency of internal users who rely on SEI for their daily administrative tasks. Searching for information often becomes a time-consuming process, leading to delays and, in some cases, the inadvertent duplication of processes when original procedures cannot be located. Such duplication can result in divergent decisions and reduce institutional efficiency.

To address these challenges, it is necessary to consider advanced strategies for processing unstructured data. Extracting information from SEI documents and transforming it into a structured or semi-structured format would enhance search engines' capability to retrieve information semantically or by fields such as names, dates, and locations. Natural Language Processing (NLP) techniques, particularly Named Entity Recognition (NER), offer promising solutions. NER can label specific entities in text, associating them with document types and topics of interest, thereby converting unstructured SEI content into a more accessible and navigable form.

This paper presents an NLP pipeline to extract and structure information in SEI documents using Named Entity Recognition (NER). The goal is to train an NER model to identify entities in SEI documents, transforming unstructured natural language text into structured data and enabling entity-based search. Since NER is a supervised NLP task, a corpus of public SEI documents from University of Brasília has been created and labeled by trained human annotators, providing reliable data for training NER models. The "SEI corpus" enables the training and evaluation of NER models specifically designed for SEI documents and their entities. These advancements enhance search accuracy and efficiency, ensuring greater transparency and alignment with the principles of LAI. Moreover, the approach is expected to improve document accessibility, support compliance with transparency requirements, and enhance both the user experience and administrative efficiency.

The following research question guided this research: *can state-of-the-art NER techniques effectively extract information from the unstructured text documents of the Brazilian Electronic Information System (SEI), considering the wide variety of entity types and document formats?*

The main contributions described in this paper are:

- SEI corpus: an annotated Brazilian Portuguese corpus of public documents extracted from SEI for NER tasks.
- An NLP pipeline to extract information from SEI documents;
- A prototype of a Web Application that demonstrates search operations on structured data extracted from SEI documents using the proposed NLP pipeline.

This research is also motivated by the need to prioritize transparency, productivity, and efficiency in Information Systems (IS), particularly in government systems. These are relevant research topics cited in the "Grand Research Challenges in Information Systems in Brazil 2016-2026" book organized by the Brazilian Computer Society (SBC). This research aims to improve the information search system in the SEI, making it more effective for the target users, consequently aiding the performance of public bodies, and making information more transparent and available to society.

The remainder of this paper is organized as follows. Section 2 introduces the fundamentals of NER. Section 3 discusses the related work. Section 4 details the proposed NLP pipeline and its constituent steps. Section 5 provides the experimental setup and analyzes the results. Section 6 presents a web application that implements the proposed approach. Finally, Section 7 summarizes the conclusions of this research and outlines future directions.

2 Background

This section explains state-of-the-art methods for NER tasks along with related concepts. This research is based on the principles of "Organizational Processing Information Theory" [11] which provides resources for managing information to enhance organizational efficiency and reduce uncertainty in processes.

2.1 Named Entity Recognition

NER is a well-explored NLP task in the literature focused on identifying and categorizing entities in text into specific tags [17], allowing the transformation of unstructured natural language text into structured data that can be stored in databases. Various models have been widely explored for NER, including Bidirectional Long Short-Term Memory (BiLSTM), Conditional Random Fields (CRF), Bidirectional Encoding Representations for Transformers (BERT), and Convolutional Neural Networks (CNN).

The following text illustrates the categorization of entities using common tags in SEI:

The $\overbrace{\textit{Dean of Community Affairs,}}^{\text{Position}}$ *in the exercise of his duties and considering the provisions of* $\overbrace{\textit{Article 143 of Law 8.112/90}}^{\text{Article number}}$ *and contained in Process* $\overbrace{\textit{23198.123456/2023-63}}^{\text{SEI's Process ID}}$, resolves...

2.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are architectures designed for sequential data, including text, time series, and biological data.

[2] https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

RNNs capture the context of the input sentence by analyzing each word in its natural order and encoding information from previous inputs into hidden states, allowing the network to retain past dependencies. Recognizing word order is important for capturing semantic insights [1].

RNNs typically consider only the information related to the past inputs of the current position of a sentence. However, in some cases, information about future states can be useful, and, for that reason, Bidirectional RNNs (BiRNNs) were created. Bidirectional networks consist of two independent layers that process the same input sequence in both forward and backward directions. As a result, the hidden state representations from both directions are combined by concatenation or addition.

Alternatively, Long Short-Term Memory (LSTM) is a type of RNN widely studied in text interpretation tasks. LSTM balances the preservation of long-term memory, the influence of short-term memory, and generalization capability [12]. These advancements are especially useful for tasks involving sequential data. Similar to BiRNNs, LSTM can also benefit from bidirectional processing, resulting in a Bidirectional Long Short-Term Memory (BiLSTM) network.

2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) were introduced in 1989 by LeCun et al. [16]. A CNN is a feedforward neural network that uses convolutional kernels to extract features from data. Initially, this architecture was primarily explored for image-processing tasks [18], including image classification, object recognition, detection, and localization.

CNN-based architectures typically encompass convolutional, pooling, and ReLU layers, followed by a fully connected layer that generates the output.The convolutional layer applies convolutional operations using kernels to filter the input data, enabling feature extraction and producing an activation map. In contrast, the pooling layer reduces the spatial dimensions of the activation map.

In the past decades, CNNs have also been studied in the context of text processing. Similar to images, which are represented as two-dimensional objects with depth defined by the number of color channels, a sentence is treated as a one-dimensional object with depth determined by its representation dimensionality [1]. In NER tasks, CNNs have been explored as an alternative method to enhance computational efficiency and performance in English and Chinese, as shown by Shen et al. (2017) [25], who proposed the CNN-CNN-LSTM architecture, and by Gui et al. (2019) [9], who introduced the Lexicon Rethinking CNN (LR-CNN) architecture.

