University of Brasília

Institute of Exact Sciences
Department of Computer Science

Multilingual Named Entity Recognition: A Design Study on Academic and Institutional Documents

Reconhecimento de Entidades Nomeadas Multilíngues: Um Estudo de Design sobre Documentos Acadêmicos e Institucionais

Mayara Chew Marinho

Thesis presented in partial fulfillment of the requirements for the degree of Master of Science in Informatics

Advisor
Prof. Dr. Vinícius Ruela Pereira Borges

Brazil

2025

University of Brasília

Institute of Exact Sciences
Department of Computer Science

Multilingual Named Entity Recognition: A Design Study on Academic and Institutional Documents

Reconhecimento de Entidades Nomeadas Multilíngues: Um Estudo de Design sobre Documentos Acadêmicos e Institucionais

Mayara Chew Marinho

Thesis presented in partial fulfillment of the requirements for the degree of Master of Science in Informatics

Prof. Dr. Vinícius Ruela Pereira Borges (Advisor)
CIC/UnB

Prof. Dr.a Nádia Felix Felipe da Silva	Dr. Luís Paulo Faina Garcia
INF/UFG	CIC/UnB

Prof.a Dr.a Cláudia Nalon
Coordinator of the Graduate Program in Informatics

Brasília, August 22, 2025

Dedication

The most important acknowledgment I would like to make is to my advisor, Professor Vinícius Borges, who has guided me with dedication and patience since the middle of my undergraduate studies. He enriched my academic journey with much more than technical knowledge — he taught me how to apply what we learn for the benefit of society and, most importantly, how to pass it on from person to person.

I would like to thank Ana Clara B. Borges, Vanessa P. Costa, and Laryssa O. Ferreira for their invaluable contributions to the corpus annotation process. It's due to their effort that we were able to create a SEI golden standard corpus, one of the key contributions of this research. I am also grateful to Ana Carolina for her dedication to improving the graphical interface for NEVis. Her expertise was essential to give the final touch to the project. I would also like to express my gratitude to Prof. Dra. Nádia Félix and Prof. Dr. Luís Paulo, who generously offered their suggestions and guidance in the formation of my thesis committee. This collaboration was essential to the development of the research.

I am deeply grateful to my parents for their unwavering support throughout this academic journey. I am also truly thankful to my friends for being by my side throughout these two years of study, especially to Leonardo Ribas and Maria Clara Mendes, whose companionship and encouraging words gave me the strength to persevere until the end. Just an overpriced coffee, a Pomodoro timer, and the right company to make everything feel manageable.

Acknowledgements

The author would like to thank the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) for the financial support (Finance Code 88887.967792/2024-00) and the Access to Journals Portal.

Abstract

Academic and institutional documents play a central role in higher education institutions, serving as formal records of students' academic trajectories, institutional decisions, and regulatory compliance. Given the large volume of documents produced and stored over time, Named Entity Recognition (NER) can be an essential Natural Language Processing (NLP) task for extracting information from unstructured documents and improving search processes within electronic information systems in educational institutions. The goal of NER is is to identify and classify text spans according to predefined categories of real-world entities, enabling the conversion of raw text into a structured format. Brazilian academic documents may contain terms in different languages, such as international events, research related activities, and locations, steering research towards multilingual NER. In this context, this research addresses multilingual NER in academic documents in scenarios of long documents, limited availability of labeled data, and the presence of low frequency entities. Several NER approaches, encompassing the classical methods and those based on Large Language Models (LLMs), have their performances compared and evaluated using quantitative metrics. Moreover, the lack of publicly available academic documents required the construction of labeled corpora for multilingual NER. Experiments were conducted to evaluate the quality of the constructed corpora and to compare the performance of state-of-the-art NER models, including CRF, BiLSTM, CNN-BiLSTM, BERT, fine-tuned LLaMA, and fine-tuned DeepSeek. The results indicated that CRF and BERT achieved the best performance on the developed multilingual corpus, with macro F1-scores above 0.9. Krippendorff's Alpha and Cohen's Kappa metrics demonstrated that the entity labels are reliable and that the corpus has high quality. Finally, to enable the analysis of the predicted categories, a visualization tool for named entities was proposed to display NER and nested NER entities.

Keywords: Natural Language Processing, Named Entity Recognition, Multilingual Texts, Language Models, Transformers, Corpus Construction, Visualization

Resumo

Documentos acadêmicos e institucionais desempenham um papel importante nas instituições de ensino superior, uma vez que são registros formais das trajetórias acadêmicas dos estudantes, das decisões institucionais e do cumprimento de normas regulatórias. Dada a grande quantidade de documentos produzidos e armazenados ao longo do tempo, o uso de Reconhecimento de Entidades Nomeadas (NER) torna-se uma tarefa essencial de Processamento de Linguagem Natural para extrair informações de textos não estruturados e melhorar os processos de busca nos sistemas de informação dessas instituições. O objetivo do NER é identificar e classificar palavras de acordo com categorias predefinidas de entidades, permitindo-se transformar texto bruto em dados estruturados. Documentos acadêmicos podem conter palavras em mais de um idioma, como nomes de eventos internacionais, atividades de pesquisa e localizações, o que caracteriza o NER multilíngue. Esta pesquisa de mestrado aborda o NER multilíngue em documentos acadêmicos em cenários que envolvem textos longos, disponibilidade limitada de dados rotulados e presença de entidades de baixa frequência. Diversas abordagens de NER, incluindo métodos clássicos e modelos baseados em LLMs, são comparadas e avaliadas com o uso de métricas quantitativas. A inexistência de *corpora* de documentos acadêmicos rotulados disponíveis publicamente demandou a criação de *corporas* multilíngues anotados para NER. Foram realizados experimentos com o objetivo de avaliar a qualidade dos *corpora* construídos e comparar o desempenho de modelos NER, como CRF, BiLSTM, CNN-BiLSTM, BERT, LLaMA e DeepSeek ajustados. Os resultados indicaram que CRF e BERT apresentaram os melhores desempenhos no corpus multilíngue desenvolvido, com macro F1-score superior a 0,9. As métricas Krippendorff's Alpha e Cohen's Kappa demonstraram que os rótulos atribuídos às entidades são confiáveis e que o corpus possui alta qualidade. Por fim, uma ferramenta foi proposta para visualizar entidades nomeadas, inluindo entidades aninhadas, permitindo uma análise detalhada dos resultados dos modelos de NER.

Palavras-chave: Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Textos Multilíngues, Modelos de Linguagem, Transformers, Criação de Corpus, Visualização

Contents

List of Figures

List of Tables

List of Acronyms

BERT Bidirectional Encoder Representations from Transformers.

BiLSTM Bidirectional Long Short-Term Memory.

BPE Byte-Pair-Encoding.

CAPES Coordination for the Improvement of Higher Education Personnel – Brazil.

CNN Convolutional Neural Network.

CRF Conditional Random Fields.

DL Deep Learning.

EIS Electronic Information Systems.

HMM Hidden Markov Model.

LLM Large Language Model.

LLMs Large Language Models.

LoRA Low-Rank Adaptation.

LSTM Long Short-Term Memory.

ML Machine Learning.

MLM Masked Language Modeling.

NER Named Entity Recognition.

NLP Natural Language Processing.

NSP Next Sentence Prediction.

PLM Pre-trained Language Model.

PLMs Pre-trained Language Models.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

SEI Brazilian Electronic Information System.

t-SNE t-Distributed Stochastic Neighbor Embedding.

UnB University of Brasília.

Chapter 1

Introduction

Academic documents, or institutional documents, are formal records existing in institutions of education, research, or educational administration. They serve to certify, communicate, or regulate information related to the academic and administrative life of students, professors, and staff. These documents are commonly stored in internal Electronic Information Systems (EIS), such as the *Sistema Integrado de Gestão de Atividades Acadêmicas* (SIGAA) and the Brazilian Electronic Information System (SEI) (*Sistema Eletrônico de Informações*) (Bezerra et al. (2022)), while others are published through the institutions' own communication channels or through official outlets, such as government gazettes.

Although the literature does not explicitly define the types of these documents, it can be assumed that some have a legal nature, while others belong to the educational domain. Examples include academic transcripts, diplomas, course syllabi, certificates of participation, enrollment declarations, meeting minutes, ordinances, regulations, memos, forms, and circulars, among others. These documents may vary in format and type of files, including free text, scanned images, PDF files, and structured forms.

Educational institutions are constantly providing activities and services to both academic and external communities. As they must record these activities into their EIS, the efficiency of search processes are affected due to the growing volume of stored documents. In systems like SEI, this retrieval is typically performed using exact matching across the entire database. This forces users to make multiple interactions to find the desired information or the correspondent administrative process of interest. Consequently, administrative tasks become less productive for those who depends on these EISs, as academic documents vary in type and users are often required to perform multiple searches daily. This scenario shows the need for intelligent search mechanisms that can support users in accessing relevant content more efficiently.

Search processes in EISs can be improved by formulating indexed queries based on document-specific keys and attributes. However, this requires extracting relevant information from unstructured texts in a structured format, a task addressed by Named Entity Recognition (NER) in Natural Language Processing (NLP). NER is formally defined as the process of locating and classifying entities into predefined semantic categories (Li et al. (2022a)). The categories, also named entities, may vary depending on the domain and purpose of the application, with the most generic being people, organizations, and geographic locations (Grishman and Sundheim (1996)).

State-of-the-art NER approaches rely on probabilistic approaches such as Conditional Random Fields (CRF) (Lafferty et al. (2001)) and neural language models for sequence labeling, including Long Short-Term Memory (LSTM) and its variants (Hochreiter and Schmidhuber (1997)). More recently, Transformer-based architectures (Vaswani et al. (2017)) such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. (2019)), have been successfully employed in NER tasks, as these models incorporate the attention mechanism allows them to capture long-range dependencies and contextual relationships between tokens. Moreover, these models are pre-trained on large-scale labeled corpora and can be fine-tuned to perform effectively on domain-specific tasks. However, fine-tuning demands a labeled corpus associated to the target domain, presenting labeled entities according to predefined categories relevant to the specific domain.

Although NER has been extensively studied by the NLP community, academic documents present specific challenges due to their structural aspects and linguistic characteristics. These documents may be written entirely in Portuguese, entirely in English, or contain mixed-language content, with sections in both languages. This variability introduces additional difficulties related to tokenization and language-specific modeling (Mayhew et al. (2024)). Regarding the documents, there is a wide variety of formats (e.g., syllabi, certificates, memos) and multiple entity types, some of which share similar textual patterns (e.g., numbers of administrative processes and number of documents) or present ambiguity. Finally, many documents may contain few named entities that are overlapping, which might characterizing nested entities (e.g., course names within program names) or some type of hierarchy among them.

The current literature addresses many of the described challenges in NER through a variety of domain-specific approaches. Several works have focused on the biomedical (Durango et al. (2023); Hu and Ma (2023); Nunes et al. (2024)) and legal domains (Darji et al. (2023); Zanuz and Rigo (2022)), where the complexity of entities and the lack of annotated data in domain-specific tasks requires the modeling of sophisticated approaches based on contrastive learning (Das et al. (2022); Mo et al. (2024); Yang et al. (2024)), weak supervision (Wang et al. (2022a)), and few-shot learning (Moscato et al. (2023)). Regarding multilingual NER, cross-lingual transfer learning (Hazem et al. (2022)) and multilingual pre-trained language models such as mBERT (Abilio et al. (2024)) have shown promising results.

Large Language Models (LLMs) have recently achieved remarkable progress in tasks related to the generation of text in natural language and IE. Particularly NER, there is still progress to be made. Studies have evaluated the performance of ChatGPT and other LLMs on NER tasks, showing that existing specialized methods are still competitive using less computational resources (Lai et al. (2023); Laskar et al. (2023); Santos et al. (2024); Zhang et al. (2023d, 2025)). This gap shows the need to investigate both classical and LLM-based NER methods in academic documents, which exhibit specific linguistic structures and entity labels, as well as diverse formats and writing styles that differ from those found in previously explored domains such as biomedical, legal, or financial texts.

Although the performance of NER models can be evaluated using quantitative metrics, visualization techniques can be employed to support explainability and enable the analysis of results by generating intuitive graphical representations of model predictions over the underlying texts (Chatzimparmpas et al. (2020); Chefer et al. (2021); Tian et al. (2021); van den Elzen et al. (2023)). Such visualizations enable researchers and NLP practitioners to better understand NER models behavior and their performance across various entity types, such as dates, names, and numerical values. However, most existing NER visualizations are limited to flat entities or rely on simple highlighting strategies, such as those offered by displacy[1], which are limited for analyzing large documents containing multiple, fine grained or nested entities.

This master's research aims to explore multilingual NER models applied to academic documents, which are usually long and contain a wide range of entity types. To address challenges such as low-frequency entities and the limited availability of labeled data, this research investigate well-known NER models, including those based on transformers and LLMs. For this purpose, some corpora of academic and institutional documents were constructed by a manual annotation process and using SEI publications, including a multilingual corpus specifically composed of academic staff records. In the latter case, the quality of entity labels and their corresponding text spans was assessed using Krippendorff's alpha, Cohen's kappa, and Levenshtein distance.

Furthermore, in order to fill a gap in the visualization of named entities, this work proposes to extend existing approaches by developing interactive and nested representations. The proposed technique supports the visual analysis of NER results by displaying the recognized entities from a JSON file in their natural textual order as blocks within a rectangular visual space. Interactive mechanisms are provided to reveal the surrounding context of each entity.

[1] https://spacy.io/usage/visualizers#displacy

1.1 Research Questions

This research aims to address the following Research Questions (RQs):

- **RQ1:** Are there publicly available multilingual corpora in Portuguese and English for training Named Entity Recognition (NER) models? If not, how can multilingual corpora for NER tasks be curated?
- **RQ2:** How do state-of-the-art NER models perform on academic documents in both flat and nested NER tasks?
- **RQ3:** How do Large Language Models (LLMs) perform on NER tasks in fine-tuning scenarios when applied to academic documents?
- **RQ4:** Is it possible to effectively visualize NER outputs in long documents with multiple overlapping or nested entities?

1.2 Goals

The goal of this master dissertation is to investigate and develop NER techniques tailored to academic and institutional documents, taking into account the presence of multiple languages and the diversity of entity types that may vary according to the structure, format, and writing style of different document types.

To accomplish the main goal, the specific goals are described below:

- Propose a methodology for constructing and validating corpora of multilingual academic documents;
- Explore state-of-the-art NER models based on language models and also LLMs;
- Investigate techniques for interactive visualization of entities in textual documents to support the visual analysis of NER results.

1.3 Contributions

The research has the following contributions in the fields of NER and visualization, aiming to address the identified gaps and advance the current state of the art:

- A Portuguese-English corpus presenting large documents and multiple entities related to the academic documents domain for nested NER;
- Baseline experiments using the corpus on state-of-the-art NER models;

- A novel visualization for named entities to support nested entities and long texts with multiple entities.

In the context of Brazilian educational institutions, this research aims to serve as a starting point for enhancing operational efficiency and advancing digital transformation initiatives. The proposed NER models enable the automatic identification of key entities, facilitating the extraction of structured data from unstructured texts. This can improve document indexing and search processes in EISs, reducing the time users spend on repetitive tasks and minimizing errors associated with manual data entry. Furthermore, the proposed approaches may benefit the broader NLP community by providing resources and baseline experiments focused on multilingual text processing (including Portuguese).

1.4 Organization

This dissertation is organized as follows. Chapter 2 presents the theoretical background, covering deep neural networks, recurrent neural networks, transformers, and named entity recognition (NER) techniques. Chapter 3 reviews the related work in the field. Chapter 4 describes the proposed methodology for Named Entity Recognition, including the construction of NER corpora based on SEI documents and the development of state-of-the-art models for both flat and nested NER. Chapter 5 outlines the requirements, tasks, and algorithms for named entity visualization. Chapter 6 presents the experimental results, including the assessment of corpus quality and the comparative performance analysis of NER models, supported by statistical evaluation. Finally, Chapter 7 provides concluding remarks, highlights the main contributions, and discusses directions for future work.

Chapter 2

Background

NER is a supervised NLP task whose goal is to identify entities that belong to predefined categories (Li et al. (2022a)), that may vary according to the domain and task objective. Entity recognition allows the transformation of unstructured natural language text into structured data that can be stored in a database. The main steps of a NER pipeline are described on Figure 2.1.

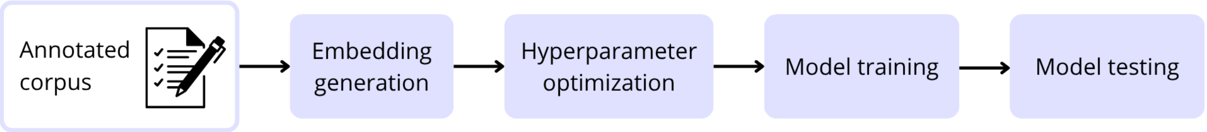


Figure (2.1) Steps comprising the traditional NER pipeline: an annotated NER corpus is split into training, validation, and testing. These raw texts are transformed in contextual embeddings prior to the hyperparameter optimization of a NER model. Finally, the best model is tested and evaluated.

As a supervised task, NER requires an annotated corpus, that is usually divided in training, validation and test data for model training and evaluation. An annotated corpus is a collection of text documents labeled with linguistic or semantic information, enabling its use in supervised NLP tasks. The annotation can be manual, when the process is performed by a trained person who knows the particularities of the data, and also by semi-automatic or automatic approaches, which are detailed in the next sections. Once the NER model has been trained on the annotated corpus, it can be used to identify entities in unknown texts that share similar characteristics with the training data. During this process, each token is either predicted as part of a predefined entity or assigned the "Other" (O) category, indicating that it does not belong to any of the specified entity labels recognized by the model.

When studying the NER task, it is essential to consider the difference between flat NER and nested NER (Wang et al. (2022b)). Flat NER, or simply NER, assigns a single category to each token, assuming that entities do not overlap and are not nested within other entities. Conversely, nested NER allows multiple categories for the same token, which is required when entities are embedded within others or have subcategories. The following examples illustrate the expected output of this task, which is to identify entities within a given text.

Figure 2.2 illustrates a sentence with Flat NER highlighted, and Figure 2.3 exemplify a sentence annotated with Nested NER. The first highlight Position, Article number and SEI's process ID entities, and the second illustrates that exists an Organization inside the Position entity.

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

Figure (2.2) Flat NER example.

$$\textit{The } \overbrace{\textit{Dean of Community Affairs of the } \underbrace{\textit{University of Brasília}}_{\text{Organization}}}^{\text{Position}}.$$

Figure (2.3) Nested NER example.

This Chapter presents the theoretical foundations of NER, introducing key concepts in NLP, a review of NER models, and visualization techniques that support data interpretation and knowledge discovery.

2.1 Corpus Annotation

Garside et al. (1997) defines corpus annotation as the process of adding interpretative and linguistic information to a raw corpus that can be created using spoken or written language data. The term "interpretative" means that the annotation is the result of human mind's understanding of the corpus (Garside et al. (1997)) while the term "linguistic" encompasses syntax, morphology, and semantic information.

The importance of annotating a corpus lies in its potential for automatic analysis and interpretation, particularly valuable in large-scale scenarios. Beyond information extraction, Garside et al. (1997) also argue that an annotated corpus is a reusable and multifunctional resource, because it can be utilized by other users and for purposes different from the original one. However, manual annotation requires human experts to label data manually, and this process is inherently expensive, time-consuming, labor-intensive, and prone to errors (Beck et al. (2020); Garside et al. (1997)).

Semi-automatic and automatic approaches are common alternatives to overcome these challenges, but trade-offs must be considered. Semi-automatic annotation consists of the combination of methods and/or tools with human supervision to reduce the time and effort required in data annotation while maintaining quality standards by human participation. Automatic annotations are performed by Machine Learning (ML) methods that annotate data without human intervention. While the automatic approach may be faster than the others, the resulting annotated data will not be reviewed and validated by a human expert.

Corpus annotation for NER is the process of designating labels for all tokens of each corpus text. A widely used tagging scheme for this task is the Inside-Outside-Beginning (IOB) format, which uses the following labels:

- **I** (Inside): identifies tokens that are part of a named entity but not the first one. The prefix “I-” is added to the entity type;
- **O** (Outside): identifies tokens that do not belong to any named entity and are labeled simply as “O”;
- **B** (Beginning): indicates the first token of a named entity. The prefix “B-” is added to the entity type (e.g., “B-PER” for the beginning of a person name);

According to Li et al. (2022a), a better understanding of language ambiguity requires a focus on the quality and consistency of data annotation, as errors in the annotation process can lead to model misclassifications, even when working with corpora from the same domain. In this sense, some strategies can be applied for quality control of the annotation process, which is particularly important when different specialists annotate the same corpus. Known by Inter-Annotator Agreements (IAA) metrics, the Cohen’s Kappa coefficient (Cohen (1960)) and Krippendorff’s Alpha (Krippendorff (1970, 2011)) may be employed to investigate the consistency between annotators.

Cohen’s Kappa is an IAA metric to calculate the level of agreement between categorical labels annotated by two annotators. This metric is not suitable for more than two annotators and in cases with missing annotations. The formula is presented in Equation (2.1).

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{2.1}$$

whereκ is the agreement coefficient between two annotators, P_o is the observed agreement, i.e., the proportion of items on which both annotators assigned the same category, and P_e is the expected agreement by chance, calculated based on the marginal probabilities of each category assigned by the annotators.

The value of κ is interpreted as follows:

$$\kappa = \begin{cases} 1, & \text{perfect agreement} \\ 0, & \text{agreement equivalent to chance} \\ < 0, & \text{less agreement than expected by chance} \end{cases}$$

Krippendorff's Alpha is also a commonly used metric for IIA that presents more flexible aspects than Cohen Kappa. It can handle incomplete data, support any number of annotators, and can also be applied in other scenarios beyond classification. The formula is presented in Equation (2.2).

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{2.2}$$

where:

α is the reliability coefficient between annotators, D_o is the observed disagreement, in other words, the weighted sum of differences between the labels assigned by annotators to the same items, and the expected disagreement by chance, which is calculated based on the marginal distribution of the assigned categories.

The value of α varies according to the level of agreement between annotators:

$$\alpha = \begin{cases} \text{perfect agreement,} & \text{if } \alpha = 1 \\ \text{agreement equivalent to chance,} & \text{if } \alpha = 0 \\ \text{less agreement than expected by chance,} & \text{if } \alpha < 0 \end{cases}$$

Regarding the annotation process, it is also relevant to consider that the creation of an annotated corpus is a costly process due to the high amount of time, resources, and human effort to perform annotation tasks with the quality and consistency needed (Li et al. (2022a)). Thus, combinations of prompt-based, weak supervision and human annotation can be investigated to accelerate corpus annotation and decrease the costs as shown by the research of Oliveira et al. (2024).

2.2 Deep Learning in Natural Language Processing

In NER tasks, the Deep Learning (DL) techniques are extensively explored. In this sense, this section presents the main concepts related to the theoretical base of DL models within the scope of NER. Special attention is given to the most common architectures in NER pipelines, such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and transformer-based models.

2.2.1 Deep Neural Networks

Neural networks were created with the aim of simulating the biological learning process that occurs in the nervous system (Hobson Lane and Hapke (2019)). The basic architecture of neural networks consists of input nodes, bias, activation functions, and output nodes. This section describes single-layer and multi-layer neural networks. The simplest neural network is the perceptron, a neural network that contains a single input layer and an output node, as described next.

Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ be a dataset, where each instance $\mathbf{x}_i \in \mathbb{R}^M$ is a vector described by M attributes, that is, $\mathbf{x}_i = [a_{i,1}, \ldots, a_{i,M}]$. Each instance is associated with a label y_i. The perceptron is a linear model that receives $\mathbf{x}_i$ as input and produces an output o_i computed as shown in Equation (2.3):

$$o_i = b + \sum_{j=1}^{M} a_{i,j} w_j, \tag{2.3}$$

where b is the bias term, and w_j is the weight associated with the j-th attribute.

Activation Functions

In the Neural Networks context, activation functions are mathematical functions that define the activation of a neuron, adding non-linearity to the model, as represented by Equation (2.4). In this sense, the importance of non-linearity in DL models lies in their ability to learn more complex patterns, which means that the non-linearity increases the model's learning potential (Dubey et al. (2022)).

$$\hat{y_i} = activation\_function(o_i), \tag{2.4}$$

where $\hat{y_i}$ is the predicted output of the model for the i-th input.

Some commonly used activation functions include the Rectified Linear Unit (ReLU), and the sigmoid function represented by Equations (2.5 and 2.6), respectively. While these functions differ in output range, the ReLU outputs only positive values, returning

0 for all negative inputs. The sigmoid function maps inputs to a range between 0 and 1, whereas the hyperbolic tangent maps inputs to a range between -1 and 1.

$$\mathrm{ReLU}(x) = \max(0, x). \tag{2.5}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{2.6}$$

Learning Process

The perceptron learning process, also known as training, consists of adjusting the parameters $w_1, \ldots, w_M$ and the bias b based on patterns identified in the dataset. Training a neural network involves minimizing an objective function E, also referred to as the loss function, as shown in Equation (2.7):

$$E = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{2.7}$$

Each data instance $\mathbf{x_i}$ fed the network to generate a prediction $\hat{y}_i$, and the weights are updated at each iteration to minimize the loss function, as defined in Equation (2.8):

$$w_j^{h+1} = w_j^h + \lambda a_{i,j}(\hat{y}_i - y_i). \tag{2.8}$$

In other words, the weights $\mathbf{w}$ are adjusted to improve the model's predictions by reducing the number of misclassifications. The parameter λ is the learning rate hyperparameter, which controls the magnitude of the weight updates. The variable h represents the time step or iteration. In a binary classification problem, the weights are updated only when the predicted output $\hat{y}_i$ is incorrect, which characterizes a misclassification.

The perceptron training is an iterative process that aims to minimize the error function E, where Equation (2.8) is repeatedly applied until an "acceptable" error level is reached. A complete pass through all training instances is called an epoch. Therefore, the minimization process may require multiple epochs, depending on the complexity of the task.

Multilayer Perceptron

Multilayer neural networks have been proposed to solve problems that cannot be addressed by a single-layer perceptron. These networks are composed of multiple perceptron-type neurons, with connections between neurons in adjacent layers. Figure 2.4 illustrates a multilayer feedforward architecture consisting of an input layer, one or more hidden layers, and an output layer. The number of hidden layers and the number of neurons in each layer are hyperparameters that must be tuned independently during training.

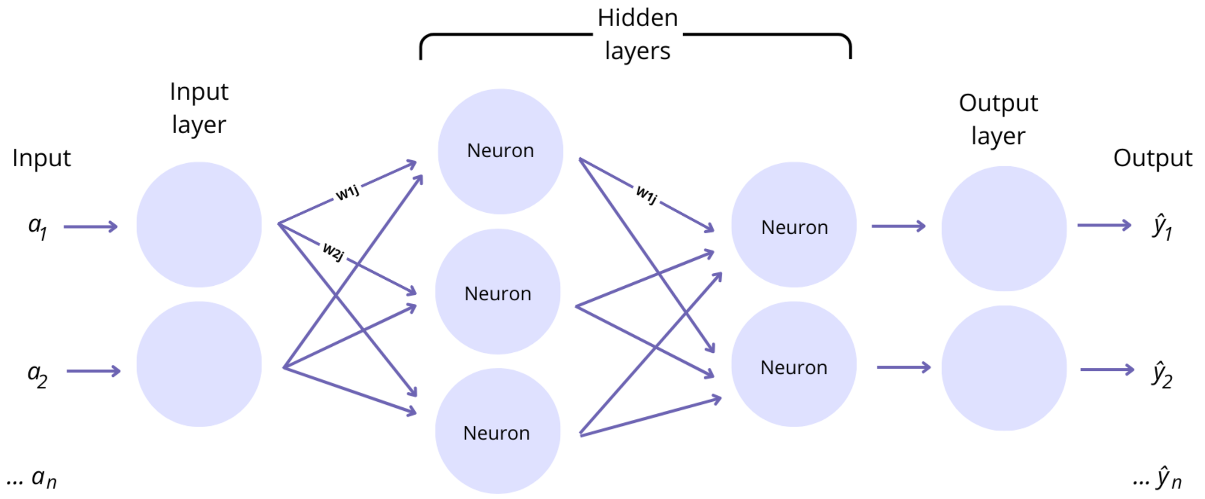


Figure (2.4) Feedforward Neural Network architecture.

Multi-layer neural network training uses the backpropagation algorithm, which is composed of two steps: propagation and backpropagation. In the propagation step, the input data is passed through the network, generating outputs to calculate a loss function. In the backpropagation step, the loss function gradient is calculated, which is used to adjust the network parameters (neuron weights) from the output layer to the input layer. To make gradient calculations feasible, it is common to use differentiable activation functions in hidden layers.

Recurrent Neural Network

A Recurrent Neural Network (RNN) is an architecture designed to handle sequential data, such as text, time series, and biological signals. In the context of Natural Language Processing, the sequential nature of the data is evident in the dependencies between words within a sentence. Furthermore, it is important to recognize that word order plays a key role in capturing semantic meaning (Aggarwal (2018)). The following pair of sentences illustrates this idea:

The cat eat the dog.
The dog eat the cat.

Notice that the same words are used in both sentences, but the meaning had changed significantly. The importance of the order of each word in a sentence highlights the need to design a neural network that process the inputs in the same order as they are in the original sentence and that treat all inputs at each position in a similar manner to previous inputs (Aggarwal (2018)).

The main difference between the RNN and the traditional Feedforward Neural Network (FNN) is the contribution to dealing with sequential data by allowing an input x_t to interact directly with hidden states created from inputs at previous time stamps $x_t, ..., x_{t-2}, x_{t-1}$ (Aggarwal (2018)). The term "recurrent" refers to the repetition of the same architecture over time. Therefore, traditional RNNs value memories of previous states according to time, which means that short-term memorization is favored.

Figures 2.4 and 2.5 illustrate the differences between FNN and RNN, which consists mainly by the addition of recurrent connections in neurons of hidden layers.

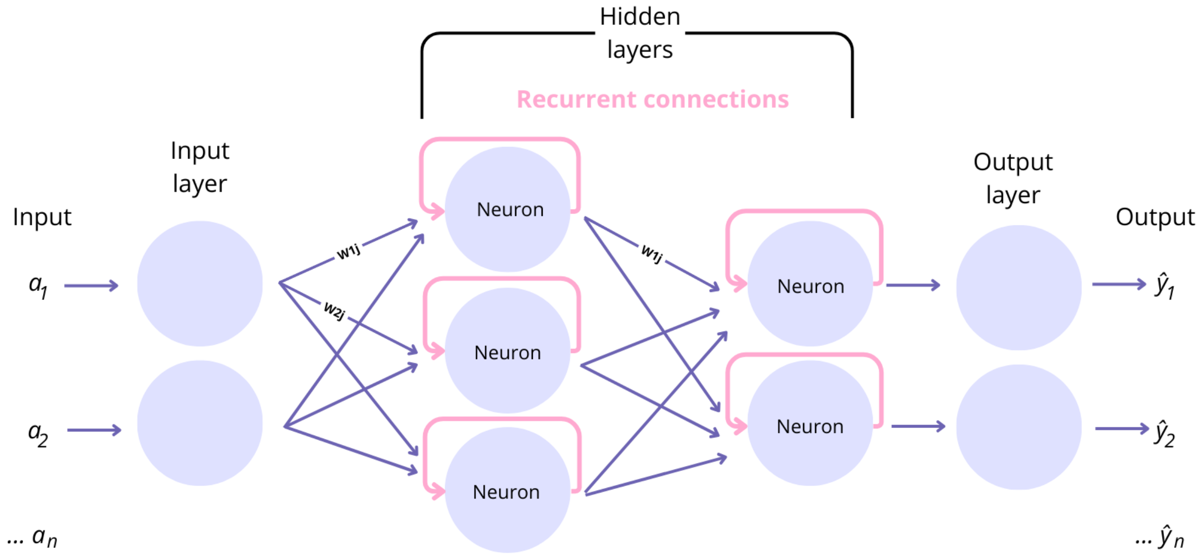


Figure (2.5) Recurrent Neural Network architecture.

Although RNNs have made significant contributions to the field of NLP, it was susceptible to vanishing or exploding gradient problems during training due to the use of backpropagation through time. The vanishing gradient problem occurs when gradients become progressively smaller as they are propagated back through time, while the exploding gradient problem occurs when error gradients accumulate and result in large updates to the network weights. This combination may result in unstable behavior in gradient-descent step size. As a result, RNNs are effective at capturing short-term dependencies but struggle with modeling long-term dependencies.

Several solutions have been proposed to address this problem, including strong regularization of parameters, gradient clipping, proper initialization strategies, batch normalization, and the integration of recurrent networks with internal memory mechanisms, such as the one introduced by the Long Short-Term Memory (LSTM) architecture, which is considered the most effective approach (Aggarwal (2018)).

Bidirectional Recurrent Network

RNNs consider only information from past inputs up to the current position in a sentence. However, in some cases, information about future tokens can also be useful. To address this limitation, Bidirectional Recurrent Neural Networks (BRNNs) were introduced, which is illustrated in Figure 2.6. Bidirectional networks separate the hidden states into forward $\overline{h}_t^{(f)}$ and backward $\overline{h}_t^{(b)}$ directions. Both hidden states receive the same input vector, but $\overline{h}_t^{(f)}$ compute states only in the forward direction and $\overline{h}_t^{(b)}$ only in the backward direction, independently, and then the generated parameters from both hidden states are computed. Consequently, a bidirectional layer enables context analysis in both left-to-right and right-to-left directions (Aggarwal (2018)).

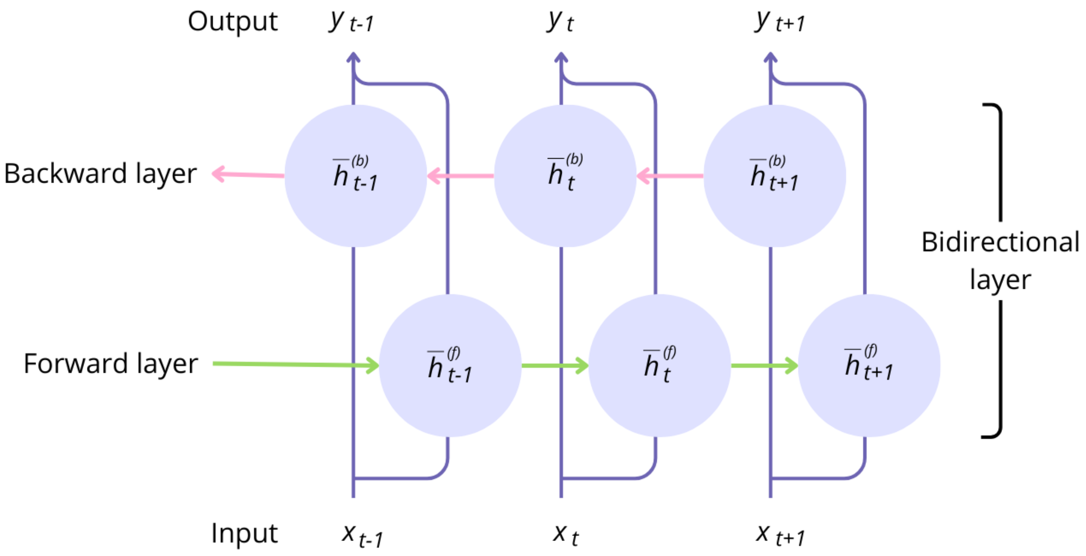


Figure (2.6) Bidirectional layer (adapted from Aggarwal (2018).

Long Short-Term Memory

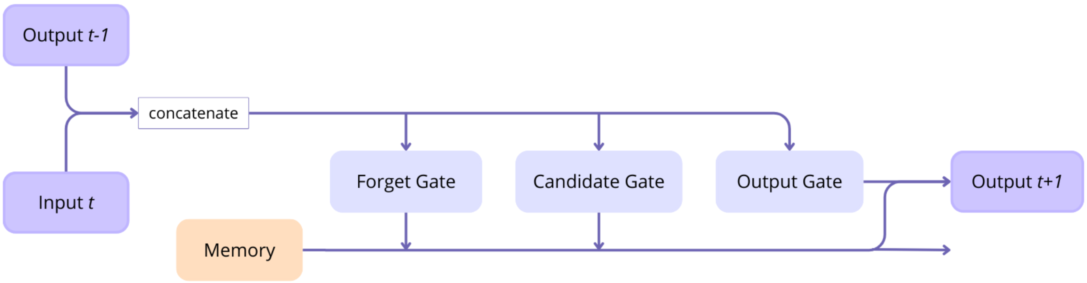


Figure (2.7) LSTM layer at time step t, adapted from Lane et al. (2019).

Long Short-Term Memory (LSTM) was proposed as a solution to the vanishing and exploding gradient problems by improving control over the information stored in long-term memory (Aggarwal (2018)). LSTM is a variant of the RNN architecture capable of balancing long-term memory retention, short-term memory influence, and generalization capacity (Hochreiter and Schmidhuber (1997)). These advancements offer benefits for tasks involving sequential data, making LSTMs widely used in the literature, particularly in text interpretation tasks. Bidirectional Long Short-Term Memory (BiLSTM) extends the LSTM architecture by adding a bidirectional layer, enabling context analysis in both forward and backward directions.

An LSTM architecture consists of three gates and a memory cell. These components are described below and illustrated in Figure 2.7.

- **Input Gate:** a feedforward network with a sigmoid activation function that decides which elements of the input are relevant to be stored in the memory cell. It works along with the candidate values (generated by a hyperbolic tangent activation) to update the cell state.
- **Memory Cell:** stores relevant information from previous time steps and affects the output in subsequent steps. This cell is represented by the "memory" rectangle in the Figure.
- **Forget Gate:** a feedforward network with a sigmoid activation function that learns which information should be discarded from the memory cell. The output is a vector of values between 0 and 1 that determines the extent to which each element of the memory is retained. Values close to 0 indicate forgetting, and values close to 1 indicate retention.
- **Candidate Values:** generated by a feedforward network with a hyperbolic tangent activation function. These candidate values represent new content to be potentially added to the memory cell, as filtered by the Input Gate.
- **Output Gate:** a feedforward network that receives the current input at time step t, the previous hidden state from time step $t-1$, and the updated memory at time t. It uses a sigmoid activation function to determine which parts of the memory are passed as the hidden state output at time t.

2.2.2 Convolutional Neural Network

The Convolutional Neural Network (CNN) terminology was introduced by LeCun et al. (1989). CNN is a feedforward neural network that uses convolution kernels to extract information. This architecture was primarily explored in the literature for image-processing tasks (Li et al. (2022b)), such as image classification and recognition, and object detection and localization.

CNN-based architecture commonly encompasses convolutional, pooling, along with a fully connected layer that generates the output. The main layer, which gave the network its name, consists of convolutional operations that filter the input data with a kernel to extract features on the output activation map, while the pooling layer reduces the spatial dimension of each activation map.

Despite the first CNN's objectives being extracting features from images, CNN has been also studied in the context of text processing. Similarly to images, that are represented as a 2-dimensional object with the depth defined by the number of color channels, a sentence is represented as a 1-dimensional object with depth defined by the representation dimensionality (Aggarwal (2018)).

2.2.3 Transformers

Recurrent models, such as LSTM, process the inputs sequentially, creating hidden states h_t in position t that depend on the previous ones h_{t-1}. This scenario does not allow parallelization and may be computationally inefficient for longer sentences. In this sense, the transformer architecture was proposed to overcome the challenges of processing long sentences with less computational effort (Vaswani et al. (2017)).

The transformer is an encoder-decoder architecture that implements the Scaled Dot-Product Attention function (Luong et al. (2015)) and the Multi-Head Attention technique. The proposed mechanisms determine which parts of an input sequence are most relevant to the output query while adding parallelism capabilities to increase computational efficiency (Bahdanau et al. (2014)).

The Scaled Dot-Product Attention is the implementation of the Self-Attention concept, which uses dot-product to determine weights for each input element, representing their relevance to the query, then normalizes and applies the softmax function to calculate the probabilities of each possible word being the result of the query.

Multi-Head Attention is a component responsible for adding parallelism to the attention layers. The output of the multiple attention mechanisms running in parallel is then concatenated into a dense vector representation, which accomplishes capturing and combining different aspects of the text. These aspects may include word relationships, contextual dependencies and patterns, and semantic meanings within the text.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language representation model introduced by Devlin et al. (2019), based on the encoder component of the Transformer architecture, which enables the model to capture the context of the processed text in both left-to-right and right-to-left directions (Hagiwara (2021)). In this sense, the attention mechanisms allows BERT to access information from older tokens along with current tokens (Hagiwara (2021)).

The training of BERT can be explained in two main steps: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled texts using the tasks of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The pre-training takes place by using a large-scale, general-purpose text corpus, providing the model with extensive linguistic representations. In the fine-tuning step, the model is adapted to specific supervised tasks with minimal additional parameters, reusing the same architecture in order to improve performance in specialized contexts (Hagiwara (2021)). Bidirectional representations are considered during the training process.

Another interesting aspect of BERT is the subword tokenization approach based on the WordPiece algorithm, which allows BERT to capture patterns related to rare or out-of-vocabulary words by splitting them into smaller units (subtokens). Prior to tokenization, BERT applies lowercasing (only for uncased models) and includes special tokens like `[CLS]` at the beginning of the sequence and `[SEP]` to delimit sentences or mark the end of input. The tag `[UNK]` is also used for words that the BERT tokenizer cannot match a word (or subword) to anything in its vocabulary. As a result, this produces word-level and character-level representations, reducing the vocabulary size while preserving semantic information. The Figure 2.8 illustrates the BERT subtokenization on a sentence extracted from SEI.

$$\underbrace{\quad}_{\text{[CLS]}}\ \underbrace{\textit{Boletim}}_{\text{[Bol, \#\#et, \#\#im]}}\ \underbrace{\textit{de}}_{\text{[de]}}\ \underbrace{\textit{Atos}}_{\text{[Ato, \#\#s]}}\ \underbrace{\textit{Oficiais}}_{\text{[Of, \#\#iciais]}}\ \underbrace{\textit{da}}_{\text{[da]}}\ \underbrace{\textit{UnB}}_{\text{[Un, \#\#B]}}\ \underbrace{\textit{em}}_{\text{[em]}}\ \underbrace{\textit{24/10/2018}}_{\text{[24,/,10,/,2018]}}\underbrace{.}_{\text{[.]}}\underbrace{\quad}_{\text{[SEP]}}$$

Figure (2.8) BERT tokenization example.