2.4 Transformers

In 2017, Vaswani et al. [33] proposed a novel network architecture named Transformer. It comprises a pair of encoder and decoder stacks composed of attention mechanisms, normalization layers, and a softmax function in the output layer that generates probabilities as the model output.

In order to capture information from both characters and sentence context, Zhu et al. [34] proposed in 2019 a Convolutional Attention Network (CAN), which is character-based CNN with attention-based layers, to avoid segmentation errors and out-of-vocabulary problems, that are common in Chinese NER.

2.5 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. [6], is a Pre-trained Language Model (PLM) built upon the Transformer architecture. It is based on the principle that pre-trained representations provide a more efficient alternative to task-specific architectures that can reduce development time by leveraging prior training on large datasets.

BERT implementation can be explained in two main steps: pre-training and fine-tuning. The pre-training step involves training the model on unlabeled data within Next Sentence Prediction and Masked Language Modeling tasks. The fine-tuning step involves specializing the model in specific supervised tasks. Bidirectional representations are considered during the training process.

2.6 Conditional Random Fields

Conditional Random Fields (CRF) is a probabilistic model that learns weights in order to maximize the conditional probability of the correct label sequence given an input sequence [28]. CRF was proposed by Lafferty et al. [15] in 2001 and can still be considered as a method with satisfactory results nowadays [20].

3 Related Works

To the best of our knowledge, few studies in the literature investigate information extraction from text documents obtained from EIS, whereas NER remains a widely researched topic. This section discusses some selected works that addressed similar tasks involving the structuring of electronic system data and NER.

3.1 Information Extraction from EIS Documents

Regarding information extraction from text documents obtained from electronic systems, the research by Juez-Hernandez et al. [13] and Varagnolo et al. [31] can be mentioned. However, the corpora are not in Portuguese and their texts have different characteristics compared to Brazilian SEI.

Juez-Hernandez et al. presented AGORA [13], a tool that anonymize sensitive data in text documents by identifying named entities. The main features are anonymize, extract, and visualize named entities and the methodology behind its solution is based on NER models. Varagnolo et al. [31] presented a tool that extract information from web documents based on visualization.

3.2 Named Entity Recognition

Guimarães et al. [10] designed a tool to classify acts and perform NER on the official gazette of the Federal District of Brazil. It combines rule-based text classification with Machine Learning for NER, comparing three models: CRF, CNN-CNN-LSTM, and CNN-BiLSTM-CRF. This Python CLI tool enhances public transparency by allowing users to track government actions, contracts, and procurements. Furthermore, it can be refined for integration with other

public information processing systems, such as those related to topics of interest.

Trends in NER have been reviewed and applied to Indonesian datasets by Budi and Suryono [3] in tasks such as news extraction, flood monitoring, and traffic analysis. Their study highlights the need for improved data collection, cleaning, and model selection in NER tasks. Theoretical and practical applications include detecting illegal Fintech entities from social networks. However, the limited number of studies on Indonesian corpora suggests an opportunity for future research to develop models and libraries that account for the region's unique linguistic challenges.

Araujo et al. [19] created a Portuguese corpus, LeNER-Br, composed of 66 legal documents from Brazilian courts and four legislative documents for NER. The documents were segmented into sentences, each associated with a token. They then applied the IOB tagging scheme with labels for persons, places, time entities, organizations, legislation, and legal cases. Experiments were conducted, and model performance was evaluated using the F1-score, demonstrating that an LSTM-CRF model trained with this corpus achieved a satisfactory average F1-score of 92%.

Aiming to improve the NER system for the Romanian legal domain, Păis et al. [32] made a manually annotated corpus making an annotation process in which each annotator was responsible for 100 Romanian documents, considering the following tags: person, location, organization, time expressions, and legal document references. A crawler was used to extract the documents from their source and reduce manual labor. Experimental results showed that the CRF model yielded an improvement when compared to the previous NER system, with an 84% F1-score, and proved to be useful in improving the NER system.

To the best of our knowledge, the SEI documents present specific aspects that have not been addressed in the literature, such as the diversity of entities and the variations in writing style across different document types. To fill this gap and inspired by previous NER applications in the literature, this research aims to investigate information extraction from the Brazilian Electronic Information System so that unstructured text can be transformed into structured or semi-structured formats. This will enable the development of more efficient search approaches within SEI.

4 Methodology

This section presents an NLP pipeline to recognize named entities to transform unstructured information from SEI publications to a structured format, enabling more effective and efficient information searching, as presented in Figure 1. Subsection 4.1 describes process of information extraction from the website. Subsection 4.2 describes the process of annotation and validation to create the SEI corpus. Subsection 4.3 details the experiments conducted to decide the most suitable model to the NLP pipeline and Subsection 4.3.3 presents the evaluation metrics applied on the experiments.

The proposed pipeline consists of the following main steps: data extraction from the SEI website, corpus annotation of NER entities, corpus validation, information extraction using NER models, storage of structured information in an SQL database, and an entity-based search engine. The proposed methodology is illustrated in Figure 1.

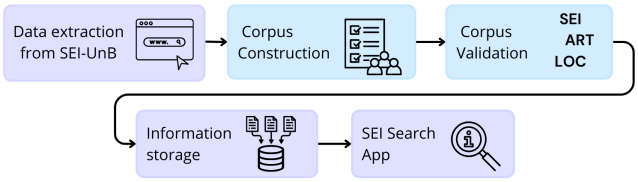


Figure 1: The proposed NLP pipeline with its constituent steps, in which the NLP tasks, corpus construction and validation, are highlighted.

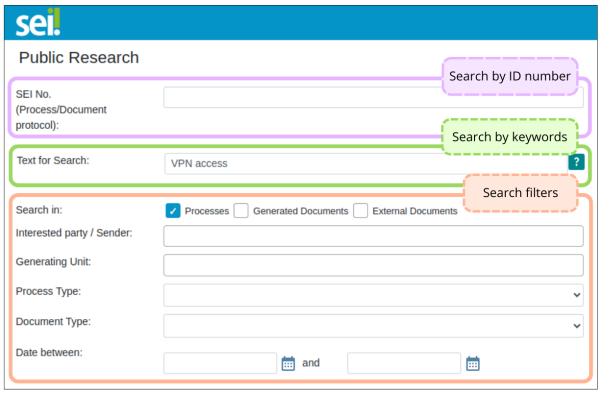


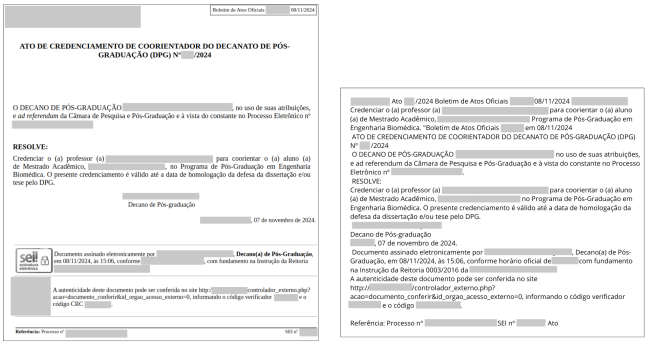
Figure 2: Public search for processes and documents in SEI (translated from Portuguese by Google Translate).

4.1 Data Extraction

A web crawler was developed to collect publications from the Brazilian Electronic Information System of University of Brasília (UnB). To achieve this, a Python script with the Selenium API was implemented. The script takes the URL of the UnB's public document search page as input and allows it to retrieve publications from UnB's departments and divisions over different periods. Figure 2 depicts the public webpage of UnB's SEI, highlighting the fields available in its search engine that was considered to set the web crawler. Users can search by process ID (purple rectangle), keywords using an exact-match approach (green rectangle), or apply filters (orange rectangle).

Although the SEI system at the University of Brasília contains various types of documents, a selection was made based on their frequency and significant presence in the public SEI collection. The chosen document types are described below:

- **Act *(Ato)***: a public decision, usually establishing rules or outlining specific actions to be taken. Examples of relevant entities for this type include the SEI process ID, names of individuals, document number, location, and beginning and ending dates, etc;
- **Standard Announcement *(Edital)***: a public notice describing about a specific event or opportunity. Some relevant entities include regulation items, department names, type of standard announcement, position or subject of the announcement, SEI process ID, etc.;



(a) Web page of a SEI document.

(b) Extracted text.

Figure 3: Text extraction with web crawler.

- **Resolution *(Resolução)***: a public decision by the university that may establish policies or rules. Examples of relevant entities include process ID, document number, date of issue, meeting number, among others;
- **Leave *(Afastamento)***: a formal document approving or revoking time off requested by a civil servant. Some of the targeted entities are names of individuals, location, department names, beginning and ending dates, and civil servant position, etc.

The complete list of entities per type of SEI document is provided in our repository[3].

Figure 3 illustrates the process of text extraction from an SEI document. Figure 3a displays an original SEI document as viewed by users on the public SEI website of UnB. The title of the publication is located at the top of the document and includes key information such as the publication number and the responsible department. The content region consists of the text between the title and the two rows at the bottom, which are associated with the electronic signature and the document's authenticity. The NER model is applied to both the title and the content, as these sections contain the majority of the relevant information, including names, registration IDs, dates, secondary process IDs, and document subjects, among others. Figure 3b presents the extracted raw text from the full publication.

4.2 Corpus Construction

Although the literature presents some public corpora in Portuguese [24] [7] [19] [26] [2], their underlying documents do not present similar characteristics to the documents in SEI. This limitation motivated the creation of a specific corpus for SEI, consisting of annotated SEI documents according to predefined entities. The annotation process involved manually labeling text spans by trained annotators, focusing on the targeted entities for each document type.

4.2.1 Annotation Guidelines. To annotate the SEI corpus, the open-source tool LabelStudio[4] was used as an interface for annotators to label the assigned documents. Five annotators participated in this process, performing both annotation and peer review in accordance with the annotation guidelines. In the first round, each annotator independently labeled their assigned SEI documents. In the peer review round, a different annotator reviewed the annotations to resolve disagreements and reduce bias. LabelStudio allows an annotator to associate a term in the raw text with a tag and also allows simultaneous accesses, which is essential to accelerate the both annotation and review. This method was chosen to certify the quality of the corpus.

4.2.2 Nested Entities. Annotators were instructed to identify and label nested entities when specified in the guidelines as some cases are exemplified below. Example (1) illustrates that may exist an organization inside a full position, and example (2) illustrates a date inside a Brazilian Official Gazette information.

(1) *The* Position[*Dean of Community Affairs of the* Organization[*University of Brasília*]]*.*

(2) *DOU n. 66 de* Brazilian Official Gazette information[*05 de abril de 2017* Date]*, seção 1, página 13.*

The SEI Corpus contains 293 annotated documents in Inside-Outside-Beginning (IOB) tagging format and 143, 230 tokens, each annotated as one of 23 entities. Tokens that do not represent an entity were tagged as 'O'. The number of documents selected for annotation was determined by the limited availability of annotators, as manual labeling is complex, costly, and prone to errors, requiring a peer review of the annotations. Each document has one specific type according to its purpose, such as Act, Standard Announcement, Resolution, or Leave. The dictionary with each NER tag and its meaning is available in our repository.

Table 1 presents a statistical comparison of token and sentence counts across different document types in the SEI Corpus. In general, Resolution documents are the longest, while Leave documents are the shortest. It is worth noting that the corpus presents an imbalance in entity distribution, as shown in Figure 4.