BERT can be pre-trained on a variety of corpora depending on the target language or domain. Devlin et al. (2019) introduced a mulitilingual BERT model named bert-base-multilingual-cased and Souza et al. (2020) a Portuguese BERT model named bert-base-portuguese-cased (BERTimbau). Moreover, several variants have been developed to address specific limitations or optimize its usage, such as the Robustly Optimized BERT Approach (RoBERTa) (Zhuang et al. (2021)) and DistilBERT (Sanh et al. (2019)).

2.2.4 Large Language Models

Large Language Models (LLMs) are transformer-based neural language models trained on large volumes of unlabeled text using tasks such as Next Token Prediction, Next Sentence

Prediction, or Masked Language Modeling (Minaee et al. (2024)). The term “LLM” was introduced to refer to large-scale PLMs, which may contain tens or even hundreds of billions of parameters (Zhao et al. (2024)). Unlike traditional PLMs, such as Google’s BERT and RoBERTa, LLMs do not require supervised fine-tuning for each specific task. In turn, an instruction prompt is typically sufficient to define the task (Minaee et al. (2024)).

The most known LLMs are GPT (Generative Pre-trained Transformers) (Brown et al. (2020); OpenAI et al. (2024)), LLaMa (Large Language Model Meta AI) (Touvron et al. (2023a)), Mistral (Jiang et al. (2023)), PaLM (Pathways Language Model) (Chowdhery et al. (2024)), Phi (Abdin et al. (2024)), and DeepSeek (DeepSeek-AI et al. (2025a)), each one with its particularities.

LLaMA 3.2 and DeepSeek R1 use tokenization methods based on Byte-Pair-Encoding (BPE) (Touvron et al. (2023b), Grattafiori et al. (2024), DeepSeek-AI et al. (2025b)), which defines tokens using subwords to reduce vocabulary size while preserving semantic information. The BPE algorithm merges the most frequent token pairs to define new tokens, for example, merging [“l”, “e”] to form “le” during the first iteration, and then merging [“le”, “t”] to “let” on the next. These newly formed tokens are added to the vocabulary. Each iteration is repeated on all subwords until the vocabulary reaches its predefined size. This method allows the model to represent rare words and misspellings without relying on the `[UNK]` token. Special tokens may also appear in the tokenization, such as `[PAD]` for padding.

The Figures 2.9 and 2.10 illustrate the LLaMA 3.2 and DeepSeek-R1-Distill-Qwen tokenizations on a sentence extracted from SEI. The main difference between the examples lies in how numerical tokens are handled: DeepSeek tokenizes numbers digit by digit, while LLaMA groups them into larger numerical units.

$$\underbrace{\quad}_{\text{<s>}}\ \underbrace{\textit{Boletim}}_{\text{[Bo, let, im]}}\ \underbrace{\textit{de}}_{\text{[Ġde]}}\ \underbrace{\textit{Atos}}_{\text{[ĠAt, os]}}\ \underbrace{\textit{Oficiais}}_{\text{[ĠOf, icia, is]}}\ \underbrace{\textit{da}}_{\text{[Ġda]}}\ \underbrace{\textit{UnB}}_{\text{[ĠUn, B]}}\ \underbrace{\textit{em}}_{\text{[Ġem]}}\ \underbrace{\textit{24/10/2018}}_{\text{[Ġ, 24, /, 10, /, 201, 8]}}\ \underbrace{.}_{\text{[.]}}\ \underbrace{\quad}_{\text{</s>}}$$

Figure (2.9) LLaMA 3.2-1B tokenization example.

$$\underbrace{\quad}_{\text{<s>}}\ \underbrace{\textit{Boletim}}_{\text{[Bo, let, im]}}\ \underbrace{\textit{de}}_{\text{[Ġde]}}\ \underbrace{\textit{Atos}}_{\text{[ĠAt, os]}}\ \underbrace{\textit{Oficiais}}_{\text{[ĠOf, icia, is]}}\ \underbrace{\textit{da}}_{\text{[Ġda]}}\ \underbrace{\textit{UnB}}_{\text{[ĠUn, B]}}\ \underbrace{\text{em}}_{\text{[Ġem]}}\ \underbrace{\textit{24/10/2018}}_{\text{[Ġ, 2, 4, /, 1, 0, /, 2, 0, 1, 8]}}\ \underbrace{.}_{\text{[.]}}\ \underbrace{\quad}_{\text{</s>}}$$

Figure (2.10) DeepSeek-R1-Distill-Qwen-1.5B tokenization example.

LLMs can be used in different ways depending on the target task and the available computational resources. The most popular and accessible option is prompting, in which a user provides a crafted input to the model to guide its response (Schulhoff et al. (2024); Zamfirescu-Pereira et al. (2023)). In zero-shot learning, the model is expected to perform a task solely based on the information in the prompt, even if it has never been explicitly trained for that task (Kojima et al. (2022)). Alternatively, LLMs can be integrated into downstream architectures, acting as backbones, in which their outputs (or internal representations) are used as features for additional components, such as a fully connected layer appended on top for classification or regression tasks. In this case, fine-tuning appears as a potential strategy since it allows the LLM to adapt its parameters to a specific domain, improving its performance in the specific context.

However, full fine-tuning of LLMs is computationally expensive and demands high memory usage, making it unfeasible for use on traditional hardware. To address this challenge, efficient training techniques have been developed, such as partial fine-tuning and the introduction of adapter layers (Zhang et al. (2023c)), which allow for the modification of only a small subset of model parameters. These methods are capable of reducing the usage of computational resources while attempting to preserve most of the model's performance, making fine-tuning more accessible to a wider range of users.

Among these techniques, Low-Rank Adaptation (LoRA) (Hu et al. (2021)) has become popular as an effective and lightweight parameter-efficient fine-tuning (PEFT) strategy. LoRA inserts trainable low-rank matrices into specific layers of the model, enabling subtle updates without altering the original weights. This approach reduces the number of trainable parameters and memory usage, making it feasible to fine-tune LLMs even on low resources scenarios, including limited GPU memory. These low-rank matrices are incorporated into the weights associated with the attention mechanism (such as the query and value) in transformer-based models. The literature has presented several variants of LoRA since then, such as QLoRA (Dettmers et al. (2023)), which combines quantization with LoRA for even greater memory efficiency, and AdaLoRA (Zhang et al. (2023b)), which dynamically adjusts the rank of the adaptation during training to balance performance and the use of resources.

The large number of parameters, in the order of billions, makes loading them into memory also unfeasible. To overcome this shortcoming, researchers and developers have employed quantization approaches in LLMs (Egashira et al. (2024)). The goal is to reduce the precision of model parameters, usually representing numbers from 32-bit floating point to a lower representations of bits, such as 8-bit or 4-bit integers. As a result, the quantization process decreases memory usage while minimally affecting model performance.

2.3 Conditional Random Fields

Conditional Random Fields (CRF) is a probabilistic undirected graphical model introduced by Lafferty et al. (2001) as a variant of a Markov Random Field, widely used in tasks such as keywords extraction, NER, sentiment analysis, classification, speech recognition, and part-of-speech tagging. In the literature, CRF have been explored to perform NER tasks as a standalone model and also as the final layer of DL architectures (Sutton and McCallum (2012), Luz de Araujo et al. (2018)).

The CRF learns weights in order to maximize the conditional probability of the correct label sequence given an input sequence. As a discriminative model, CRF's predictions rely on a conditional distribution $p(\mathbf{y}|\mathbf{x})$, a conditional probability of a sequence of labels $\mathbf{y}$ given the sequence $\mathbf{x}$. Regarding sequence processing, the recommended CRF model is the Linear-Chain Conditional Random Field, which is defined in Equation 2.9.

Given $X = (x_1, x_2, \ldots, x_n)$ as a sequence of words and their respective labels (IOB tags) $Y = (y_1, y_2, \ldots, y_n)$, and $\mathcal{F}$ as a set of feature functions, the linear-chain CRF's distribution $p(\mathbf{y}|\mathbf{x})$ is described as shown by Eq. (2.9):

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}, \tag{2.9}$$

where $f_k(y_{t-1}, y_t, X, t)$ is feature function k, which depends on the adjacent labels y_{t-1}, y_t, the sequence X and the position t. θ_k is the weight parameter learned for feature function f_k during training. $Z(\mathbf{x})$ is the normalization function, applied to ensure the distribution p sums to 1, as specified in Equation 2.10:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}. \tag{2.10}$$

The model is trained by maximizing the log-likelihood of the labeled training data, adjusting the parameters $\theta = \theta_k$ accordingly. Then, given an input sequence X, the most probable label sequence $\mathbf{y}^*$ is predicted by maximizing the conditional probability $p(Y \mid X)$:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) \tag{2.11}$$

This prediction step involves the Viterbi algorithm, that compute the most probable sequence and choose the $\mathbf{y}^*$ based on the higher probability found. For each position t, a score ranging from 0 to 1 quantifies how likely the transition from y_{t-1} to y_t is, based on the previously learned patterns.

2.4 Evaluation Strategies

Selecting appropriate evaluation metrics and analyzing the results from a statistical perspective are essential steps to conduct a robust and trustworthy analysis of the results. The evaluation metrics and statistical analysis of this research are described next.

2.4.1 Evaluation Metrics

The most common evaluation metrics for NER tasks are accuracy and F1-score. When choosing the ideal metric for each experiment, it is important to analyze which one will reflect reality. Even though accuracy is easier to calculate and more interpretable, it is not recommended if the data is unbalanced (Hagiwara (2021)) regarding the class labels. In this case, it is more appropriate to use the F1-score metric.

Equations 2.12 to 2.15 exemplifies accuracy, precision, recall and F1-score metrics respectively, considering a corpus with n entities.

$$Accuracy = \frac{\sum_{i=1}^{n} \frac{TP(i)+TN(i)}{,} TP(i) + TN(i) + FP(i) + FN(i)}{n} \tag{2.12}$$

$$Precision(i) = \frac{TP(i)}{TP(i) + FP(i)}, \tag{2.13}$$

$$Recall(i) = \frac{TP(i)}{TP(i) + FN(i)}, \tag{2.14}$$

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

$$f1\_score = \frac{\sum_{i=1}^{n} \frac{2*Precision(i)*Recall(i)}{Precision(i)+Recall(i)}}{n} \tag{2.15}$$

The formulation in Equation 2.15 corresponds to the Macro F1-score, which is the unweighted average of the F1-scores computed individually for each class. This metric consider all classes equally, regardless of their frequency in the corpus, being appropriate to evaluate performance on imbalanced data, as it happens in NER tasks.

2.4.2 Statistical Analysis

Statistical tests are fundamental strategies for comparing ML models and verifying if a specific model presents improved performance over others, especially when experiments are repeated multiple times on different datasets (Demšar (2006)). Their relevance lies in defining whether a result is statistically significantly better than others, rather than relying on traditional procedures, such as comparing directly their explicit F1-scores. To achieve this, a null hypothesis is established, stating that all algorithms perform equivalently, and any observed differences are merely due to random chance.

For comparisons involving more than two classifiers over multiple data sets, non-parametric tests have became a suitable choice due to common violations of assumptions like normality and sphericity in ML data (Demšar (2006)). Among these, the Friedman test alongside its corresponding post-hoc tests are capable to evaluate the chi-square distribution to determine whether there are significant differences among multiple models by ranking their performance on each dataset separately. If the null hypothesis is rejected by the Friedman test, indicating that there are significant differences, a post-hoc test, like the Nemenyi test, is applied to determine which specific pairs of classifiers differ significantly.

For example, considering $k = 3$ models, namely x_1, x_2, and x_3, the Nemenyi post-hoc test performs pairwise comparisons between all classification models. In this case, three pairwise comparisons are performed, and an adjusted p-value is computed for each of them:

- p-value$_{1,2}$ is calculated comparing x_1 to x_2
- p-value$_{1,3}$ is calculated comparing x_1 to x_3
- p-value$_{2,3}$ is calculated comparing x_2 to x_3

If the p-value$_{a,b}$ is greater than the significance level α (commonly set to 0.05), the difference between the models a and b is not considered statistically significant. Otherwise, if the p-value$_{a,b}$ is less than α, such difference is considered statistically significant.

The significance of model performances can also be determined by comparing the absolute difference between the average ranks of multiple classifiers to a computed critical difference (CD), instead of analyzing only on individual p-values. If the rank difference between any two models is equal to or greater than the CD, their performance is considered significantly different. The results can be visually summarized using a CD diagram, which plots the average ranks of the models and connect groups of algorithms that are not significantly different as illustrated by Figure 2.11. In this sample, classifiers connected by a horizontal line form a group of models whose performance differences are not statistically significant at the chosen significance level ($\alpha = 0.05$). The $CD = 2.728$ represents the minimum rank difference required for significance according to the Nemenyi post-hoc test.

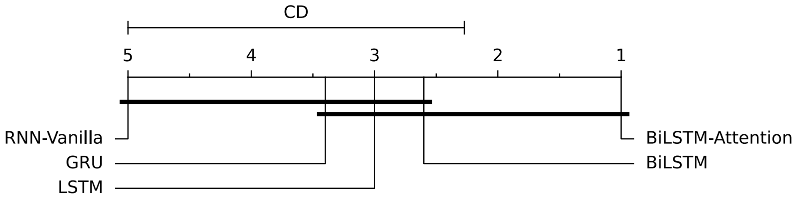


Figure (2.11) CD diagram comparing the average ranks of multiple classifiers over cross-validation folds.

In this research, the statistical analysis was carried out following the steps described below, using K-Fold Cross-Validation with the F1-score computed for each model in each fold:

1. For each of the M classifiers and each of the K cross-validation folds, compute the F1-score. Let $c_{j,i}$ denote the F1-score of the j-th classifier on the i-th fold.
2. For each of the K folds, rank the M classifiers based on their performance scores. The model with the best performance in a fold receives rank 1, the second in the rank is 2, and so on. Let $r_{j,i}$ be the rank of the j-th classifier in the i-th fold.
3. For each classifier, compute its average rank for all folds:

$$R_j = \frac{1}{K}\sum_{i=1}^{K} r_{ji},$$

   where R_j is the average rank of the j-th classifier.
4. Use the average ranks to calculate the Friedman statistic and test the null hypothesis that all classifiers perform equally. If the null hypothesis is rejected (i.e., there is evidence of significant differences among classifiers), proceed to the next step.
5. To identify which specific pairs of classifiers differ significantly, apply the Nemenyi test. It compares the absolute difference between the average ranks of each pair of classifiers against a critical difference (CD), defined as:

$$CD = q_\alpha\sqrt{\frac{M(M+1)}{6K}},$$

   where q_α is the critical value based on the Studentized range distribution for the chosen significance level α. If the absolute difference between any two average ranks is greater than or equal to CD, the difference is considered statistically significant.

2.5 Text Visualization

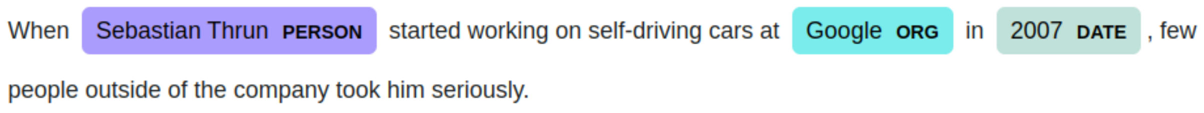
Text visualization techniques are designed to represent textual information graphically in order to support interpretation, exploration, and communication of insights. The main goal when using visualizations in NLP tasks are to help users and experts in identifying patterns, trends, and structures within large or complex text corpora (Liu et al. (2018)). Designing visualizations for raw texts presents challenges such as handling long texts, capturing meaningful patterns beyond word frequency, and avoiding clutter or misinterpretation due to the lack of semantic context (Alencar et al. (2012)).

Among the most well-known techniques for raw text visualization are word clouds, which emphasize word frequency by scaling font sizes. The word cloud algorithm first computes the frequency of each word and then places them within a limited layout using spiral positioning, ensuring that the most frequent words appear larger and are centrally located. However, word clouds have limitations, since word order, context, and semantic relationships are not taken into account, which can lead to misleading interpretations.

2.5.1 Visualization of Named Entities

Regarding named entities, visualizations are essential to analyze named entities in sentences. Small documents' entities can be visualized by colored tags, as illustrated in Figure 2.12. For longer documents, a graph-based visualization could be an alternative strategy to extract meaningful information.

Figure 2.12 illustrates an example of how named entities can be visualized in short sentences, highlighting detected tags such as Person (PERSON), Organization (ORG), and Date (DATE) using colors and tag identifiers. It is important to note that the recognized entities are limited to those predefined in the corpus, and this definition depends on the specific objectives of the task.



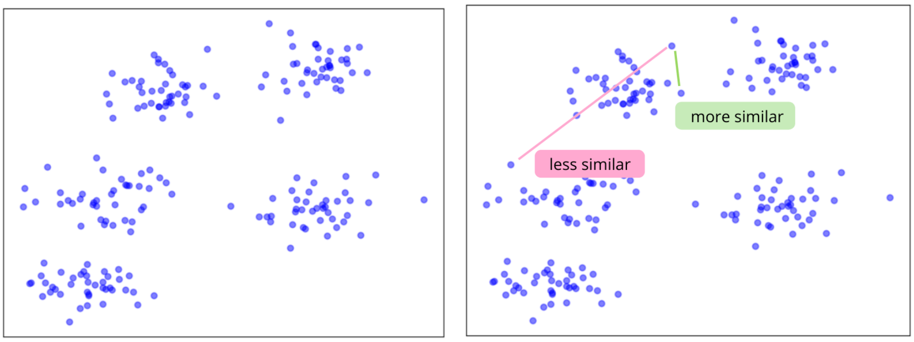
When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously.

Figure (2.12) Named entity visualization of short sentences using displacy.

The visualization techniques mentioned are often integrated with other approaches or systems to support users in performing NLP tasks.

2.5.2 Point placement based on t-SNE

Point-placement is a strategy for visualizing high-dimensional objects as points in a two-dimensional space, where the position of each point reflects the similarity between data instances (Paulovich and Minghim (2008)). In the case of texts, each sample can be represented by a contextual embedding, which typically has high dimensionality. This requires the use of dimensionality reduction techniques to enable visualization in a two-dimensional graphical representation (also called as layout). In this representation, the closer two points are, the more similar the corresponding data instances are. On the other hand, greater distances indicate greater dissimilarity. Figures 2.13a and 2.13b illustrate a point-placement visualization of generated sample data. Each blue point corresponds to a data instance, and clusters of points represent similar instances, as indicated by the low distances between points within each group.



(a) Point-placement visualization of sample data.

(b) Relation between distance and similarity in point-placement visualization.

Figure (2.13) Point-placement visualization.

Among the various techniques available for point-placement visualization (Nonato and Aupetit (2018)), t-Distributed Stochastic Neighbor Embedding (t-SNE) has been widely adopted for text representation tasks (Van der Maaten and Hinton (2008)) that requires semantic-level analysis. t-SNE is a non-linear dimensionality reduction method that maps high-dimensional data to a lower-dimensional space using the Student's t-distribution to compute similarities between points. It aims to preserve both local and global structures in the data—meaning that it maintains the relative distances between nearby data points as well as broader topological relationships.

Algorithm 1 details the steps of t-SNE, in which cosine similarity is typically used to measure the similarity between pairs of embedding vectors. The normal distribution is commonly adopted to determine the conditional distribution of data samples in the high dimensional space.

The possibility of identifying similar instances and clusters using point-placement and t-SNE has motivated its use in NLP tasks such as topic modeling (Atzberger et al. (2023)) and document clustering (Sherkat et al. (2019)). Moreover, it can support annotation tasks through visual inspection (Benato et al. (2021)), assist in the interpretation of

Algorithm 1: t-SNE: t-Distributed Stochastic Neighbor Embedding.

```
Compute neighborhood similarities using the desired similarity metric;
Randomly initialize the low-dimensional embedding;
for $e$ in elements do
    Compute conditional neighborhood similarities in the low-dimensional embedding;
    Compute the gradient of the Kullback-Leibler divergence between the original and conditional neighborhood similarities;
    Update the low-dimensional embedding using gradient descent;
    Apply dimensionality reduction using the t-Student measure;
```

classification models (Zeiler and Fergus (2013)), and enable human involvement in the learning process (Coscia and Endert (2024)).

2.6 Final Considerations

This chapter introduced the foundational concepts that support the development of this research. First, the classical NER pipeline was described, which is important for formulating the proposed methodology and for selecting appropriate techniques for recognizing entities in multilingual academic and institutional documents. The fundamentals of DL were also discussed, in which neural network architectures were discussed such as recurrent and transformer-based models that have proven effective in various NLP tasks. Furthermore, the chapter explored methods for visualizing textual data, including those tailored for named entities and point placement based on t-SNE, allowing users to gain insights over the texts patterns and the outputs of NER models.

Chapter 3

Related Works

This chapter explores research that address state-of-the-art methods for NER focused on multilingual texts. It discusses the availability and use of multilingual corpora, techniques for automatic corpus annotation and construction, and novel multilingual approaches devised to improve the performance of LLMs. Furthermore, the review emphasizes strategies to address challenges related to code-mixed language scenarios, in which texts can present multiple languages, as well as those tailored to Portuguese NER.