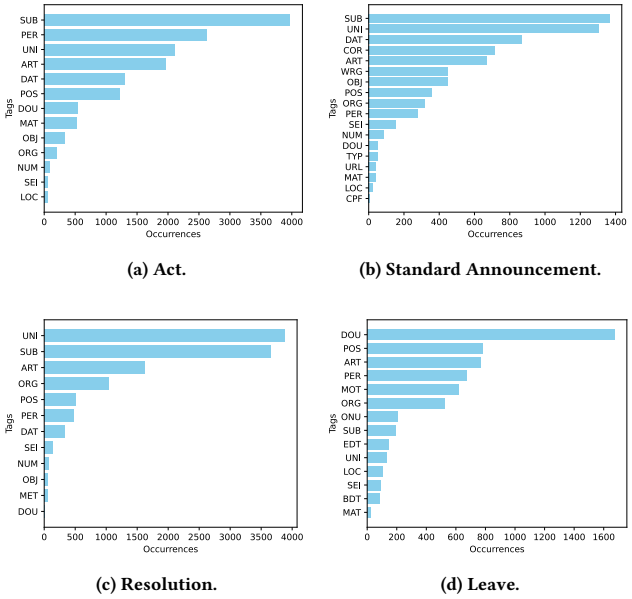
Visualizations can provide valuable insights in data analysis and interpretation. Bearing this in mind, t-SNE visualizations [30] were created to highlight some important characteristics of each document type present in the corpus, as illustrated by Figure 5. t-SNE is a non-linear dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space attempting to preserve local data structures rather than exact distances or global positioning. As a result, closer points generally suggest higher similarity between the underlying documents, though this interpretation is limited to local neighborhoods [8].

To generate the visualizations, the input texts were processed with the following steps: removal of "not a number" (nan) values and stopwords, conversion to lowercase, and stemming to reduce words to their base form by removing suffixes. These operations were

[3]https://gitlab.com/gvic-unb/sbsi-2025-could-you-tell-me-the-process-id
[4]https://labelstud.io/

Table 1: Basic statistics of SEI Corpus.

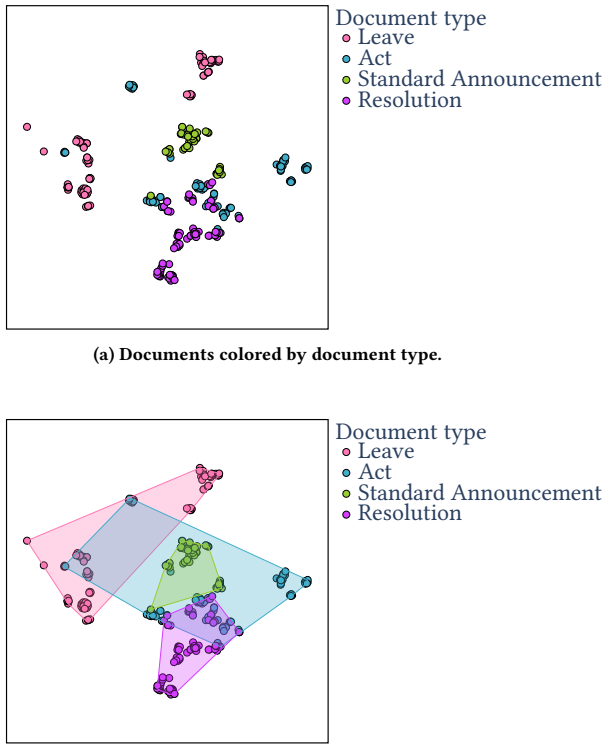
Document type	Tags	Annotated entities	Tokens	Documents	Avg. Tokens/Document
Act	13	1500	28003	84	333.37
Standard Announcement	18	7237	29172	48	607.75
Resolution	12	11835	67364	79	852.71
Leave	14	6020	18691	82	227.94



(a) Act. **(b) Standard Announcement.**

(c) Resolution. **(d) Leave.**

Figure 4: Tag frequency in SEI documents.



(a) Documents colored by document type.

(b) Documents delimited by polygons.

Figure 5: SEI document visualization using point placement.

performed using the NLTK library[5]. Subsequently, term frequency-inverse document frequency (tf-idf) was computed to transform the document texts into numerical feature vectors, capturing the importance of terms across the collection. After that, cosine distance between each pair of documents was computed to measure their similarity in the original high-dimensional space. Finally, t-SNE was applied to generate the final point-placement layout based on the computed dissimilarity matrix.

Figure 5(a) illustrates key characteristics of the SEI Corpus. In point-placement visualizations, each point represents a SEI document, and its position reflects local relationships among documents.

The dispersion of points in Act documents suggests a generic structure of this type of document, with similar structures to Resolutions, as illustrated by the point positioning in Figure 5(b), due to the legal nature of the sentences, with laws, regulations, and decrees in almost the entire text. Some Leave documents are directed to the Dean of People Management and others to the Rector, and this destination not only implies different titles but also indicates a slightly different information arrangement, reflecting on the generation of two visual subgroups. Finally, Leaves, Standard Announcements, and Resolutions are visually separated as illustrated by the overlapping polygons, meaning their structures are consistent in documents of each type and distinct from each other.

4.3 Corpus Validation

NER experiments were conducted to evaluate and determine an appropriate method for entity recognition for each document type individually. These experiments also established a baseline using state-of-the-art NER models to validate the SEI Corpus. All experiments were executed on an NVIDIA A100 GPU, and the proposed pipeline was implemented in Python using TensorFlow, Scikit-Learn, and Pandas.

4.3.1 Evaluation Strategies. Hyperparameter optimization for the NER models was performed using the Holdout strategy, chosen for its simplicity and efficiency, where the full corpus was split

[5]https://www.nltk.org/

into training (70%), validation (10%), and test (20%) sets. The performance evaluation strategy for the NER models with the optimal hyperparameters was based on the Stratified Group K-Fold method with $K = 5$, which divides the corpus into five stratified folds with non-overlapping groups for training and testing while preserving tag distributions. The training set is used to build a model that generalizes well, while the test set assesses its performance on unseen data, enabling further statistical analysis.

4.3.2 NER Models Setup. Hyperparameter optimization is an important step to improve the performance of deep neural networks. The Adam algorithm [14] was used in the training step. Adam is widely used in the literature and was chosen because it improves convergence by computing individual adaptive learning rates for different parameters from moments of the gradient estimation, being a better option in our context of limited computational resources. The hyperparameters tuning in the language models was performed with KerasTuner[6] and in the CRF model with RandomizedSearchCV[7]. Dropout layer and the Early Stopping, with patience equal to 10, were techniques employed to avoid overfitting to validation data. Therefore, the optimization was made with the validation data set from the search space predefined below:

- **CRF:**
  - Regularization parameter c1: exponential distribution with scale of 0.5;
  - Regularization parameter c2: exponential distribution with scale of 0.05.
- **BiLSTM:**
  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Droupout rate: [0.1,0.2,0.3];
  - Learning rate: [10^{-2},5.10^{-3}].
- **CNN-BiLSTM:**
  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Droupout rate: [0.1,0.2,0.3];
  - Learning rate: [10^{-2},5.10^{-3}].
- **BERT**
  - Model: bert-base-portuguese-cased [27];
  - Dropout rate: [0.1,0.2];
  - Learning rate: [10^{-4}, 5.10^{-5}].

4.3.3 Evaluation Metrics. The F1-score was used to evaluate the results, as it is a more suitable metric for unbalanced data, representing the harmonic mean of precision and recall. In this context, instances are classified as positive or negative based on the presence or absence of the target entity. Precision measures the proportion of correctly classified positive instances among all instances predicted as positive, as shown in Equation (1). Recall quantifies the proportion of correctly classified positive instances relative to all actual positive instances in the corpus, as given in Equation (2). Finally, the F1-score is defined in Equation (3).

$$precision = \frac{TP}{TP + FP}, \tag{1}$$

$$recall = \frac{TP}{TP + FN}, \tag{2}$$

[6]https://keras.io/keras_tuner/

[7]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

where TP is the quantity of correct positive predictions of an entity, FP is the quantity of wrong positive predictions of an entity, TN is the quantity of correct negative predictions of an entity, FN is the quantity of wrong negative predictions of an entity.

$$F1\text{-}score = \frac{2 \times precision \times recall}{precision + recall}. \tag{3}$$

The macro F1-score was used in the experiments to evaluate the models' performance across all tags equally, as SEI Corpus presents multiple entity labels with an imbalanced distribution.

5 Results and Discussion

Experiments were conducted on the SEI Corpus to quantitatively evaluate the results of the NER models of the proposed methodology. The source code and the SEI corpus are available in our repository. An overview of the models performance on each document type is presented in Table 2. In these experiments, three NER models were compared: CNN-BiLSTM, BERT, and CRF, selected based on their successful prior use in the literature.

In order to evaluate the performance differences between the models, the non-parametric Friedman test was conducted, considering a significance level of $\alpha = 5\%$. To enable this statistical analysis, the folds were generated under the same conditions, with the same random seed. The following null and alternative hypotheses were formulated:

H_0: No differences between all the models average macro F1-scores.

H_a: At least one model average macro F1-scores differs from the others.

In the Friedman test, the p-value for Act, Standard Announcement, Resolution and Leave documents were 0.00285, 0.00285, 0.00182, and 0.00698, respectively. Thus, the null hypothesis is rejected, and a Nemenyi pairwise comparison was performed with a significance level of $\alpha = 5\%$. The significant results are shown in Table 3.

Table 2: Macro F1-score results for NER task in SEI Corpus using Stratified K-Fold Cross Validation (K = 5): average and standard deviation of Macro F1-Score across all folds.

Document type	CRF	BiLSTM	CNN-BiLSTM	BERT
Act	0.69±0.07	0.29±0.14	0.45±0.07	**0.74±0.04**
Standard Announcement	**0.54±0.06**	0.30±0.07	0.09±0.07	0.23±0.02
Resolution	0.71±0.09	0.44±0.06	0.22±0.14	**0.77±0.04**
Leave	**0.93±0.01**	0.54±0.03	0.57±0.06	**0.93±0.01**

According to the Nemenyi test, CRF and BERT showed a superior performance than BiLSTM for Act documents, CRF outperformed CNN-BiLSTM for Standard Announcement documents, BERT outperformed CNN-BiLSTM in Resolution documents, and BERT also outperformed BiLSTM for Leave documents. These results indicate that statistically significant differences in F1-scores were found only for the model comparisons listed above.

Table 3: Nemenyi post-hoc test significant results.

Document type	Pairwise Model Comparison	P-Value
Act	CRF vs BiLSTM	0.036
Act	BERT vs BiLSTM	0.003
Standard Announcement	CRF vs CNN-BiLSTM	0.001
Resolution	BERT vs CNN-BiLSTM	0.001
Leave	BERT vs BiLSTM	0.036

It is interesting to note that Leave documents achieved the highest results among all document types, followed by Act documents. These document types contain fewer tokens per document, as shown in Table 1, with averages of 227.94 and 333.37 tokens, respectively. In other words, these documents typically present concise and direct information, yielding better entity recognition.

In contrast, Standard Announcement documents had the lowest F1-score in the experiments. This result may be attributed to the nature of their content, as announcements can encompass diverse documents establishing policies or rules for events and opportunities that are relevant to students, professors, and university staff. Consequently, these documents can vary significantly in length and structure, which may have negatively impacted entity recognition.