3.1 The evolution of state-of-the-art methods for NER

This research follows the categorization of NER models proposed by Keraghel et al. (2024), which divide the models into 5 distinct periods: Knowledge-based methods, Feature Engineering-based methods, Deep Learning-based methods, Transformer-based language methods, LLM-based methods. This section will introduce each period with references of the most popular NER models and approaches.

3.1.1 Knowledge-based methods

The identification of named entities using knowledge-based methods relies on predefined linguistic rules and lexical resources, such as patterns in capitalizing letters or particular terms and prefixes (Keraghel et al. (2024)). Regular expressions and dictionaries are often used to verify the predefined patterns and consequently identify the named entities. For example, if a text spam matches the pattern $< month >< day >< year >$, according to the following regular expression $regex = \backslash b(\text{January}|\ldots|\text{December})\backslash s + \backslash d\{1, 2\}, \backslash s + \backslash d\{4\}\backslash b$, It could be labeled as a “DATE” entity. Similarly, if there is a name close to the terms “Inc.” or “Ltda.” there is a chance of being an “ORGANIZATION” name.

3.1.2 Feature Engineering-based methods

Baum et al. (1970) proposed a Hidden Markov Model (HMM), a probabilistic method used for sequence labeling tasks such as NER that considers the previous state to classify the entity of the current state. Then, Lafferty et al. (2001) proposed the CRF, a model that has the advantage of considering the context to make predictions of named entities. Sutton and McCallum (2012) also suggest using CRF as the final layer of the neural network to accomplish an inference in classification-related tasks. This method still presented satisfactory results in NER tasks nowadays, as proved by M. C. Guimarães et al. (2024).

3.1.3 Deep Learning-based methods

The rise of neural networks has allowed the creation of models with the potential to capture more complex patterns in textual data. This period has remarkable proposals, such as convolutional (LeCun et al. (1989)), recurrent (Bengio et al. (1994)), and bidirectional neural network approaches. Additionally, the LSTM network (Hochreiter and Schmidhuber (1997)), a variation of RNN, was introduced by incorporating a memory mechanism to handle sequential data.

CNNs, initially proposed to extract patterns in images, were also explored as an alternative method to improve computational efficiency and performance in English and Chinese languages, as shown by Shen et al. (2017) with the CNN-CNN-LSTM architecture, and by Gui et al. (2019) with the Lexicon Rethinking CNN (LR-CNN) architecture. The literature also shows the use of BiLSTM+CRF (Ju et al. (2018a)) and LSTM+CNN (Parsaeimehr et al. (2023)) architectures in NER tasks.

3.1.4 Transformer-based language models

The development of the transformer architecture (Vaswani et al. (2017)) represents an important breakthrough in the NLP and leads to the beginning of attention-based approaches. The following Small Language Models were proposed in this period: multilingual BERT (mBERT) (Devlin et al. (2018)), RoBERTa (Liu et al. (2019)), XLM-RoBERTa + CRF (Conneau et al. (2020)), and multilingual BART (mBART) (Liu et al. (2020)).

3.1.5 Large Language Model-based methods

With the advance of the language models, the PLMs trained with tens or hundreds of billions of parameters are called Large Language Models (Zhao et al. (2024)). Recently, LLM have been a widely explored and discussed topic in NLP due to its impressive performance in various tasks, including text summarization, machine translation, and question & answering. The publication of ChatGPT 3.5 and 4, based on Open AI's GPT (Generative Pre-trained Transformers) (Brown et al. (2020); OpenAI (2023)) marks the transition of research focus from Pre-trained Language Model to LLM approaches. Then, other LLMs were proposed LLaMa (Large Language Model Meta AI) (Touvron et al. (2023a)), Mistral (Jiang et al. (2023)), PaLM (Pathways Language Model) (Chowdhery et al. (2024)), Phi (Abdin et al. (2024)), and DeepSeek (DeepSeek-AI et al. (2025a)).

3.2 Multilingual NER corpora

Multilingual corpora were searched in the literature to identify suitable corpora for our research. However, the majority of the identified multilingual corpora that included Portuguese were not code-mixed, which is an important characteristic in this research. Popular open corpora for multilingual NER include CoNLL-2002 and 2003 (Tjong Kim Sang (2002); Tjong Kim Sang and De Meulder (2003)), and MultiCoNER I (Malmasi et al. (2022)). Some multilingual corpora that includes Portuguese have also been found in the literature, such as WikiAnn (Pan et al. (2017)), MultiNERD corpora (Tedeschi and Navigli (2022)), MultiCoNER II (Fetahu et al. (2023a)), EMBER (Fetahu et al. (2021)), and WikiNER (Nothman et al. (2013)), which are widely explored corpora in the multilingual scenario.

CoNLL-2002 corpus (Tjong Kim Sang (2002)) includes news data in Spanish and Dutch, annotated in a variant of the IOB tagging scheme with the following entity categories: person's names, organizations, locations and miscellaneous entities that do not belong to the previous groups. Only the top level entity was annotated, resulting in non-overlapping entities, which characterizes it as a Flat NER corpus. Similarly, CoNLL-2003 corpus (Tjong Kim Sang and De Meulder (2003)) contributes by providing annotated corpora for English and German languages. A sample sentence from CoNLL-2003 is illustrated in Figure 3.1.

$$\underbrace{\textit{U.N.}}_{\text{ORGANIZATION}} \quad \textit{official} \quad \underbrace{\textit{Ekeus}}_{\text{PERSON}} \quad \textit{heads for} \quad \underbrace{\textit{Baghdad}}_{\text{LOCATION}}.$$

Figure (3.1) CoNLL-2003 example (English).

Malmasi et al. (2022) presented the MultiCoNER II corpora in the SemEval 2023 Task 2, a challenge of Multilingual Complex Flat NER, that is compound of 11 languages: Bangla, Chinese, English, Spanish, Farsi, French, German, Hindi, Italian, Portuguese, Swedish, Ukrainian. It includes individual corpora for each language, and also multilingual corpus in which each sentence belongs to a different language. The annotation scheme was follows the CoNLL format and the annotated categories are location, creative work, group, person's name, product and medical. Sample sentences from MultiNERD are illustrated in Figure 3.2.

O plantio de $\overbrace{\textit{café}}^{\text{DRINK}}$ *logo despontou como a cultura principal e se tornou fator de desenvolvimento.*

Translated: *Coffee plantations soon emerged as the main crop and became a factor in development.*

(a) MultiCoNER II example (Portuguese).

He inspired medical student $\underbrace{\textit{alexander rich}}_{\text{PERSON}}$ *to pursue an academic career.*

(b) MultiCoNER II example (English).

Figure (3.2) Sample sentences from the MultiCoNER II.

Similarly, the MultiNERD is also a multilingual corpus, but has a separate set of documents for each language and focus on Wikipedia and WikiNews articles. MultiNERD is composed of 10 languages: Chinese, Dutch, English, French, German, Italian, Polish, Portuguese, Russian and Spanish; and 15 entities were considered: person's name, location, organization, animal, biological entity, celestial body, disease, event, food, instrument, media, plant, mythological entity, time and vehicle. NER and Entity Disambiguation annotations have been generated automatically by Babelscape[1]. A sample sentence from MultiNERD is illustrated in Figure 3.3.

Dentre as naturais, destaca-se a $\underbrace{\textit{mamona}}_{\text{PLANT}}$*.*

Translated: *Among the natural ones, castor bean stands out.*

Figure (3.3) MultiNERD example (Portuguese).

The EMBER corpus (Fetahu et al. (2021)) is distinguished by its code-mixed sentences, in which each sentence combines two languages from the following set: English, Dutch, Russian, Turkish, Farsi, and Korean. The sentences of EMBER corpus were obtained by Bing search queries, which results in a corpus constructed with short sentences. This

[1] https://babelscape.com/

corpus has specific characteristics, such as small and limited context, non-well-formed sentences, and ambiguous entities. Moreover, since the corpus does not include the Portuguese language, it was not suitable for use in this research. The authors also proposed the mLOWNER, a low context multilingual NER corpus. A sample sentence from EMBER is illustrated in Figure 3.4.

What is $\underbrace{\textit{same girl (jennifer lopez şarkısı).}}_{\text{CREATIVE WORK}}$

Translated: *What is same girl (jennifer lopez song).*

Figure (3.4) EMBER example (English-Turkish).

Nothman et al. (2013) created the WikiNER, a multilingual corpus with silver-standard annotations for NER with texts extracted from Wikipedia website. The corpus contains 7200 Wikipedia articles across 9 languages: English, German, French, Polish, Italian, Spanish, Dutch, Portuguese and Russian. There are code-mixed texts, that contain Portuguese and English words in the same sentence. The annotation process involved an initial named entity manual labeling, followed by inferences based on internal text links. For corpus evaluation, 149 articles were annotated by 3 annotators, achieving a Fleiss' Kappa of 0.83 on named entities tokens. A cross-lingual approach was employed and achieved up to 95% of accuracy. A sample sentence from WikiNER is illustrated in Figure 3.5.

Os latinos em geral têm acrescentado à cultura local suas artes, artesanato, e criações diversas que foram se integrando, e hoje são copiadas e utilizadas pelo $\underbrace{\textit{Brasil}}_{\text{LOCATION}}$ *afora como "hand made in* $\underbrace{\textit{Brazil}}_{\text{LOCATION}}$ *".*

Translated: *Latinos in general have added their arts, crafts and various creations to the local culture, which have been integrated and today are copied and used throughout Brazil as "hand made in Brazil".*

Figure (3.5) WikiNER example (Portuguese-English).

After searching in the literature by annotated and open corpora for multilingual and code-mixed NER, the multilingual corpora MultiCoNER II, MultiNERD were found, along with the multilingual code-mixed corpus EMBER and the WikiNER corpus that has some code-mixed sentences. However, no publicly available annotated corpus for multilingual code-mixed data including Brazilian Portuguese was found.

3.3 Corpus annotation for NER

As NER is a supervised task, it is required to build labeled corpora to enable training, fine-tuning, and evaluation of NLP systems and models. The following research was selected based on similar approaches for curating corpora and describing annotation processes. Corpora in other languages were also explored, particularly if they belonged to domains comparable to academic texts, for example, legal documents, which often contain entities with similar characteristics to the academic ones.

The HAREM corpus (Santos et al. (2006)) was proposed for an advanced NER evaluation contest for Portuguese that included 1202 documents from Brazil, Portugal, Asia and Africa. The corpus has a golden collection and also documents tagged by systems. A HAREM' sub-collection of 129 documents has a golden standard annotation, with 5132 manually annotated entities from web, newspaper, e-mail, oral, expository, fiction, technical and political text genres. In the golden collection, 10 categories were considered: person's name, organization, time, location, product/work, event, abstraction, object, value and other.

Luz de Araujo et al. (2018) created the LeNER-Br, a Portuguese corpus 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%.

Albuquerque et al. (2022) constructed an annotated corpus for NER on Brazilian legislative documents motivated by the lack of official texts from the Chamber of Deputies written. The annotation process involved trained annotators divided into three groups, following predefined entity categories (such as dates, legal foundations, and organizations), and was conducted using the INCEpTION tool to support collaborative tagging. The quality of the annotations was assessed using inter-annotator agreement metrics, with Cohen's Kappa scores yielding to 91%, 94%, and 88%, one for each annotation team. State-of-the-art NER techniques such as CRF, HMM, and BiLSTM-CRF with GloVe embeddings were considered to evaluate and compare their performances. The experiments showed that CRF model achieved the best performance, with an F1-score of approximately 81%.

Silva et al. (2023) created a corpus named CachacaNER that contains over $180,000$ tokens labeled in 17 entity categories in the domain of a popular Brazilian beverage named "Cachaça". The categories include person's name, organization, location, time, value, abstraction, event, thing, title, and other. The authors conducted an annotators agreement experiment, which resulted in the Kappa metric of 0.857, that defines a high level of agreement among the annotators. Furthermore, in the experimental evaluation, we obtained a micro-F1 value equal to 0.933. The corpus is publicly available at Github[2].

Aiming to improve the NER system for the Romanian legal domain, Vasile Păiș and Onuț (2023) created a manually annotated corpus making an annotation process in which each annotator was responsible for 100 Romanian documents, considering the following categories: 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 F1-score 84%.

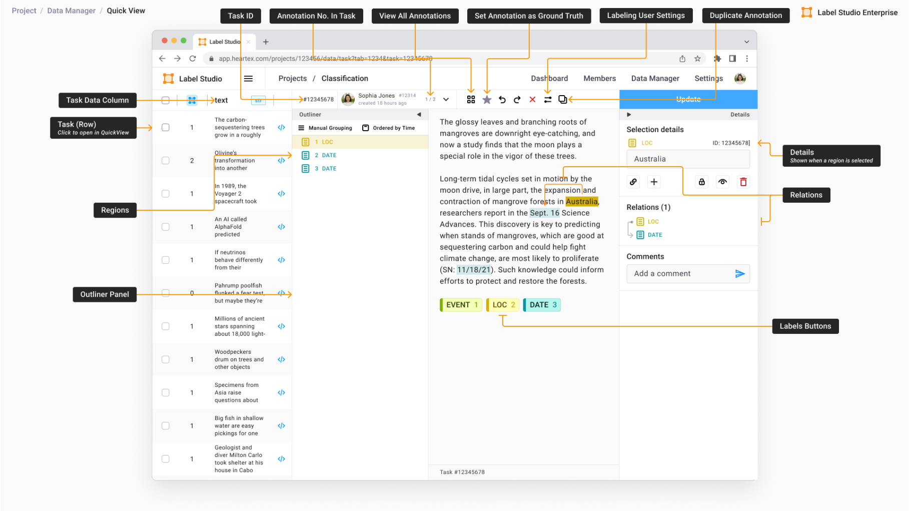
Guimarães et al. (2024) designed a tool to classify acts and perform NER on the official gazette of the Federal District of Brazil. It combines rule-based text classification with Machine Learning for NER, comparing three models: CRF, CNN-CNN-LSTM, and CNN-BiLSTM-CRF. This Python CLI tool enhances public 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.

Tkachenko et al. (2025) proposed LabelStudio, an open tool for corpus annotation, with code available on Github[3]. Allow personalization of NER categories adapted to the use case domain, multiple annotation types, including classification and NER, multiple data types like images, audio, text, time series, multi domain and video; multiple annotator profiles to annotate and review the texts, registering the creation and update datetime with version control. The tool permits the annotation of overlapping entities, which is applicable in Nested NER scenarios. Furthermore, LabelStudio also has integrations with Machine Learning and inter-annotator agreement capabilities as part of its paid subscription plans. The project and labeling interfaces are shown in Figures 3.6 and 3.7.

The creation of an annotated corpus for NER involves significant challenges, including high costs, time, and relevant human effort, and expertise in some domain-specific cases (Jehangir et al. (2023)). In this context, some research address the challenge of creating corpora for NER with reduced costs associated with corpus annotation (Nothman et al. (2013); Tedeschi et al. (2021)). LLM-based methods are also being explored in zero-shot (Ma et al. (2022); Sainz et al. (2024)), one-shot and few-shot (Agrawal et al. (2022); Zhang et al. (2023a)) scenarios, which are tasks with prompts consisting of task instructions alongside with zero, one or a few examples, respectively.

[2] https://github.com/LabRI-Information-Retrieval-Lab/CachacaNER

[3] https://github.com/HumanSignal/label-studio



The glossy leaves and branching roots of mangroves are downright eye-catching, and now a study finds that the moon plays a special role in the vigor of these trees.

Long-term tidal cycles set in motion by the moon drive, in large part, the expansion and contraction of mangrove forests in Australia, researchers report in the Sept. 16 Science Advances. This discovery is key to predicting when stands of mangroves, which are good at sequestering carbon and could help fight climate change, are most likely to proliferate (SN: 11/18/21). Such knowledge could inform efforts to protect and restore the forests.

Figure (3.6) LabelStudio's project page (from LabelStudio's official documentation).

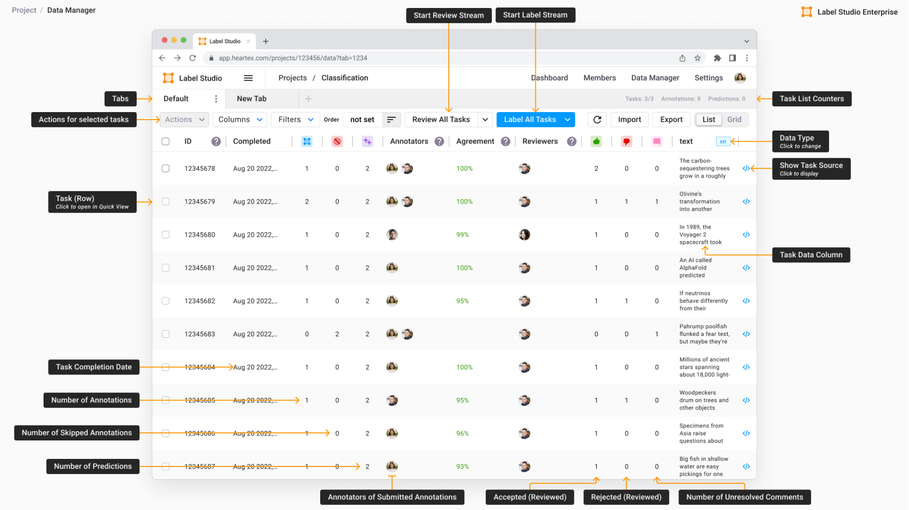


Figure (3.7) LabelStudio's labeling interface (from LabelStudio's official documentation).

3.4 Multilingual NER

Some research in the literature has explored innovative alternatives to improve the results of multilingual NER, including cross-lingual transfer, data augmentation, and multimodal approaches. Cross-lingual transfer is the concept of using other languages' knowledge to enhance the performance of a particular language. The data augmentation technique explores the expansion of the corpus based on existing data. Lastly, multimodal approaches explore the combination and interaction of various forms of information.

Coria et al. (2022) conducted cross-lingual transfer experiments on MultiCONER I (Malmasi et al. (2022)) with the multilingual BERT model (Devlin et al. (2018)), which was trained on Wikipedia data from 104 different languages using masked language modeling and next-sentence prediction tasks. The model was followed by a two-layer feedforward classifier with hidden size 768 and the ReLU activation function. The architecture also includes a dropout layer with $p = 0.1$, and the Adam optimization algorithm with a learning rate of 5×10^{-5}. The authors concluded that, despite some forgotten information, the forward transfer between languages persists.

The SemEval 2023 challenge instigates solutions for multilingual complex NER using the MultiCoNER II corpora (Fetahu et al. (2023b)), which includes the Portuguese monolingual track. The competitors Lovon-Melgarejo et al. (2023) used the XLM-RoBERTa (xlm-roberta-large) model, a pre-trained model, to compute contextualized representations of tokens. Each entity was expanded by concatenating a category descriptor, which adds more information about the category, including its definition and the fine-grained taxonomy's hypernymy. The macro-averaged F1-score obtained with the proposed method increased in comparison to the baseline XLM-RoBERTa, indicating that the addition of context-relevant information about the NER category can improve the token representations of a pre-trained model.

Multimodal approaches are also being studied in multilingual scenarios. Wang et al. (2024) combined multilingual text and image representations to create a multimodal contrastive learning method for NER. The framework used the multilingual BERT encoder to create text representations, ViT and ResNet encoders to create image representations based on patch and convolution features, an attention-based mechanism to allow interaction between text and image representations, and a CRF layer for category prediction.

3.5 Nested NER

Nested NER involves detecting entities embedded within other entities, reflecting a hierarchical structure and enabling richer, more fine-grained semantic representations. For instance, the name of a state might appear within an organization entity, which may include a person's name, resulting in multiple levels of nesting. The first approaches for nested NER comprised rule-based systems and traditional machine learning models such as Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), and CRF (Wang et al. (2022c)). These methods typically handled nested entities by first identifying inner entities and then applying post-processing rules to detect outer entities.

One of the first challenges in this field was the representation of nested entities, since traditional IOB assumes a flat, non-overlapping structure, which presents limitations for capturing nesting. To address this, Wang et al. (2022c) discussed alternative representations capable of modeling the hierarchical and overlapping nature of nested entities. Region-based tagging represents entities by labeling spans of text directly, allowing multiple overlapping entities to be captured without relying on token-level tags. Hypergraph-based representations model entities and their relationships using hyperedges, enabling the structured encoding of nested and overlapping entities. Layered tagging introduces multiple layers of flat annotations over the same text, with each layer capturing a distinct level of nesting. Finally, span-based representations treat entity recognition as a span classification task, where spans of text are scored and labeled independently.

Recent research in nested NER has been motivated by the gradual increasing availability of proper corpora. ACE 2005 (and ACE 2004) (Walker et al. (2006)) cover English newswire and conversational data with multi-level nested entities. GENIA[4] is a biomedical corpus of scientific abstracts with richly annotated nested structures, such as "human T cell leukemia virus type I tax gene", which includes nested entities like "human T cell", "T cell leukemia virus", and "tax gene". The GermEval 2014 corpus (Benikova et al. (2014)) is a German-language NER resource based on Wikipedia and news citations, containing over $31,000$ sentences and $590,000$ tokens, and comprising four main entity categories and subcategories. However, there is a lack for multilingual corpora for nested NER, even considering Portuguese among them.