Given that each document type has its own entities and particular characteristics, the discussion should also address these details. Tables 4, 5, 6, and 7 provide a more detailed view of the models' effectiveness for each entity.

In Act documents, shown in Table 4, almost all entities related to dates (DAT), positions (POS), numbers (NUM, DOU, ART), and organization (ORG) entities were well predicted by CRF or BERT. However, the 'object' (OBJ) entity of an Act was not correctly detected despite the approximate 200 occurrences in the corpus. This intriguing result was also observed in Standard Announcements, meaning that this type of information is more challenging to obtain. The subjectivity and personal writing styles of publications in the same corpus may have further influenced this outcome.

Table 4: Mean and standard deviation of the F1-score for the NER task on Act documents using Stratified K-Fold Cross-Validation ($K = 5$).

Tag	CRF	BiLSTM	CNN-BiLSTM	BERT
ART	0.84±0.05	0.51±0.23	0.68±0.09	**0.90±0.04**
DAT	0.95±0.01	0.65±0.30	0.78±0.06	**0.96±0.02**
DOU	0.88±0.07	0.66±0.33	0.77±0.09	**0.96±0.01**
LOC	**0.97±0.03**	0.66±0.34	0.87±0.06	0.80±0.06
MAT	**0.45±0.29**	0.01±0.01	0.09±0.10	0.37±0.33
NUM	**0.96±0.03**	0.00±0.00	0.84±0.06	0.90±0.06
OBJ	0.00±0.00	0.00±0.00	0.00±0.00	**0.22±0.13**
ORG	**0.68±0.15**	0.15±0.09	0.34±0.09	0.42±0.16
PER	0.41±0.29	0.21±0.15	0.17±0.08	**0.73±0.04**
POS	0.87±0.13	0.53±0.28	0.50±0.26	**0.92±0.07**
SEI	0.89±0.08	0.01±0.01	0.32±0.22	**0.91±0.09**
SUB	0.41±0.09	0.10±0.05	0.09±0.06	**0.75±0.05**
UNI	0.66±0.14	0.28±0.16	0.38±0.20	**0.74±0.03**

In Standard Announcement and Resolution documents, described in Tables 5 and 6, the subject of the announcement (SUB) achieved low F1-scores, indicating a significant misclassification rate despite being present in all analyzed announcements and resolutions. In Leave documents (Table 7), the SUB entity achieved an F1-score of 0.99. However, there are only two options for leave: authorization and cancellation. In contrast, announcements and resolutions present a wider variety of subjects.

Table 5: Mean and standard deviation of the F1-score for the NER task on Standard Announcement documents using Stratified K-Fold Cross Validation ($K = 5$).

Tag	CRF	BiLSTM	CNN-BiLSTM	BERT
ART	**0.49±0.20**	0.29±0.18	0.20±0.12	0.20±0.08
COR	0.13±0.17	**0.39±0.29**	0.00±0.00	0.00±0.00
CPF	**1.00±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
DAT	0.61±0.06	0.39±0.07	0.14±0.13	**0.67±0.08**
LOC	**0.50±0.30**	0.10±0.13	0.00±0.00	0.00±0.00
MAT	**0.33±0.37**	0.25±0.43	0.00±0.00	0.00±0.00
NUM	**0.81±0.05**	0.55±0.05	0.33±0.23	0.33±0.29
OBJ	**0.32±0.22**	0.00±0.00	0.00±0.00	0.09±0.14
ORG	**0.67±0.09**	0.34±0.18	0.00±0.00	0.00±0.00
PER	**0.59±0.21**	0.49±0.21	0.00±0.00	0.28±0.20
POS	**0.72±0.12**	0.43±0.19	0.11±0.14	0.10±0.04
SEI	0.91±0.11	0.50±0.26	0.26±0.22	**0.97±0.03**
SUB	0.31±0.07	0.17±0.04	0.06±0.05	**0.48±0.08**
TYP	**0.87±0.12**	0.56±0.14	0.14±0.24	0.00±0.00
UNI	**0.67±0.09**	0.45±0.10	0.13±0.11	0.30±0.05
URL	**0.40±0.26**	0.25±0.43	0.00±0.00	0.00±0.00
WRG	**0.17±0.22**	0.11±0.10	0.00±0.00	0.16±0.14

Table 6: Mean and standard deviation of the F1-score for the NER task on Resolution documents using Stratified K-Fold Cross Validation ($K = 5$).

Tag	CRF	BiLSTM	CNN-BiLSTM	BERT
ART	0.82±0.07	0.68±0.06	0.32±0.16	**0.83±0.06**
POS	**0.82±0.05**	0.54±0.14	0.31±0.19	0.59±0.06
DAT	0.75±0.11	0.46±0.09	0.07±0.13	**0.77±0.21**
MET	0.97±0.04	0.72±0.11	0.03±0.06	**0.98±0.02**
NUM	0.92±0.03	0.24±0.05	0.12±0.13	**0.93±0.04**
ORG	**0.74±0.07**	0.56±0.10	0.19±0.03	0.58±0.09
PER	0.68±0.22	0.56±0.18	0.27±0.22	**0.93±0.05**
SEI	0.74±0.24	0.26±0.12	0.12±0.11	**0.92±0.05**
SUB	0.37±0.07	0.14±0.02	0.05±0.07	**0.76±0.03**
UNI	0.52±0.11	0.28±0.10	0.18±0.09	**0.67±0.04**

Table 7 shows the results of the SEI's Leave documents. It can be observed that there is only one entity predicted with an F1-score below 0.80, the justification for the leave requested (MOT). This entity typically comprises actions such as participation in conferences, educational activities, travel, and technical training. The wide variety of justifications appears to affect the learning

patterns of this entity. A possibility to improve these outcomes is to provide more samples of the Leave document type to the NER models.

Table 7: Mean and standard deviation of the F1-score for the NER task on Leave documents using Stratified K-Fold Cross Validation ($K = 5$)**.**

Tag	CRF	BiLSTM	CNN-BiLSTM	BERT
ART	0.84±0.07	0.73±0.17	0.81±0.08	**0.95±0.02**
BDT	**0.94±0.10**	0.64±0.17	0.58±0.12	**0.94±0.10**
DOU	**0.99±0.01**	0.79±0.04	0.82±0.07	0.95±0.02
EDT	0.96±0.03	0.17±0.10	0.53±0.21	**0.97±0.04**
LOC	**0.84±0.13**	0.03±0.05	0.06±0.08	0.79±0.06
MAT	0.93±0.13	0.00±0.00	0.00±0.00	**1.00±0.00**
MOT	0.77±0.11	0.43±0.03	0.35±0.21	**0.78±0.03**
ONU	**0.92±0.05**	0.84±0.13	0.52±0.30	**0.92±0.04**
ORG	**0.96±0.02**	0.76±0.07	0.77±0.06	0.95±0.03
PER	**0.97±0.02**	0.58±0.02	0.61±0.07	**0.97±0.02**
POS	**0.98±0.01**	0.81±0.06	0.74±0.06	0.96±0.03
SEI	**0.98±0.01**	0.02±0.04	0.57±0.35	**0.98±0.03**
SUB	**0.99±0.03**	0.96±0.04	0.77±0.21	**0.99±0.02**
UNI	**0.89±0.05**	0.79±0.16	0.87±0.07	0.82±0.09

Finally, the experiments validate the SEI Corpus and demonstrate that CRF and BERT achieved the highest macro F1-scores, with CRF presenting the shortest training time. This makes CRF the most suitable model for production, as it requires fewer computational resources compared to the other models analyzed.

6 SEI Search App

SEI Search App is a web application that simulates the interface between SEI users and the result of the NLP pipeline. The technologies used in the development of this application include Python with Streamlit [8] for frontend and MySQL [9] to store the documents in a structured format. The SEI Search App is available in our repository.

A screenshot of the web application is presented in Figure 6. The main page contains filters and text input to search by entities in documents, and the other page has the original text of the selected document with highlighted entities. Entities are defined based on CRF model predictions.

7 Conclusion

This paper introduced an NLP pipeline for extracting information from documents of the Brazilian Electronic Information System (SEI). Specifically, a NER-based approach was chosen to identify key entities in the unstructured texts, enabling their transformation into a structured format. A corpus of SEI documents from four distinct document types was created and manually annotated by five trained annotators. Experiments were conducted to validate the corpus and compare the performance of three state-of-the-art NER models, with CRF and BERT achieving the best overall macro F1-Score.

[8]https://streamlit.io/
[9]https://www.mysql.com/

The manual annotation approach provided reliable entities in most of document types that are specific for the complex nature of the SEI documents. However, the results also showed that the more complex ones require additional samples to enhance the representativeness of entities, once the diversity of texts in some cases affected their appropriate recognition by NER models.

The main contributions of this paper include the SEI Corpus, a new labeled corpus of documents in SEI for NER tasks in the Brazilian Portuguese language, an NLP pipeline to search for information in SEI, and the SEI Search App, a Web Application that allows searching SEI documents by entities. The findings in this research provide possibilities for developing effective searching and retrieving documents on information systems.

For future work, we emphasize the investigation of combinations of prompt-based, weak supervision, and human annotation to accelerate corpus annotation and decrease the costs as shown by the research of Oliveira et al. [21]. In this sense, Large Language Models as Open AI's Chat GPT (Generative Pre-trained Transformers) [22], Google's PaLM (Pathways Language Model) [5], and Meta's Llama (Large Language Model Meta AI) [29] are potential models to be further considered.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 88887.967792/2024-00, and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Process Number 800019/2024-5.

References

[1] Charu C. Aggarwal. 2018. *Neural Networks and Deep Learning*. Springer. https://doi.org/10.1007/978-3-319-94463-0
[2] Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Computational Processing of the Portuguese Language*. Springer International Publishing, Cham, 3–14.
[3] Indra Budi and Ryan Randy Suryono. 2023. Application of named entity recognition method for Indonesian datasets: a review. *Bulletin of Electrical Engineering and Informatics* 12, 2 (2023), 969–978. https://doi.org/10.11591/eei.v12i2.4529
[4] Ana Camarinha, António Abreu, Marcelo Júnior, and Ivone Cardoso. 2023. Users' perception of satisfaction of the eletronic information system – SEI in the Instituto Federal de Rondônia. *Journal of Information Systems Engineering and Management* 8 (January 2023), 18354. https://doi.org/10.55267/iadt.07.12744
[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian et al. Gehrmann. 2024. PaLM: scaling language modeling with pathways. *The Journal of Machine Learning Research* 24, 1 (March 2024).
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805
[7] Mariana Dias, João Boné, João C. Ferreira, Ricardo Ribeiro, and Rui Maia. 2020. Named Entity Recognition for Sensitive Data Discovery in Portuguese. *Applied Sciences* 10, 7 (2020). https://doi.org/10.3390/app10072303
[8] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. 2021. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics* 27, 3 (2021), 2153–2173. https://doi.org/10.1109/TVCG.2019.2944182
[9] Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. CNN-Based Chinese NER with Lexicon Rethinking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 4982–4988. https://doi.org/10.24963/ijcai.2019/692