Ju et al. (2018b) proposed a novel transition-based approach to address nested NER using a shift-reduce parsing framework. In their proposal, a neural model incrementally constructs labeled and nested structures by predicting actions (e.g., shift, reduce) regarding the words in order to deal with the hierarchical entities. This architecture implements stack-LSTMs and character-level embeddings, and it is not based on span enumeration. Experiments on the GENIA corpus show that the proposed model achieves a F1-score of 74.7% against 73.0% and 72.2% of FOFE-based model and hypergraph-based model, respectively. Regarding ACE 2005, their model reached 72.2% compared to 62.5% and 63.1%, respectively.

Zhou et al. (2022) introduced PANNER, a part-of-speech (POS)-aware model based on a heterogeneous graph neural network. The model constructs a multi-layered heterogeneous graph incorporating word, span, and POS nodes to capture both syntactic and semantic relationships. This architecture enables joint reasoning over entity characteristics and their linguistic contexts through message passing across heterogeneous node types. Experimental results show that PANNER achieved an F1-score of 75.4% on the GENIA corpus and 74.8% on ACE 2005, outperforming previous approaches in the literature at the time.

Wan et al. (2022) proposed a span-based approach for nested NER consisting of a conflict resolution mechanism to handle overlapping entities effectively. The proposed approach works by generating candidate spans and providing independent scores, and a novel conflict resolution strategy is applied to select a consistent set of nested entities. The

[4] `https://www.geniaproject.org/home`

architecture leverages pretrained language models, making it flexible and easily adaptable to different contexts. The authors performed experiments to demonstrate the superiority of their model, which achieves an F1-score of their BERT-based model obtained an F1-score of 79.30% on GENIA, which is competitive than additionally supervised BERT-MRC (83.75%) (Li et al. (2020)) and NER-DP (Yu et al. (2020)) (80.50%) approaches. Moreover, considering ACE 2005, the proposed model yielded 85.11% against 86.59% of BERT-MRC and 86.70% of NER-DP.

Sui et al. (2022) developed a novel graph neural network (GNN) architecture which incorporates “triggers” to guide the recognition of nested entities. Entity triggers are defined as a set of words from a sentence that provide complementary annotation for recognizing entities. The proposed Trigger-GNN model constructs a directed graph where each word in the sentence is a node, and edges are added based on lexicon. It also designs a graph-level node to gather information from nodes and edges. This architecture enables multiple graph-based interactions among words, entity triggers, and the whole sentence semantics. Experimental results demonstrate that Trigger-GNN achieves competitive performance, by outperforming other strategies in literature, obtaining F1-scores 86.53% on JNLPBA, and 93.45% on BC5CDR.

3.6 Large Language Models in Multilingual Scenario

Despite the mentioned advances, recent research defend that LLMs, with emphasis on GPT, have to be improved to solve some challenging tasks with reliability (Laskar et al. (2023)), especially the ones that require more complex reasoning abilities like Named Entity Recognition (Lai et al. (2023)).

Lai et al. (2023) highlight that GPT’s performance in zero-shot learning is generally worse than state-of-the-art supervised learning methods. Xie et al. (2023) also proved that GPT’s zero-shot performance had results below BERT-based models on NER task. Furthermore, Zhang et al. (2023d) conducted some experiments with GPT and concluded that GPT was not designed to take advantage of the multilingual corpus due to its inability to create a single multilingual conceptual representation analogous to compound multilingualism.

A recent proposal by Zaratiana et al. (2023) involves the creation of a generative model named GLiNER, based on Bidirectional Language Models such as BERT. Experiments were performed on MultiCoNER I corpora and they demonstrate GLiNER’s ability to handle languages that were not even encountered during training. The GLiNER multilingual version (mdeberta-v3-base) outperforms zero-shot ChatGPT in the following Latin languages: German, Spanish, and Dutch.

Another research that addresses zero-shot NER is the UniversalNER proposed by Zhou et al. (2024). The proposal is to explore targeted distillation in order to reduce ChatGPT into small models, called UniversalNER models, that have the same ability to recognize entity types, better F1-score results, and less complexity and parameters.

As an effort to reduce the costs of traditional fine-tuning methods and improve performance, the low-rank adaptation (LoRA) is explored by Nunes et al. (2025). The proposed method consists in an ensemble approach with fine-tuning BERT models on Portuguese Legal NER and LLMs. This research obtained F1-scores of 88.49% for the LeNER-Br corpus and 81.00% for the UlyssesNER-Br corpus, achieving results that are comparable to state-of-the-art benchmarks.

3.7 Visualization of Named Entities

Text visualization techniques are currently widely employed to convey implicit textual patterns in large corpora. These techniques leverage advancements in NLP by transforming complex and extensive textual data into comprehensible visual representations, enabling more effective analysis and interpretation of significant trends and insights within the data. Some works in the literature that proposed a visualization system for NER are described next.

Kucher et al. (2018) introduced DoSVis (Document Stance Visualization) to support the analysis of lengthy text documents. DoSVis provides an overview of multiple stance categories detected by a classifier at the utterance level and a detailed text view annotated with classification results, thus supporting both distant and close reading tasks.

Lo et al. (2022) investigated word segmentation (WS) and part-of-speech tagging (POS) in Chinese NER tasks through visualizations. They developed CNERVis, a visual tool enabling users to analyze NER models and their respective predictions interactively. The underlying NER model first determines the words' boundaries for an input text. After the Chinese characters are recognized as words, a POS model tags them according to their function, such as verbs, nouns, and adjectives.

Sultanum et al. (2019) introduce Doccurate, a visualization-based system to support physicians in reviewing extensive patient medical records and large clinical corpora. Physicians can examine specific parts of text documents while observing the original text and access customizable and automated tools. They can also create thematic filters based on medical taxonomies, enhancing the efficiency and transparency of text processing. A qualitative evaluation with six domain experts revealed Doccurate's potential to improve the integration of automated tools in clinical workflows, highlighting the importance of customization, trust in automation, and effective data visualization. Use case scenarios further demonstrate how the system can benefit various clinical tasks.

Rodrigues et al. (2022) describes an NLP pipeline for information extraction in forensic texts employing pre-trained BERT NER models, which were subsequently fine-tuned on custom corpora, such as CONLL, WikiNER, LeNER, etc. After that, the relationships between entities are identified to build a knowledge graph. A classical graph-based visualization is employed, where nodes and edges depict the entities and their relationships. The graph visualization provides an overview of the relevant insights obtained from the input data, making it easy to create filters, plots, and detailed reports.

The literature has some works that explore the visualization of named entities in specific domains and languages, employing interactivity, timelines, graphs, and point placement strategies. However, there is still a lack of visualizations that support the analysis of Nested NER entities, entities presenting long texts and multiple entities in different domains.

3.8 Final Considerations

After reviewing state-of-the-art methods for NER in multilingual scenarios, approaches based on feature engineering, deep learning, and transformers continue to demonstrate strong baselines for this task. Although several studies have evaluated the use of LLMs for NER, these models have not yet reached the performance levels of traditional methods such as CRF, LSTM, CNN, and BERT. Nevertheless, LLM-based approaches are frequently used for automatic corpus annotation due to their lower cost, especially when open-source LLMs are employed.

Despite existing research on Portuguese and multilingual NER, certain gaps remain in this area. To the best of our knowledge, there are no publicly available annotated corpora for NER that specifically address the presence of code-mixed content, such as the occurrence of words from other languages within Portuguese texts (or vice versa). Consequently, no studies have evaluated the performance of traditional NER models or LLMs in such contexts. This gap motivated the creation of a new corpus to support research on code-mixed Portuguese texts.

Moreover, NER visualization efforts to date have been limited to specific languages and domains. Currently, no visualization tool supports the analysis of NER in multiple languages, which is a barrier to research in multilingual and code-mixed contexts. In this sense, this research will develop a NER visualization tool for multilingual entity analysis.

Chapter 4

Named Entity Recognition Methodology

This chapter presents a pipeline to recognize named entities to transform unstructured information from academic documents to a structured format, enabling more effective and efficient information searching, as presented in Figure 4.1. Section 4.1 describes process of information extraction from the website. Section 4.2 describes the process of annotation and validation to create the corpora. Section 4.3 details the experiments conducted to decide the most adequate model for the NLP pipeline, including data splitting technique, model configuration, and evaluation metric applied to the experiments. The proposed methodology is illustrated in Figure 4.1.

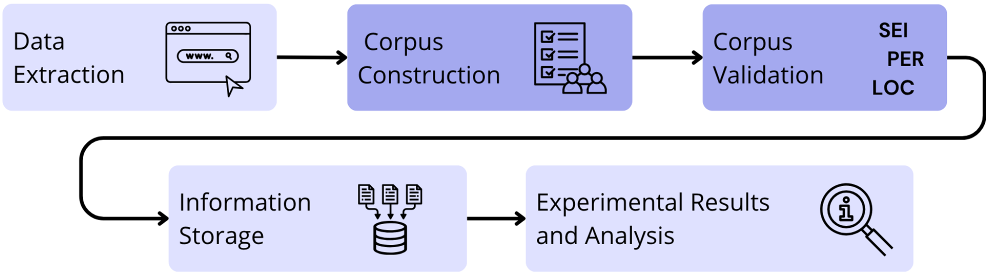


Figure (4.1) The proposed NLP pipeline with its constituent steps, in which the NLP tasks, corpus construction, and named entity recognition, are highlighted in dark lilac.

4.1 Data Extraction

The lack of available corpora reflecting the specific characteristics of academic documents motivated the construction of dedicated datasets. The types of documents selected for this study were chosen because they are representative of the academic environment, including declarations, forms, official decisions, circulars, and others. The documents were initially analyzed separately in order to better capture their individual properties, as each type presents specific and varied entity types.

4.1.1 Brazilian Electronic Information System

A web crawler was developed to collect publications from the Brazilian Electronic Information System (SEI) of University of Brasília (UnB). A Python script with Selenium API receives the URL of the University's public document search page as input and allows it to retrieve publications from UnB departments and divisions over different periods in text format. Although the UnB's SEI contains various types of documents, the web crawler was set to retrieve only leave documents, due to the presence of code-mixed sentences in this document type.

Figure 4.2 illustrates the text extraction process from a Brazilian Electronic Information System (SEI) document. Figure 4.2a displays an original SEI document as viewed by users on the public SEI website of the 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 department responsible for it. 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.

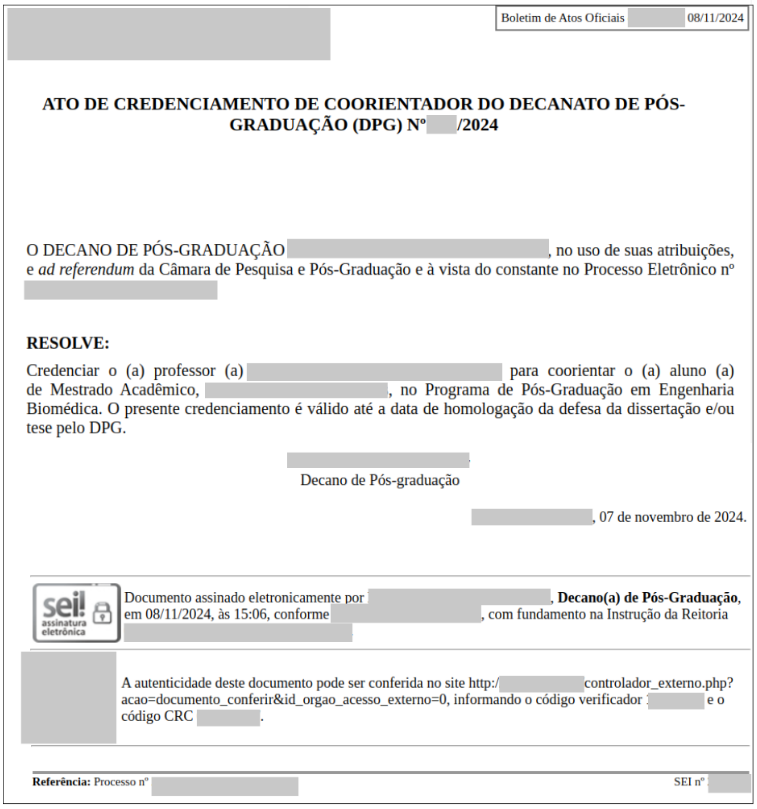
The NER model is applied to both the title and the content, as these sections contain the majority of the relevant information, including names, registration IDs, dates, secondary process IDs, and document subjects, among others. Figure 4.2b illustrates the full publication in the raw text as a result of the extraction step.

4.2 Corpus Construction

Although the literature presents some multilingual corpora (Fetahu et al. (2023a, 2021); Malmasi et al. (2022); Pan et al. (2017); Tedeschi and Navigli (2022); Tjong Kim Sang (2002); Tjong Kim Sang and De Meulder (2003)), the underlying documents do not present comprehend code-mixed texts with Portuguese and English. This limitation motivated the creation of a code-mixed corpora, consisting of a sample of WikiNER corpus and new annotated academic documents.

4.2.1 Code-mixed and Brazilian Portuguese Corpora

After searching for code-mixed corpora in the literature, it was noted that some sentences from the WikiNER corpus included both Portuguese and English, and these were selected to compose a code-mixed corpus. Therefore, a code-mixed and a Brazilian Portuguese



[redacted]

Boletim de Atos Oficiais [redacted] 08/11/2024

ATO DE CREDENCIAMENTO DE COORIENTADOR DO DECANATO DE PÓS-GRADUAÇÃO (DPG) Nº [redacted]/2024

O DECANO DE PÓS-GRADUAÇÃO [redacted], no uso de suas atribuições, e *ad referendum* da Câmara de Pesquisa e Pós-Graduação e à vista do constante no Processo Eletrônico nº [redacted]

RESOLVE:

Credenciar o (a) professor (a) [redacted] para coorientar o (a) aluno (a) de Mestrado Acadêmico, [redacted], no Programa de Pós-Graduação em Engenharia Biomédica. O presente credenciamento é válido até a data de homologação da defesa da dissertação e/ou tese pelo DPG.

[redacted]

Decano de Pós-graduação

[redacted], 07 de novembro de 2024.

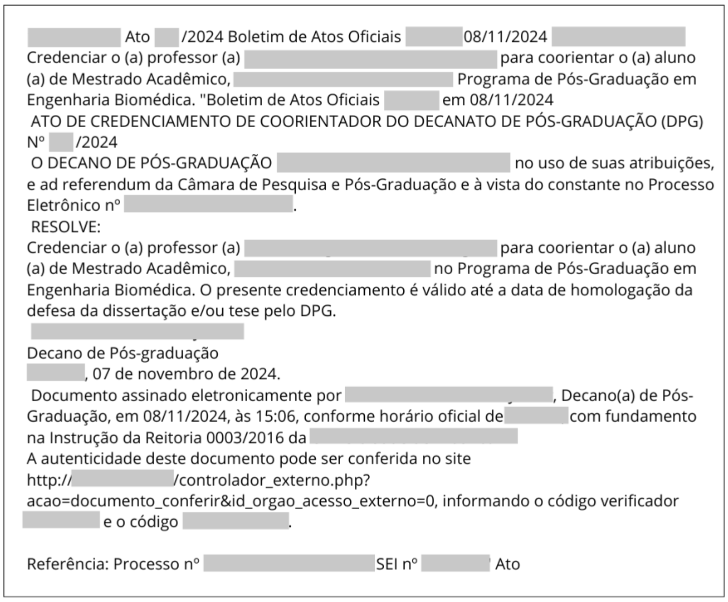
Documento assinado eletronicamente por [redacted], **Decano(a) de Pós-Graduação**, em 08/11/2024, às 15:06, conforme [redacted], com fundamento na Instrução da Reitoria [redacted]

[redacted]

A autenticidade deste documento pode ser conferida no site http:/[redacted]controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador [redacted] e o código CRC [redacted].

Referência: Processo nº [redacted] SEI nº [redacted]

(a) Web page of a SEI document.



[redacted] Ato [redacted] /2024 Boletim de Atos Oficiais [redacted] 08/11/2024 [redacted]
Credenciar o (a) professor (a) [redacted] para coorientar o (a) aluno (a) de Mestrado Acadêmico, [redacted] Programa de Pós-Graduação em Engenharia Biomédica. "Boletim de Atos Oficiais [redacted] em 08/11/2024
ATO DE CREDENCIAMENTO DE COORIENTADOR DO DECANATO DE PÓS-GRADUAÇÃO (DPG) Nº [redacted] /2024
O DECANO DE PÓS-GRADUAÇÃO [redacted] no uso de suas atribuições, e ad referendum da Câmara de Pesquisa e Pós-Graduação e à vista do constante no Processo Eletrônico nº [redacted].
RESOLVE:
Credenciar o (a) professor (a) [redacted] para coorientar o (a) aluno (a) de Mestrado Acadêmico, [redacted] no Programa de Pós-Graduação em Engenharia Biomédica. O presente credenciamento é válido até a data de homologação da defesa da dissertação e/ou tese pelo DPG.
[redacted]
Decano de Pós-graduação
[redacted], 07 de novembro de 2024.
Documento assinado eletronicamente por [redacted], Decano(a) de Pós-Graduação, em 08/11/2024, às 15:06, conforme horário oficial de [redacted] com fundamento na Instrução da Reitoria 0003/2016 da [redacted]
A autenticidade deste documento pode ser conferida no site http://[redacted]/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador [redacted] e o código [redacted].

Referência: Processo nº [redacted] SEI nº [redacted] Ato

(b) Extracted text.

Figure (4.2) Text extraction with web crawler.

corpora based on SEI were generated, and sentences from the WikiNER corpus that contain code-mixed text with Portuguese compose the WikiNER-ptmulti corpus. The categories of each corpus are explained in Tables 4.1, 4.2, 4.3, 4.4 , and 4.5. Each corpus is described next:

- **WikiNER-ptmulti:** Code-mixed sentences extracted from WikiNER (Nothman et al. (2013)) corpus. The sentences are in Portuguese, but some words are in other languages, mainly English. The original annotation was used. The corpus contains 101 documents.
- **SEI-leave-pten *(Afastamento)*:** Code-mixed leave documents extracted from SEI/UnB and manually annotated. Leave documents are official public documents for approving or revoking a time off requested by a University of Brasília (UnB) civil servant. The documents are in Portuguese but there are some words in English, such as universities, departments, programs, actions and papers. The corpus contains 172 documents. This corpus is available in two formats: a Flat version, which includes only 1-level entities, and a Nested NER version with up to two levels of nested entities.
- **SEI-leave-pt *(Afastamento)*:** Leave documents in Portuguese extracted from SEI/UnB and manually annotated. Leave documents are official public documents for approving or revoking a time off requested by a UnB civil servant. Some of the targeted entities are names of individuals, location, department names, beginning and ending dates, position, etc. The corpus contains 172 documents. This corpus is available in two formats: a Flat version, which includes only 1-level entities, and a Nested NER version with up to two levels of nested entities.
- **SEI-act-pt *(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. The corpus contains 84 documents. This corpus is available in two formats: a Flat version, which includes only 1-level entities, and a Nested NER version with up to two levels of nested entities.
- **SEI-announcement-pt *(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. The corpus contains 48 documents. This corpus is available in two formats: a Flat version, which includes only 1-level entities, and a Nested NER version with up to two levels of nested entities.

- **SEI-resolution-pt** ***(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. The corpus contains 79 documents. This corpus is available in two formats: a Flat version, which includes only 1-level entities, and a Nested NER version with up to two levels of nested entities.

Table (4.1) Predefined categories of the WikiNER-ptmulti.

Category	**Meaning**
LOC	Country and region, iconic building, natural landscape, transport lines and networks, celestial bodies
PER	Name and family, fictional characters, nationality and ethnicity
ORG	Organization, institution, government bodies, political parties UMP, companies Microsoft, musical bands, higher education institutions, military organizations
MISC	Titles of works, events, historical periods and regimes, software and hardware, conventions and documents, ships and rockets, and brands

Table (4.2) Predefined categories of SEI-leave-pten and SEI-leave-pt corpora.

Category	**Meaning**
SEI	The number of SEI process
LOC	Location
ORG	Organization's name or acronym
PER	Person's full name
NUM	Document number
BDT	Initial date of leave
EDT	Final date of leave
REG	Enrollment number that identifies a civil staff
SUB	Subject of the publication
UNI	Name and Acronyms of Departments, Secretariats, Divisions, Programs
ONU	Define if there are costs to the university or organization
POS	Full role of a person
DOU	Brazilian Official Gazette information
MOT	Leave justification
EVE	Name of events, including conferences, workshops, meetings
ART	Regulation or Article number, which includes the document type alongside its number

Table (4.3) Predefined categories of the SEI-act-pt corpus.

Category	Meaning
SEI	The number of SEI process
ORG	Organization's name or acronym
PER	Person's full name
LOC	Location
DAT	Standalone date
NUM	Document number
MAT	Identifies a civil staff
SUB	Subject of the publication
UNI	Name and Acronyms of Departments, Secretariats, Divisions, Programs
OBJ	Object of a contract or a decentralized execution term (TED)
POS	Full role of a person
ART	Document type alongside number
DOU	Brazilian Official Gazette information

Table (4.4) Predefined categories of the SEI-resolution-pt corpus.