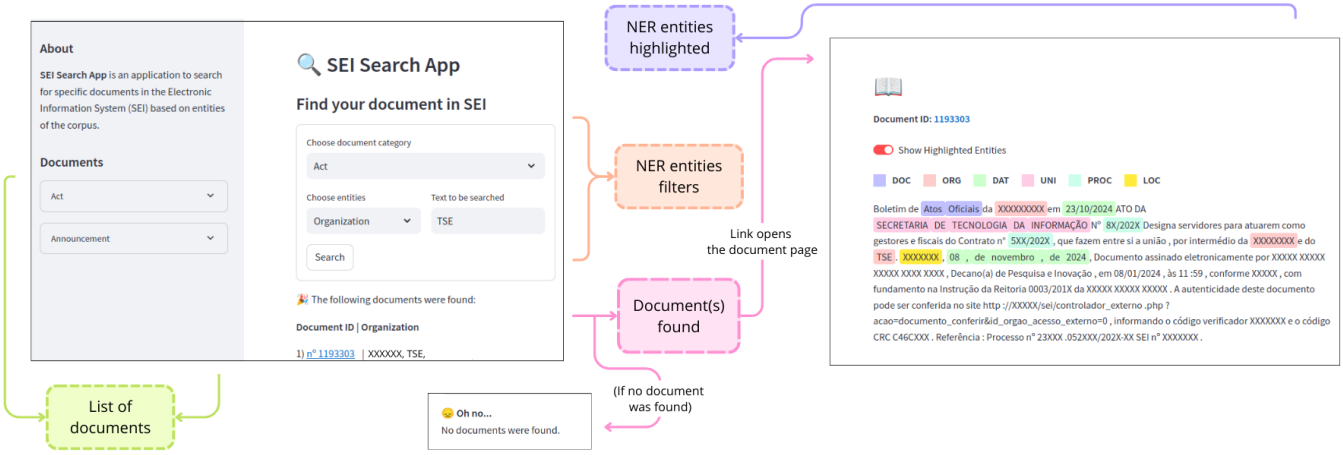


Figure 6: SEI Search App.

[10] Gabriel M.C. Guimarães, Felipe X.B. da Silva, Andrei L. Queiroz, Ricardo M. Marcacini, Thiago P. Faleiros, Vinicius R.P. Borges, and Luís P.F. Garcia. 2024. DODFMiner: An automated tool for Named Entity Recognition from Official Gazettes. *Neurocomputing* 568 (2024), 127064. https://doi.org/10.1016/j.neucom.2023.127064

[11] Clemens Hausmann, Yogesh K. Dwivedi, Krishna Venkitachalam, and Michael D. Williams. 2012. *A Summary and Review of Galbraith's Organizational Information Processing Theory*. Vol. 2. Springer New York, New York, NY, 71–93. https://doi.org/10.1007/978-1-4419-9707-4_5

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[13] Rodrigo Juez-Hernandez, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. AGORA: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. *Applied Soft Computing* 145 (2023), 110540. https://doi.org/10.1016/j.asoc.2023.110540

[14] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (December 2014).

[15] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1, 4 (1989), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

[17] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2022), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

[18] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2022. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems* 33, 12 (2022), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

[19] Pedro Henrique Luz de Araujo, Teofilo de Campos, Renato Oliveira, Matheus Stauffer, Samuel Couto, and Paulo De Souza Bermejo. 2018. *LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*. 313–323. https://doi.org/10.1007/978-3-319-99722-3_32

[20] Gabriel M. C. Guimarães, Felipe X. B. da Silva, Lucas A. B. Macedo, Victor H. F. Lisboa, Ricardo M. Marcacini, Andrei L. Queiroz, Vinicius R. P. Borges, Thiago P. Faleiros, and Luis P. F. Garcia. 2024. Legal Document Segmentation and Labeling Through Named Entity Recognition Approaches. *Journal of Information and Data Management* 15, 1 (February 2024), 123–131. https://doi.org/10.5753/jidm.2024.3368

[21] Vitor Oliveira, Gabriel Nogueira, Thiago Faleiros, and Ricardo Marcacini. 2024. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law* (February 2024), 1–21. https://doi.org/10.1007/s10506-023-09388-1

[22] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and Sam Altman et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[23] Aloir Pedruzzi Junior, Jonimar da Silva Souza, and Nubiana de Lima Irmão Pedruzzi. 2024. Sistema Eletrônico de Informações (SEI) como ferramenta para modernização da gestão documental na administração pública. *Revista de Gestão e Secretariado* 15, 1 (January 2024), 309–319. https://doi.org/10.7769/gesec.v15i1.3352

[24] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy.

[25] Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 252–256. https://doi.org/10.18653/v1/W17-2630

[26] Priscilla Silva, Arthur Franco, Thiago Santos, Mozar José de Brito, and Denilson Pereira. 2023. CachacaNER: a dataset for named entity recognition in texts about the cachaça beverage. *Language Resources and Evaluation* (2023), 1–19.

[27] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer International Publishing, Cham, 403–417.

[28] Charles Sutton and Andrew McCallum. 2012. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning* 4, 4 (April 2012), 267–373. https://doi.org/10.1561/2200000013

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv* abs/2302.13971 (2023).

[30] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[31] Davide Varagnolo, Dora Melo, and Irene Pimenta Rodrigues. 2021. A Tool to Explore the Population of a CIDOC-CRM Ontology. *Procedia Computer Science* 192 (2021), 158–167. https://doi.org/10.1016/j.procs.2021.08.017 Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 25th International Conference KES2021.

[32] Carol Luca Gasan Alexandru Ianov ă Corvin Ghit Vlad Silviu Coneschi Vasile Păiș, Maria Mitrofan and Andrei Onuț. 2023. LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain. *Miscellaneous* 15, 3 (2023), 831–844. https://doi.org/10.3233/sw-233351

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., 6000–6010.

[34] Yuying Zhu and Guoxin Wang. 2019. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3384–3393. https://doi.org/10.18653/v1/N19-1342

Received 21 November 2024; revised 4 February 2025; accepted 18 February 2025