Category	Meaning
SEI	The number of SEI process
ORG	Organization's name or acronym
PER	Person's full name
NUM	Document number
SUB	Subject of the publication
UNI	Name and Acronyms of Departments, Secretariats, Divisions, Programs
DAT	Standalone date
MET	Meeting identifier number
POS	Full role of a person
ART	Document type alongside number

4.2.2 Annotation Process

To ensure data consistency and quality, an annotation guideline was developed and was previously provided to the annotators. This guideline was designed to provide volunteers with information related to a dictionary with the categories and respective meanings, instructions for defining the beginning and end offsets of entities within the text, and illustrative annotation examples. The process involved an initial mapping of categories with supervision of an expert, followed by the creation of the guideline and then the annotation period. The Annotation Guideline is in Appendix I. The annotation comprises manual labeling of text spans by a trained annotators.

The annotation included nested entities, so the annotated corpora is appropriate for

Table (4.5) Predefined categories of the SEI-announcement-pt corpus.

Category	**Meaning**
SEI	The number of SEI process
ORG	Organization's name or acronym
PER	Person's full name
POS	Full role of a person
DAT	Standalone date
URL	The online address of announcement
TYP	Rectification and result
NUM	Number of document
WRG	Wrong information that is going to be corrected in the next lines
COR	Correction of wrong information
SUB	Subject of the publication
UNI	Name and Acronyms of Departments, Secretariats, Divisions, Programs
OBJ	Object of the announcement
ART	Document type alongside number
MAT	Identifies a civil staff
LOC	Location
CPF	Identifies a citizen

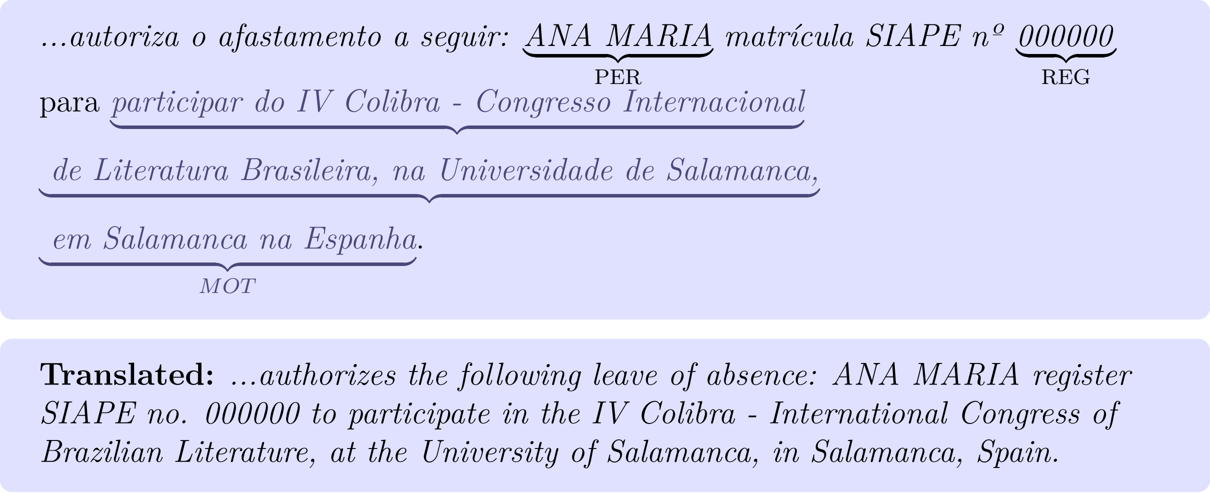
both flat and nested NER tasks. In the annotated documents, a total of two levels of nested entities were identified, where the categories ONU, MOT, and ART contain the nested entities. For each annotated corpora, two versions are available: one focused on entities at the first hierarchical level (Flat NER), and another that includes the first and second hierarchical level entities (Nested NER). Both examples are illustrated in Figures 4.3 and 4.4.

To annotate the corpus, the open-source tool LabelStudio[1] was used by the annotators as an interface to label the assigned documents, its annotation panel is presented in Figures 4.5 and 4.6. This tool allows an annotator to associate a token in the raw text with a category and also creates a history of all annotations, which is essential for the reviewing process. This method was chosen to guarantee the quality of the corpus.

4.2.3 Annotation Evaluation Metrics

An annotation evaluation process was conducted to analyze inter-annotators' agreement, including common metrics used in the literature such as Cohen's Kappa and Krippendorff coefficients. In order to capture differences between the beginning and ending offsets of

[1] `https://labelstud.io/`



...autoriza o afastamento a seguir: $\underbrace{\textit{ANA MARIA}}_{\text{PER}}$ *matrícula SIAPE nº* $\underbrace{\textit{000000}}_{\text{REG}}$ para $\underbrace{\textit{participar do IV Colibra - Congresso Internacional de Literatura Brasileira, na Universidade de Salamanca, em Salamanca na Espanha}}_{\textit{MOT}}$.

Translated: *...authorizes the following leave of absence: ANA MARIA register SIAPE no. 000000 to participate in the IV Colibra - International Congress of Brazilian Literature, at the University of Salamanca, in Salamanca, Spain.*

Figure (4.3) SEI-leave-pt in Flat NER version.

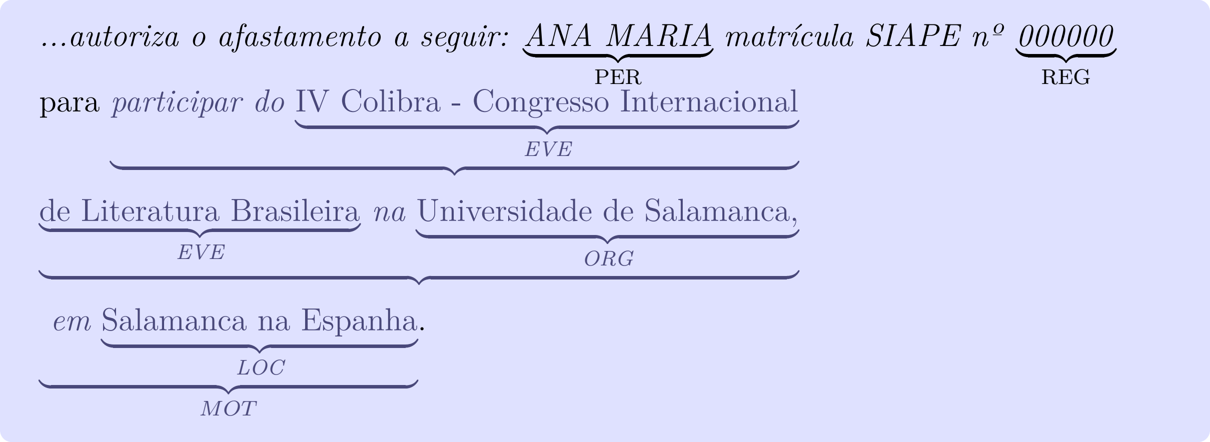
each entity, a greedy algorithm based on Sweep Line and Levenshtein distance was used to identify how similar the boundaries of the annotated entities are.

Algorithm 2 addresses the task of comparing two sets of entity annotations for a given sequence of tokens, whose goal is to extract aligned pairs of entity strings, one from each annotator, considering the first longest sequence of each from the moment there is a token with the same category in both sequences. The longest sequence is obtained greedily and follows the steps of the sweep line.

Another considered metric for the annotation evaluation is the Levenshtein distance. This metric, also known by Edit Distance, measures the similarity between two strings by computing the minimum number of operations such as insertions, deletions or substitutions required to transform one string into another. Its application involves measuring the distance between each set of extracted entities, according to Algorithm 3, whose output is the final distance of the corpus and the mean distance by entity. The lower the distance, the greater the level of annotators' agreement about the entities' boundaries.

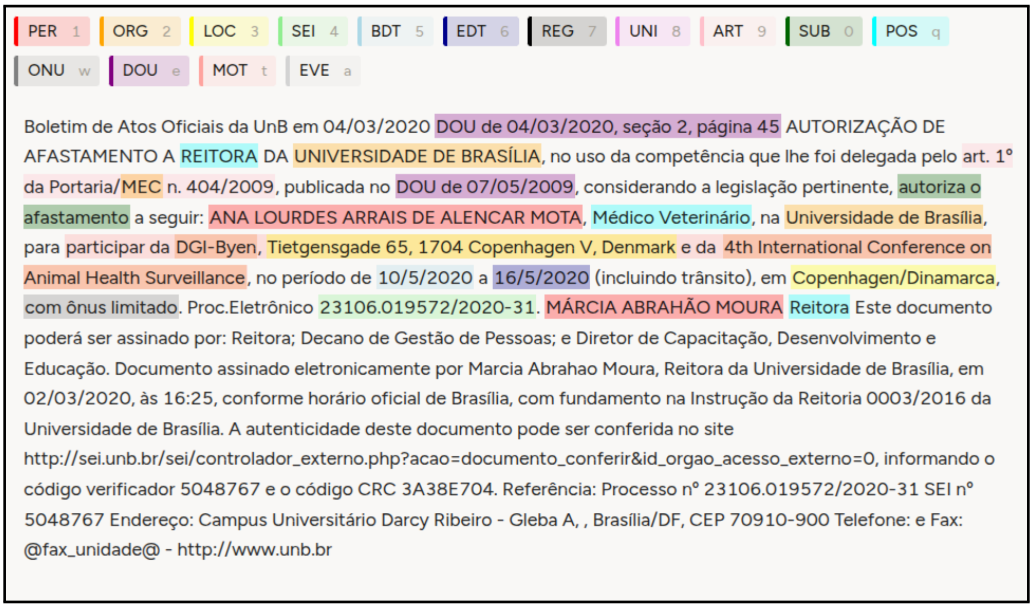
4.3 Corpus Validation

Flat and Nested NER experiments were conducted to evaluate and define an appropriate method for detecting entities for each document type, separately. The Flat NER setup used only 1-level categories, whereas the Nested NER considered both 1-level and 2-level annotations. In this sense, the experiments also establish a baseline regarding state-of-the-art NER models, and LLM-based strategies to validate the corpora.


...autoriza o afastamento a seguir: ANA MARIA matrícula SIAPE nº 000000
PER
REG
para participar do IV Colibra - Congresso Internacional
EVE
de Literatura Brasileira na Universidade de Salamanca,
EVE
ORG
em Salamanca na Espanha.
LOC
MOT

Translated: *...authorizes the following leave of absence: ANA MARIA register SIAPE no. 000000 to participate in the IV Colibra - International Congress of Brazilian Literature, at the University of Salamanca, in Salamanca, Spain.*

Figure (4.4) SEI-leave-pt in Nested NER version.

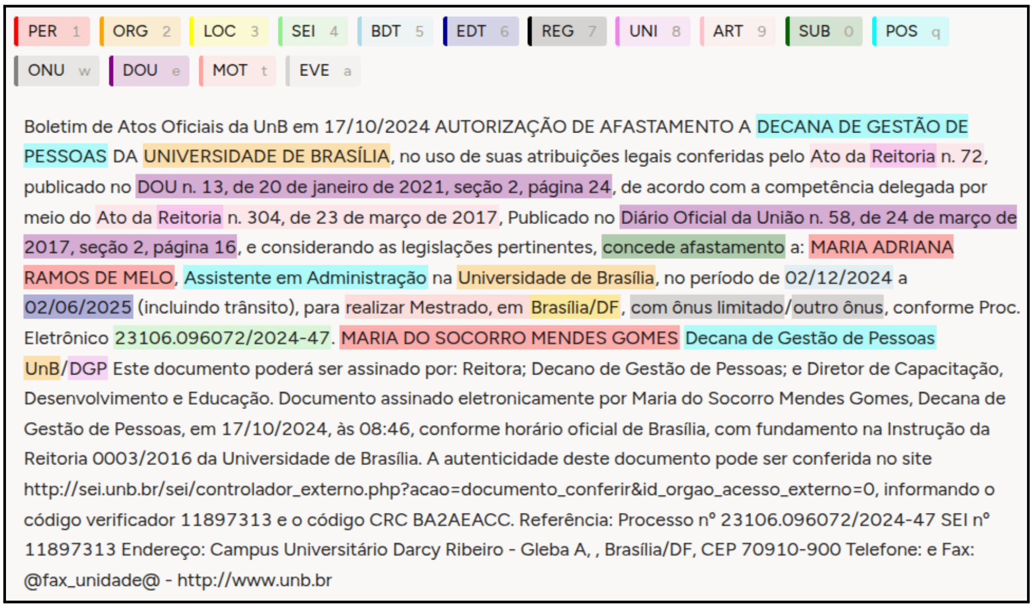


Boletim de Atos Oficiais da UnB em 04/03/2020 DOU de 04/03/2020, seção 2, página 45 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA, no uso da competência que lhe foi delegada pelo art. 1° da Portaria/MEC n. 404/2009, publicada no DOU de 07/05/2009, considerando a legislação pertinente, autoriza o afastamento a seguir: ANA LOURDES ARRAIS DE ALENCAR MOTA, Médico Veterinário, na Universidade de Brasília, para participar da DGI-Byen, Tietgensgade 65, 1704 Copenhagen V, Denmark e da 4th International Conference on Animal Health Surveillance, no período de 10/5/2020 a 16/5/2020 (incluindo trânsito), em Copenhagen/Dinamarca, com ônus limitado. Proc.Eletrônico 23106.019572/2020-31. MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação, Desenvolvimento e Educação. Documento assinado eletronicamente por Marcia Abrahao Moura, Reitora da Universidade de Brasília, em 02/03/2020, às 16:25, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília. A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 5048767 e o código CRC 3A38E704. Referência: Processo n° 23106.019572/2020-31 SEI n° 5048767 Endereço: Campus Universitário Darcy Ribeiro - Gleba A, , Brasília/DF, CEP 70910-900 Telefone: e Fax: @fax_unidade@ - http://www.unb.br

Figure (4.5) LabelStudio's annotation panel of a document from SEI-leave-pten.

4.3.1 State-of-the-art Models with Hyperparameters Tuning

Hyperparameter optimization is an essential step to improve the performance of Deep Neural Networks (Bakhashwain and Sagheer (2021)). In this step, a set of potentially suitable hyperparameters are chosen and experimented to find the subset that maximizes the evaluation metric. The set is generally chosen according to data structure, techniques used, and similar works in the literature. In this step, the corpus was split according to the Holdout method, due to its simplicity and efficiency. The full corpus was divided into



Boletim de Atos Oficiais da UnB em 17/10/2024 AUTORIZAÇÃO DE AFASTAMENTO A DECANA DE GESTÃO DE PESSOAS DA UNIVERSIDADE DE BRASÍLIA, no uso de suas atribuições legais conferidas pelo Ato da Reitoria n. 72, publicado no DOU n. 13, de 20 de janeiro de 2021, seção 2, página 24, de acordo com a competência delegada por meio do Ato da Reitoria n. 304, de 23 de março de 2017, Publicado no Diário Oficial da União n. 58, de 24 de março de 2017, seção 2, página 16, e considerando as legislações pertinentes, concede afastamento a: MARIA ADRIANA RAMOS DE MELO, Assistente em Administração na Universidade de Brasília, no período de 02/12/2024 a 02/06/2025 (incluindo trânsito), para realizar Mestrado, em Brasília/DF, com ônus limitado/outro ônus, conforme Proc. Eletrônico 23106.096072/2024-47. MARIA DO SOCORRO MENDES GOMES Decana de Gestão de Pessoas UnB/DGP Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação, Desenvolvimento e Educação. Documento assinado eletronicamente por Maria do Socorro Mendes Gomes, Decana de Gestão de Pessoas, em 17/10/2024, às 08:46, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília. A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 11897313 e o código CRC BA2AEACC. Referência: Processo nº 23106.096072/2024-47 SEI nº 11897313 Endereço: Campus Universitário Darcy Ribeiro - Gleba A, , Brasília/DF, CEP 70910-900 Telefone: e Fax: @fax_unidade@ - http://www.unb.br

Figure (4.6) LabelStudio's annotation panel of a document from SEI-leave-pt.

training (70% of full data), validation (10% of the full data), and test (20% of full data).

The most common types of hyperparameter tuning are Grid Search and Random Search (Bergstra and Bengio (2012)). The first one is an extensive search, that tries all possible combinations of hyperparameters that were given, while the second tries a fixed number of random combinations, not every one. The main advantage of Random Search is that it is capable of finding a "good" result in a reasonable time. This means that, when analyzing the results, this technique can find a local maximum, which is not necessarily the global maximum. The Grid Search was used to find the optimal hyperparameters for CRF and Random Search was used for the language models due to its complexity.

The Adam algorithm (Kingma and Ba (2014)) 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 Keras-Tuner[2] and in the CRF model with RandomizedSearchCV[3]. Therefore, the optimization was made with the validation data set from the search space predefined below for each models' hyperparameters:

- **CRF:**
    - Regularization parameter c1: exponential distribution with scale of 0.5;

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

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

Algorithm 2: Extract aligned entity sequences from dual annotations.

```
Input: Annotated token sequence df with columns chars, annotator1,
       annotator2
Output: List of entity pairs sequences between annotators
sequences, currSeq1, currSeq2, longestSeq1, longestSeq2 ← empty lists;
oldTag1, oldTag2 ← "O";
flagSave, seq1Saved, seq2Saved ← False;
foreach row in df do
    token, tag1, tag2 ← row.chars, row.annotator1, row.annotator2;
    if tag1 changes from oldTag1 then
        if flagSave and not seq1Saved then
            store currSeq1 as longestSeq1; seq1Saved ← True;
        currSeq1 ← [token];
    else if tag1 $\neq$ “O” then
        append token to currSeq1;
    oldTag1 ← tag1;
    if tag2 changes from oldTag2 then
        if flagSave and not seq2Saved then
            store currSeq2 as longestSeq2; seq2Saved ← True;
        currSeq2 ← [token];
    else if tag2 $\neq$ “O” then
        append token to currSeq2;
    oldTag2 ← tag2;
    if flagSave and seq1Saved and seq2Saved then
        append (join(longestSeq1), join(longestSeq2)) to sequences;
        reset flagSave, seq1Saved, seq2Saved;
    if tag1 = tag2 and tag1 $\neq$ “O” then
        flagSave ← True;
return sequences;
```

  - Regularization parameter c2: exponential distribution with scale of 0.05.

- **BiLSTM:**
  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Dropout rate: [0.1,0.2,0.3];
  - Learning rate: [1.10^{-2},3.10^{-2},5.10^{-2},5.10^{-3}].

- **CNN-BiLSTM:**

Algorithm 3: Compute average Levenshtein distance between entity pairs.

```
Input: List of entity string pairs sentences
Output: Sum and average edit distance
totalDistance ← 0;
foreach (seq1, seq2) in sentences do
    distance ← Levenshtein(seq1, seq2);
    totalDistance ← totalDistance + distance;
avgDistance ← totalDistance / len(sentences);
return totalDistance, avgDistance;
```

  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Dropout rate 1: [0.1,0.2];
  - Dropout rate 2: [0.1,0.2,0.3];
  - Learning rate: [1.10^{-2},3.10^{-2},4.10^{-2},5.10^{-3}].

- **BERT**
  - Model: bert-base-portuguese-cased (Bertimbau, Souza et al. (2020)), bert-base-multilingual-cased (mBERT, Devlin et al. (2018));
  - Dropout rate: [0.1,0.2];
  - Learning rate: [10^{-4}, 5.10^{-5}].

It is worth noting that BERT's tokenization is based on a subword tokenizer (WordPiece), as described in Section 2.2.3. This means that a single word may be split into multiple subtokens, which can affect the correct entity labeling for each subtoken. To align these subtokens with the original entity labels, a common strategy is to assign the entity label only to the first subtoken of each word and mask the rest during loss computation. During training, the labels are aligned to the tokenized input by propagating the label of a word to its first subtoken and using a specific value (-100 is a common choice) for the remaining subtokens to ensure they are ignored when computing loss function.

4.3.2 Fine-tuned LLMs

Besides the state-of-the-art NER techniques, experiments with fine-tuned LLMs on academic data were also conducted. Standard fine-tuning of LLMs is computationally expensive and complex due to the large number of parameters involved. To address this, the Parameter Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) (Hu et al. (2021)) was considered, a technique that reduces the number of trainable parameters, allowing for more efficient fine-tuning with fewer resources. The implementation includes LoRA to the architecture, allowing the attention vectors (query, key, value, and output) to be fine-tuned efficiently while keeping the majority of the LLM parameters frozen, then the quantization is applied, which reduces the model size and computational requirements by lowering the precision of real numbers to lower-precision representations.

- **LoRA:**
  - Rank: 16,
  - Scaling Factor: 32,
  - Droupout rate: 0.05,
  - Bias: None
- **Model's training arguments:**
  - Models: [meta-llama/llama-3.2-1B (Grattafiori et al. (2024)), deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al. (2025a))],
  - Learning rate: 5e-4,
  - Per device train batch size: 4,
  - Per device evaluation batch size: 4,
  - Number of epochs: 4

In an effort to avoid licensing costs, only publicly available models were chosen for this study. The selection included LLaMA 3.2[4], an instruction-tuned model from Meta with 1 billion parameters, known for its performance in multilingual dialogue, and the DeepSeek-R1-Distill-Qwen-1.5B[5], a distilled version of DeepSeek-R1. This distilled model was chosen because it demonstrated great performance in addressing language-mixing challenges and effectively transfers the advanced reasoning patterns of its larger counterpart into a smaller model with a more efficient architecture that allows it to perform well on complex reasoning tasks despite its size.

The LLMs adopted in this work use Byte Pair Encoding (BPE) for tokenization, which splits words into subtokens. To address this, the same alignment strategy used in BERT, which links subtokens to entity tags, is applied to the NER models fine-tuned on these LLMs. Special tokens such as `[PAD]`, `[UNK]`, and `[MASK]` are also incorporated into the tagging scheme.

[4] https://huggingface.co/meta-llama/Llama-3.2-1B

[5] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

4.3.3 Overfitting

During the neural network model training, it is essential to confirm that the model is learning patterns from the training data that are generalizable to unseen test data. An unwanted problem during this process is the overfitting, when the model complexity increases, resulting in the memorization of training data instead of learning generalizable patterns, and, as a consequence, affecting the model's capacity to make acceptable predictions on unseen data (Aggarwal (2018)). Overfitting may also be a consequence of not enough available data for training (Aggarwal (2018)).

A commonly used technique to avoid overfitting is the Early Stopping, which consists of the forced termination of training if there is no conversion to an optimal solution. In other words, after a certain number of epochs, in which the error of the loss function does not decrease, the network optimization process ends. The Dropout layer is also used to avoid overfitting by deactivating neurons. In the experiments, Dropout layer and the Early Stopping, with patience equal to 10, were techniques employed to avoid overfitting to validation data.

4.3.4 Stratified Group K-fold Cross-validation

The performance evaluation strategy proposed for the NER models was the Stratified Group K-fold cross-validation with $K = 5$, which divides the corpus into five stratified folds with non-overlapping groups for training and testing while preserving category distributions. Furthermore, using this cross-validation strategy allows for the application of statistical analysis to determine if performance differences between models are statistically significant.

4.3.5 Evaluation Metrics

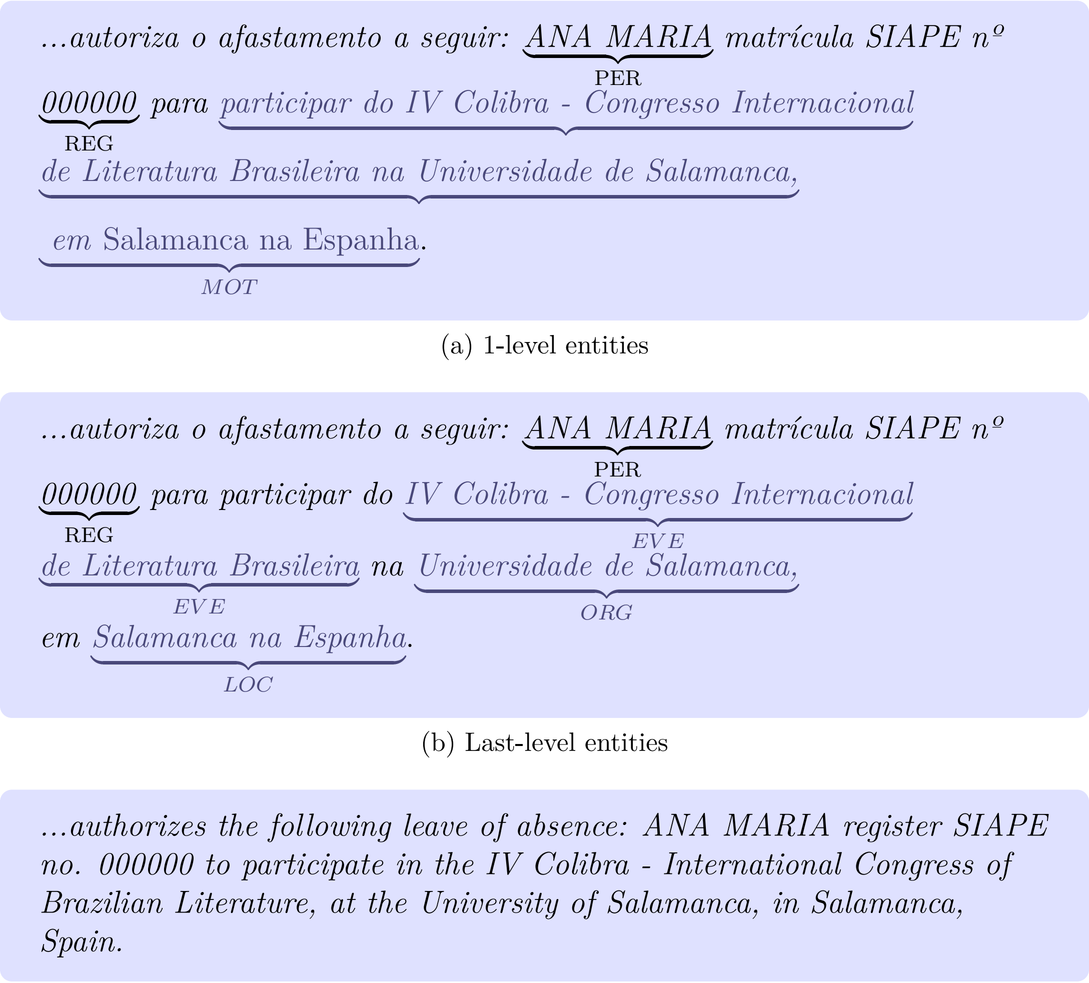
The most common evaluation metrics for NER tasks are accuracy and F1-score. When choosing the ideal metric for each experiment, it is important to analyze which one will reflect the reality. Even though accuracy is easier to calculate and more interpretable, it is not recommended if the data is unbalanced (Hagiwara (2021)), in this case, it is more appropriate to use the F1-score metric. Then, in the experiments, the F1-score was used to evaluate the results.

4.3.6 Nested NER Strategies

To validate the nested version of the corpus, three strategies were devised based on similar researches in the literature: Model-1 OR Model-n; Model-1 OR Model-n applied on filtered

1-level entities; and Model-T. These strategies were idealized based in CRF model, which was chosen for this experiment due to its simplicity and great results in previous studies, and the evaluation strategy employed was the Holdout.

In the experiments to validate the proposed strategies, the nested versions of the SEI-leave-pten and SEI-leave-pt corpora were used. A CRF model was trained on each corpus under two distinct Flat NER tasks: (1) recognition of 1-level entities and (2) recognition of last-level entities. This approach resulted in the development of two separate models. The **Model-1** was trained exclusively to identify top-level (1-level) entities, while the **Model-n** was trained only to detect the last level of entities (last-level), which includes the 2-level entities along with the 1-level entities that does not have nested entities. Figure 4.7 illustrates 1-level and last-level entities on a SEI document.



...autoriza o afastamento a seguir: ANA MARIA matrícula SIAPE nº 000000 para participar do IV Colibra - Congresso Internacional de Literatura Brasileira na Universidade de Salamanca, em Salamanca na Espanha.

(a) 1-level entities

...autoriza o afastamento a seguir: ANA MARIA matrícula SIAPE nº 000000 para participar do IV Colibra - Congresso Internacional de Literatura Brasileira na Universidade de Salamanca, em *Salamanca na Espanha*.

(b) Last-level entities

...authorizes the following leave of absence: ANA MARIA register SIAPE no. 000000 to participate in the IV Colibra - International Congress of Brazilian Literature, at the University of Salamanca, in Salamanca, Spain.

(c) Translation

Figure (4.7) Examples of annotated sentences with 1-level and last-level entities.

After that, three strategies were compared to create a method that can predict both 1-level and 2-level entities. The strategies are described next:

- **Model-1 OR Model-n**: this strategy merges the predictions from two distinct models: one trained to recognize top-level entities (Model-1) and another trained to identify last-level entities (Model-n). An entity is considered predicted if it is identified by either model.
- **Model-1 OR Model-n applied on filtered 1-level entities**: similar to the previous approach, but the last-level model is only applied to entities whose categories are known to frequently contain nested entities (`MOT`, `ART`) or that are themselves nested entities (`EVE`, `ORG`, `UNI`, `ART`, and `LOC`).
- **Model-T**: a transfer learning-based approach in which the nested entity recognition model is initialized using the parameters of the Model-1 and is retrained on last-level entities, allowing it to leverage previously learned knowledge and potentially improve generalization to nested structures.

4.4 Final Considerations

The investigation of available corpora for multilingual NER **(RQ1)** was conducted, however, no corpus was found in the literature that accomplishes the requirements, which motivates the creation of new code-mixed corpora that include Brazilian Portuguese. The creation of the SEI Corpus contributes to the development of more robust NER models for code-mixed corpora **(RQ1)**. A corpus evaluation experiment was conducted and an evaluation strategy was designed to measure inter-annotators agreement based on Cohen's Kappa, Krippendorff's alpha and Levenshtein distance **(RQ1)**.

According to the literature, the state-of-the-art methods for NER are combinations of the following models: CNN, LSTM, CRF, and BERT **(RQ2)**. The experiments can be performed multiple times with the Stratified K-Fold strategy, the macro F1-score can be used to measure the effectiveness of each NER model on identifying the predefined entities, and statistical tests should be employed to compare the results of each fold and conclude which model has a significant difference among the others. Experiments with Flat and Nested NER were considered **(RQ2)**. Finally, experiments with LLaMA and DeepSeek were conducted to analyze the performance of LLMs fine-tuned with LoRA on academic documents **(RQ3)**.

Chapter 5

Visualization of Named Entities

The lack of visualizations that allow a deeper analysis of NER in multiple languages was the motivation to the development of a user-friendly Interactive Web Application named NEVis, which is designed to visualize the results of NER and Nested NER models. NEVis provides a comprehensive toolset for researchers and practitioners to study, evaluate, and compare NER performance across diverse linguistic and domain-specific contexts. NEVis provides more profound insights into entity relationships and model behavior by integrating customized text visualization techniques.

5.1 Visualization Process Pipeline

NEVis was designed following the data visualization process pipeline described by Munzner (2014) and Ward et al. (2015) as illustrated in Figure 5.1. This pipeline outlines the steps required to transform raw data into effective visual representations that support user understanding and exploration. The first step is the data acquisition and transformation, where raw datasets are collected, preprocessed, and structured into a suitable format. Next, in the data mapping step, relevant data attributes are mapped to visual properties such as position, color, or size. This is followed by the visual encoding and visual representation (layout) design, where visual elements are arranged spatially and characterized using color scales, shapes, or layouts to improve interpretability through human visual perception. The view rendering step involves drawing the visual representation on the screen using technologies like SVG or Canvas. Finally, the pipeline incorporates user interaction, enabling actions like zooming, filtering, and highlighting, which allow users to refine visual analysis and support iterative exploration.

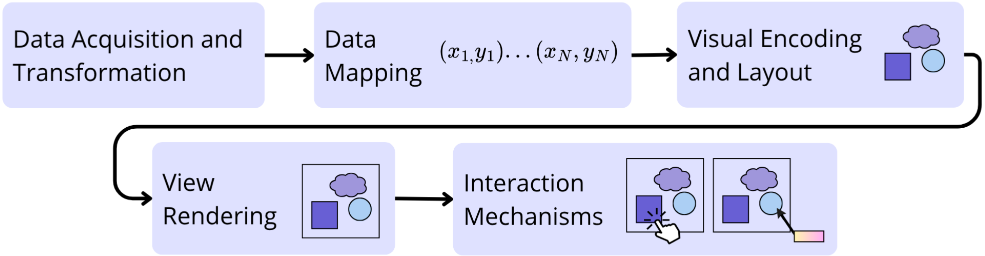

Data Acquisition and Transformation
Data Mapping
$(x_1,y_1)\ldots(x_N,y_N)$
Visual Encoding and Layout
View Rendering
Interaction Mechanisms

Figure (5.1) The steps comprising the visualization process pipeline proposed by Munzner (2014) and Ward et al. (2015).

5.2 Requirements and Tasks

In the field of visualization, defining system requirements and design tasks ensures that the resulting visualization, tool or system effectively supports users in achieving their expectations. These tasks connects the problem domain and the visual design, supporting to guide the development process in actual user needs and data characteristics. According to Munzner (2014), design tasks describe what patterns users are expected to view, explore, or compare in the data in order to accomplish their goals, while system requirements specify the functional and visual capabilities that the tool must provide. This process not only guides the choice of visualization techniques nor provide the fundamentals to customize a novel visualization, but also for evaluating whether the system accomplishes its intended purpose. It is important to note that without the definition and understanding of tasks and requirements, there are risks that the proposed visualizations would not be effective and cannot convey the expected patterns in data (Meyer et al. (2015)).

The following key features were formulated to guide the development of NEVis, having in mind that the target users are researchers and enthusiasts in NLP, who would have the support on tasks for the visual analysis of entities and the explainability of NLP model's results. Moreover, this visualization approach would be useful when integrated into interactive environments like Google Colab and Jupyter notebooks, where it can support on-site analysis alongside code and model outputs. Were identified by using the well-known tool displacy and by the systems developed by other research in literature that does not support the nature of all types of entities found in complex and large texts.

Requirements

The following requirements were designed based on key features that NLP researchers and developers would consider desirable in a visualization of named entities:

- **R1.** Simultaneously display multiple named entities with their respective labels;
- **R2.** Visualize nested entities by explicitly representing their sub-entity structures;

- **R3.** Provide contextual information for each entity;
- **R4.** Visualize entities described by long texts, preserving the visual quality of the overall layout and remaining entities;

Design Tasks

The following tasks were formulated to address the requirements, providing a meaningful and valuable visual representation of entities for the target users.

- **T1.** The visualization must display the entities in the text, using colors to identify entity labels and preserving entity ordering. The entities' positions and lengths in the input text must be preserved. (R1, R4).
- **T2.** The visualization must recursively display sub-entities, preserving their relative positions within the parent entity and the original text. (R1, R2).
- **T3.** The visualization must allow users to view surrounding words (neighboring tokens) when interacting with a specific entity. (R3).
- **T4.** The visualization must fit long entities without compromising the overall visual layout or the interpretability of the remaining entities. (R1).

5.3 General Idea

The design steps of NEVis are summarized below according to the Visualization Process Pipeline:

1. **Data acquisition and transformation:** The user loads a JSON file containing the named entities for a single document. This data is then transformed into a hierarchical dictionary structure that preserves the information about entities and their sub-entities.
2. **Data mapping:** The entities in the dictionary are sorted in ascending order based on their starting offsets. An offset denotes the position of the first character of an entity's text in the input text. Each entity is assigned a color according to its label using a predefined color scale.
3. **Visual encoding and layout design:** the proposed algorithm for visualizing named entities in NEVis renders a layout displaying entities as individual blocks and their hierarchical relationships using nested blocks. Each block includes the entity's text span and its label. If an entity contains sub-entities, the algorithm recursively

renders them inside the parent block, using a reduced height to visually represent the nested structure. The final layout preserves the natural order of entities, ensures that their widths reflect the number of characters in the original text, and uses different color coding to distinguish between entity labels. This approach enables a compact and intuitive visualization of nested named entities within a text.

4. **View rendering:** D3.js and HTML5/CSS are used to render the entity blocks on the screen, producing the final visual representation.
5. **Interaction mechanisms:** Users can interact with individual blocks to reveal additional semantic context, such as neighboring text and hierarchical relationships.

Those steps are detailed in the following sections.

5.4 Input Text Format

A hierarchical dictionary file for named entities can be structured to store both the main entities and their relevant metadata in an organized and hierarchical manner. In this format, each key represents an entity mention, while the corresponding value is a dictionary containing properties such as the entity's start and end offsets within the source text, its label or category (e.g., PERSON, LOCATION), the text span itself, and a color attribute used for visualization purposes. This structure also supports nested entities by allowing sub-entities to be represented within the parent entity's dictionary, either as a list of embedded dictionaries or as a recursive key-value structure. Table 5.1 details the key-value pairs that are used within NEVis.

Table (5.1) Key-value pairs in the hierarchical entity dictionary.

Key	**Type**	**Description**
`name`	String	Entity name or text span. For the root, it is stated as "Root".
`text`	String	Full input text (only present at the root level).
`start`	Integer	Start character offset of the entity in the text.
`end`	Integer	End character offset of the entity in the text.
`label`	String	Entity type label (e.g., PERSON, LOCATION).
`gap_left`	Integer	Number of characters between the previous entity and this one.
`gap_right`	Integer	Number of characters from the end of this entity to the next one.
`children`	List	Nested entities inside this span, each following the same structure.

The corresponding hierarchical dictionary detailing the entities of the input text is shown in Appendix II.

5.5 Block Design

Figure 5.2 illustrates the types of entity blocks rendered by the algorithm. Figure 5.2(a) shows a simple block of a flat entity with no sub-entities, in which the width `textWidth` corresponds to the number of characters in the entity's text. The block is set to a predefined height (`height` = 40 pixels), of which 1/3 is dedicated to the top rectangle showing the entity's category. The bottom rectangle displays the entity's text in font size 18 and `padding` set to 3 pixels, except blocks of sub-entities that presents `padding` set to 1. Figure 5.2(b) presents a block containing a single sub-entity, which is rendered in the exact position where its corresponding text appears within the parent entity. Before rendering the layout, the maximum sub-entity depth is determined in order to set the rectangle height for entities with depth level 0. The heights of each subsequent sub-entity rectangles are decreased by 5 pixels, thus adding the visual cue of nesting required to address tasks **T1** and **T2**.

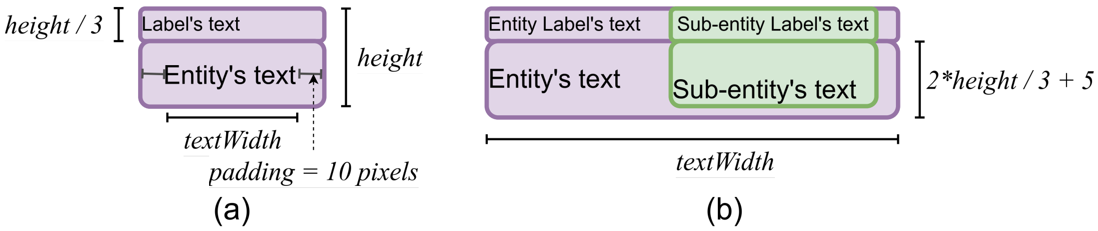


Figure (5.2) Entity blocks: (a) a simple block with no sub-entities (depth level 0); (b) a block with a single sub-entity (depth level 1).

5.6 Algorithm Design

The algorithm was designed to preserve the natural order of entities as they appear in the text in order to ensure accurate and intuitive visualization. In this setup, it is assumed a maximum of 20 entities and a maximum nesting depth of 2 levels in the hierarchy. These constraints are sufficient to avoid color ambiguity and are well-suited to the types of nested entities typically found in NLP tasks. Furthermore, the visualization must guarantee that the width of each rectangle must be proportional to the number of characters in the corresponding entity in order to convey the real entity's span in relation to the full text.

For the visual demonstration of the proposed algorithm, the following input text and the named entities are considered:

$$\text{“The }\overbrace{\text{president}}^{\text{MISC}}\text{ of }\overbrace{\text{University of California, }\overbrace{\text{Los Angeles}}^{\text{LOC}}}^{\text{ORG}}\text{ spoke at the summit at}$$

$$\overbrace{\text{Silicon Valley}}^{\text{LOC}}\text{ on }\overbrace{\text{September 23, 2023}}^{\text{DATE}}\text{.”}$$

Figure 5.3 illustrates, step by step, the algorithm processing the entities in the input sample and rendering their respective blocks. The white box represents the layout where the blocks are drawn. Initially, the layout is empty and contains a single line with a fixed width (`maxWidth`). The first entity, "president" (1), is processed because it has the lowest starting offset. Since it has no sub-entities, its corresponding block is rendered directly in the layout. The following entity, "University of California, Los Angeles" (2), contains a sub-entity, and sufficient free space is available to fit its block in the first line. The block is then rendered, and the algorithm recursively processes each sub-entity. The single sub-entity is "Los Angeles" (3), which is rendered at the corresponding position within its parent block, but with a lower height to convey a visual cue of its sub-entity condition. As there are no more sub-entities to process, the following entity, "Silicon Valley" (4), is analyzed. However, the current line has no space left since the remaining width (`remainingWidth`) does not fit appropriately within this entity (detailed criteria are provided below). This forces the algorithm to create a new line to render the corresponding block (4), which has no sub-entities. Finally, the final block (5), representing "September 23, 2023", is also rendered in this same new line.

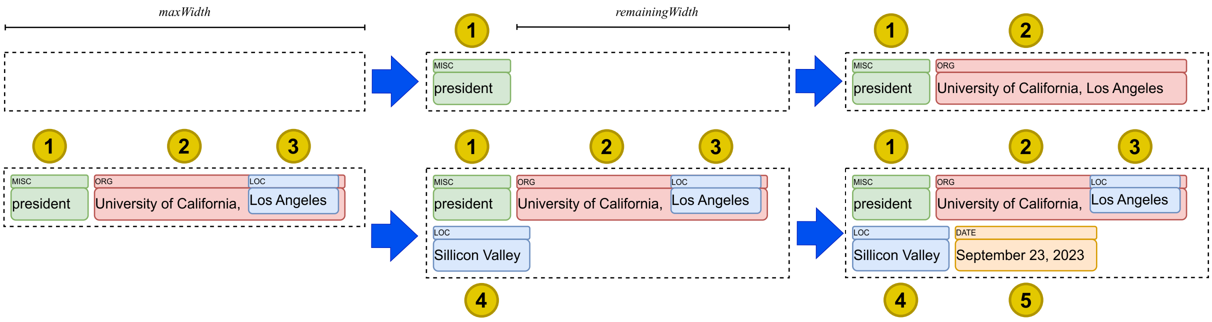


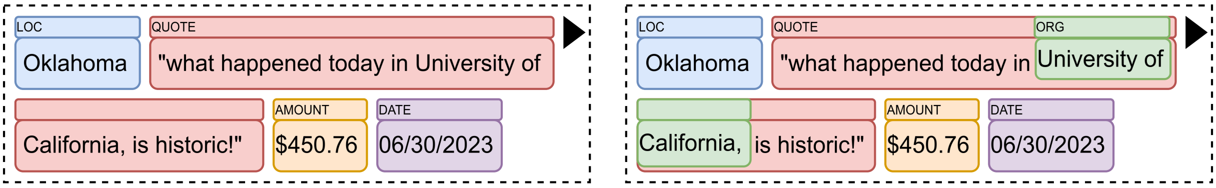
Figure (5.3) Demonstration of NEVis rendering the entities of the sample "The president of University of California, Los Angeles spoke at the summit at Silicon Valley on September 23, 2023.".

Algorithm 4 describes formally the process of creating a global layout of blocks, and the Algorithm 5 describes the process of generating the entity blocks for named entities and their sub-entities based on a hierarchical input structure. The algorithm for drawing the blocks in the multi-row layout is described below. First, entities are sorted in ascending order according to their starting offsets to preserve textual order. This sorting is recursively applied to any sub-entities. The algorithm starts from the first line and processes each top-level entity (depth 0). For each entity, it computes its corresponding text span width (`textWidth`) to determine how the block will be rendered: (1) entirely on the current line; (2) entirely on the following line; or (3) partially on the current line and continuing on subsequent lines. If `textWidth` is smaller than the available space (`remainingWidth`), the block is drawn on the same line (case 1). Otherwise, if a minimum

Algorithm 4: Create global layout of blocks.

```
Input: Hierarchical dictionary hierarchy, maximum layout width maxWidth
Output: Visual layout of named entities
fullText $\leftarrow$ hierarchy.text;
lineWidth $\leftarrow$ 0;
currentLine $\leftarrow$ new empty line in layout;
if hierarchy.children is not empty then
    foreach entity in hierarchy.children do
        entityText $\leftarrow$ fullText[entity.start:entity.end];
        entityWidth $\leftarrow$ measure width of entityText;
        if lineWidth + entityWidth $\leq$ maxWidth then
            generateEntityBlock(entity, currentLine, fullText);
            lineWidth $\leftarrow$ lineWidth + entityWidth;
        else
            currentLine $\leftarrow$ new empty line in layout;
            lineWidth $\leftarrow$ entityWidth;
            generateEntityBlock(entity, currentLine, fullText);
```

of 10 characters, or none, can not be fit in `remainingWidth`, the block is drawn entirely on the following line (case 2). Finally, in case 3, the block is split across multiple lines as illustrated in Figure 5.4(a). The algorithm renders part of the block on the current line, ending at the last word that fits and displaying (▶), and continues rendering the remaining content on the subsequent lines. If the current entity contains sub-entities, they are recursively drawn applying the same process, overlapping their corresponding text offsets within the parent entity. Figure 5.4(b) shows case 3, when a sub-entity's block is rendered alongside its entity's block in two lines. This approach of breaking a block into consecutive lines addresses task **T4**.



(a) Flat entity.

(b) Nested entity.

Figure (5.4) Line breaking of blocks.

An interaction function was implemented to support contextual exploration of named entities relative to the input text, thus addressing task **T3**. When the user hovers the mouse cursor over a level-0 entity block, a tooltip shows the surrounding text, revealing the neighboring tokens to the left and right of the entity. This information is essential

Algorithm 5: `generateEntityBlock()`: Render entity and its sub-entities as blocks

```
Input: Entity entity, layout line currentLine, full text fullText, maximum
       width maxWidth
Output: Visual layout of the entity and its sub-entities within currentLine
entityText ← fullText[entity.start:entity.end];
labelWidth ← Measure width of entityText;
if currentLine.width + labelWidth > maxWidth then
    currentLine ← Create new empty line in the layout;
Draw top rectangle for the entity label in currentLine;
Draw bottom rectangle for entityText in currentLine;
subentityRegion ← Region inside the current entity block;
Sort entity.children by starting offset;
lastIndex ← entity.start;
foreach subentity in entity.children do
    if subentity.start > lastIndex then
        Draw intermediate text from lastIndex to subentity.start inside
         subentityRegion;
    generateEntityBlock(subentity, subentityRegion, fullText);
    lastIndex ← subentity.end;
if lastIndex < entity.end then
    Draw remaining text inside subentityRegion;
```

to understanding the local context in which the entity appears. This feature helps NLP practitioners and domain experts to gain insights concerning the named entity that the NER model associates with a particular text span.

5.7 Final Considerations

NEVis is tool for visualizing named entities to support tasks related to analyzing and interpreting the outputs of NER models. The proposed technique can handle flat and nested entities, unlike existing approaches that address only flat entities. Moreover, it handles longer texts with multiple entities, which are displayed as blocks in a compact multi-row layout, effectively addressing spatial limitations not tackled by current approaches **(RQ4)**. The method also incorporates a tooltip interaction for showing surrounding text when the mouse hovers over the blocks, to provide contextual information and enhance the understanding of entities.

Moreover, evaluating the quality and usefulness of visualizations is a challenging task, because unlike traditional NLP models, in which accuracy, precision, recall, and F1-score provide objective performance indicators, visualizations are interpretive tools that support human understanding and decision-making (Munzner (2014)). Thus, their effectiveness depends on subjective factors like usability, clarity, and how well they aid users performing specific tasks, such as identifying key entities or understanding the context of an entity. These aspects are difficult to capture and asses with the aforementioned metrics. A common and widely accepted approach in the field of visualization is to conduct user studies involving both qualitative feedback and task-based performance measurements (Carpendale (2008); Islam et al. (2024); Lam et al. (2012)). Such evaluations can provide insights into how users interact with the visualizations, whether the tool improves understanding regarding the displayed entities.

As future work, we plan to conduct a formal user evaluation involving participants with varying levels of NLP expertise to assess the effectiveness of NEVis in accurately conveying the entities and associated contextual information. Furthermore, we consider investigating graph-based visualizations as an alternative visual metaphor to convey the complex hierarchical entity relationships appearing in extensive text corpora.

Chapter 6

Experimental Results

This section presents the main results obtained from the research, which include the construction of a code-mixed corpora in Portuguese and English, the implementation and evaluation of classical and state-of-the-art NER models, and the development of a visualization system for named entities, named NEVis. Experiments were conducted to validate the proposed methodology, assess the quality of the annotated corpora, and compare the performance of flat and nested NER approaches using statistical analysis. Moreover, visualizations generated by NEVis, which was implemented to support the visual interpretation of entity recognition results in long and complex documents, are presented using real-world and synthetic samples. All corpora and source code developed in this work are publicly available in the project repository[1].

6.1 Experiments

Flat NER experiments were performed on WikiNER-ptmulti, which includes code-mixed texts from diverse themes, on the institutional code-mixed corpus SEI-leave-pten, and on the Brazilian Portuguese corpora: SEI-leave-pt, SEI-act-pt, SEI-resolution-pt and SEI-announcement-pt. The nested NER task was explored on SEI-leave-pten and SEI-leave-pt corpora.

The construction of a golden standard corpus requires manual labeling, which is complex, costly, and prone to errors, and also needs a peer review of the annotations. So, the availability of annotators was the main factor in defining the number of documents selected for annotation. Figure 6.1 illustrates the annotation frequency observed during the corpus annotation phase of the corpus evaluation experiment. The annotators were volunteers, and their availability directly influenced the annotation rate, which prevented the establishment of a consistent number of documents annotated per day. Annotation

[1] https://gitlab.com/gvic-unb/multilingual-ner-on-institutional-documents/

is a labor-intensive task that requires prior domain knowledge and attention to detail. In this experiment, Annotator 1 and Annotator 2 spent approximately 7 and 12 hours, respectively, on the annotation process.

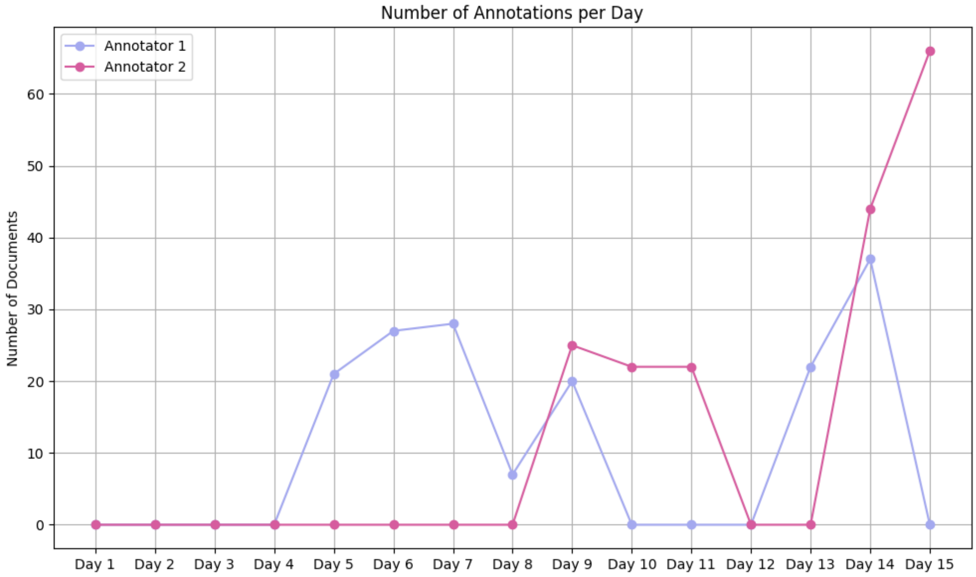


Figure (6.1) Annotation frequency during the annotation phase of the corpus evaluation experiment.

In the revision phase of the corpus evaluation experiment, the annotation agreement metrics were calculated, and some corrections were asked of the annotators. This process was repeated until the annotation coefficients in 1-level were close to the range indicated by similar studies in the literature (Fleiss' Kappa score: 86% Silva et al. (2023), Cohen's Kappa scores: 91%, 94%, and 88% (Albuquerque et al. (2022))). Finally, after finishing all disagreements, the corpus with reliable annotations of the SEI documents were generated. The annotation agreement metrics and distance measures are shown in Table 6.1, considering a maximum nesting depth of two levels in both corpora.

Table (6.1) Annotator inter-agreement metrics and distance.

Document type	**Entity's level**	**Cohen Kappa**	**Krippendorff's alpha**	**Mean Levenshtein distance per complete entity**
SEI-leave-pten	1-level	81%	99%	2.97
	2-level	24%	94%	2.36
SEI-leave-pt	1-level	93%	99%	0.66
	2-level	73%	100%	0.42

The agreement metrics indicate that both document types had a higher level of agreement between the annotators in 1-level entities. However, the Cohen's Kappa for 2-level entities of SEI-leave-pten documents was below expected, resulting in more time resolving the disagreements. This value was negatively impacted by the disagreement of the annotators about the Event (`EVE`) entities and the variation in the starting positions of the `EVE` annotations, which is a nested entity of Leave Motive (`MOT`) and may include conferences, conference's acronyms, meetings, workshops, event's editions and year, conference tagline. In this sense, Organization (`ORG`) entities were frequently confused with `EVE` entities.

Regarding Levenshtein distance, the SEI-leave-pten had a higher mean distance per complete entity, which means that one annotator tended to include more characters than the other when marking the boundaries of the same entity category, either at the beginning or the end. Then, it can be concluded that there was more disagreement about entity boundaries in SEI-leave-pten documents comparing to SEI-leave-pt documents.

Table 6.2 provides a statistical comparison between tokens and sentences of each type of document in the corpora. In general, each instance of WikiNER-ptmulti contains a short sentence, while each instance of SEI corpora contains more than one sentence, whereas SEI-leave-pt documents are longer than the SEI-leave-pten. Thus, it is also worth noting the imbalance regarding the categories in the corpus, as shown in Figures 6.2-6.7.

Table (6.2) Basic statistics of the corpora.

Corpus	**Categories**	**Annotated Entities (nested included)**	**Documents**	**Avg. Length/ Document**	**Avg. Tokens/ Document**
WikiNER-ptmulti	4	713	101	108.96	32.55
SEI-leave-pten	16	14825	172	1263.22	206.35
SEI-leave-pt	16	17957	172	1672.26	279.40
SEI-act-pt	14	20045	84	2322.5	333.37
SEI-announcement-pt	19	9132	48	4317.65	607.75
SEI-resolution-pt	13	15237	79	5967.01	852.71

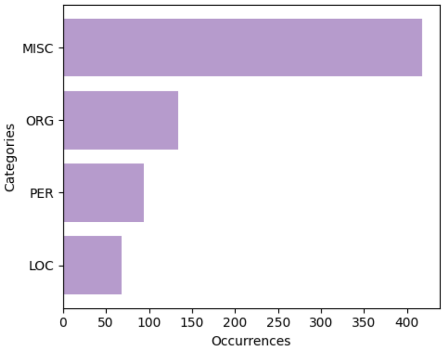
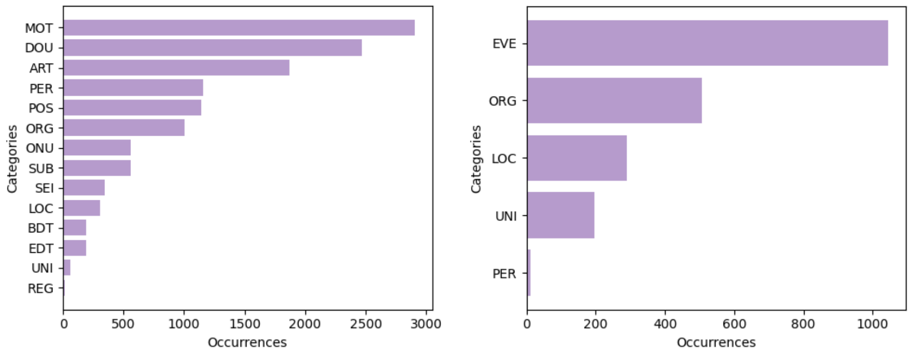


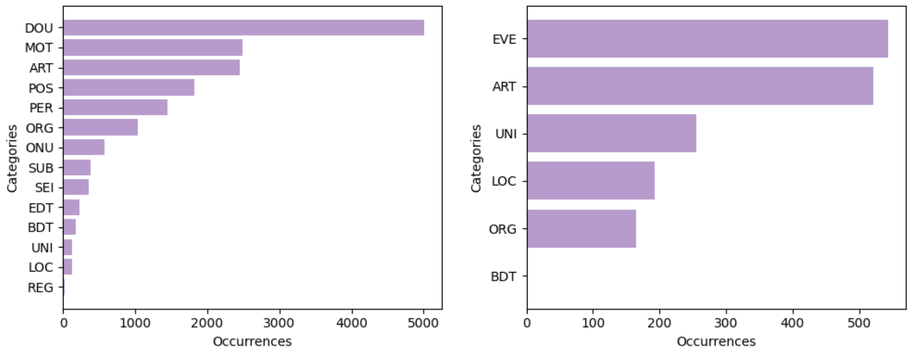
Figure (6.2) Category frequency in WikiNER-ptmulti documents.



(a) SEI-leave-pten categories (1-level).

(b) SEI-leave-pten nested categories (2-level).

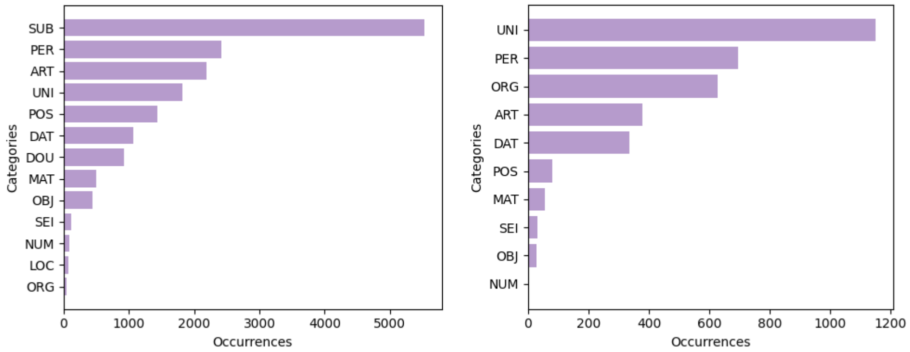
Figure (6.3) Category frequency in SEI-leave-pten documents.



(a) SEI-leave-pt categories (1-level).

(b) SEI-leave-pt nested categories (2-level).

Figure (6.4) Category frequency in SEI-leave-pt documents.

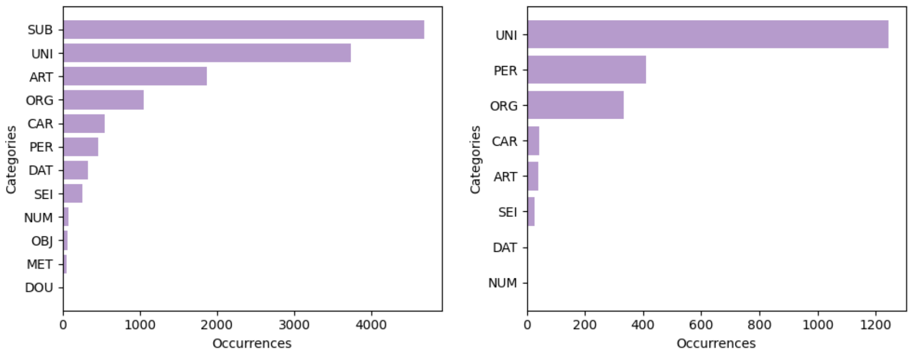


(a) SEI-act-pt categories (1-level).

(b) SEI-act-pt nested categories (2-level).

Figure (6.5) Category frequency in SEI-act-pt documents.

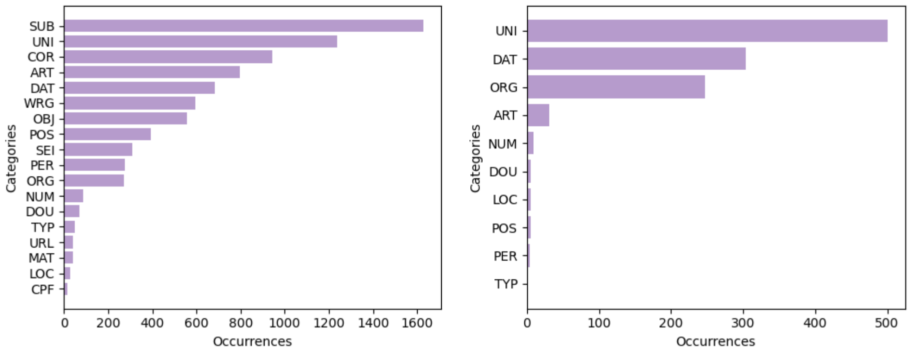
The SEI-leave-pten corpus contains, on average, 86.19 annotated entities per document, of which approximately 7.40 are in English. The most frequently occurring English words are illustrated in the word cloud shown in Figure 6.8, where word size reflects



(a) SEI-resolution-pt categories (1-level).

(b) SEI-resolution-pt nested categories (2-level).

Figure (6.6) Category frequency in SEI-resolution-pt documents.



(a) SEI-announcement-pt categories (1-level).

(b) SEI-announcement-pt nested categories (2-level).

Figure (6.7) Category frequency in SEI-announcement-pt documents.

frequency, which means that larger words appear more often in the corpus. In this visualization, the stopwords were removed.

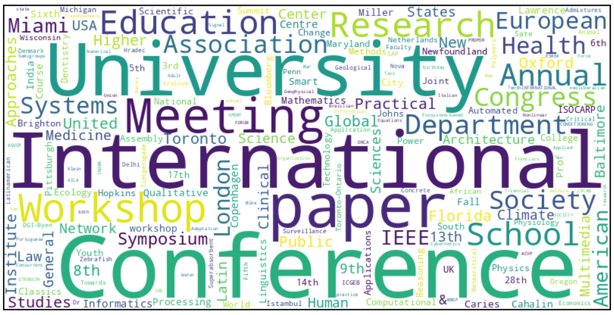
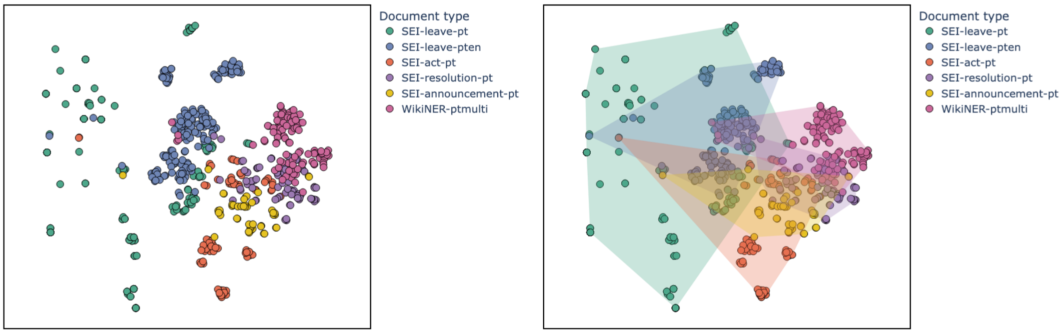


Figure (6.8) Word cloud generated with the most common English words extracted from SEI-leave-pten.

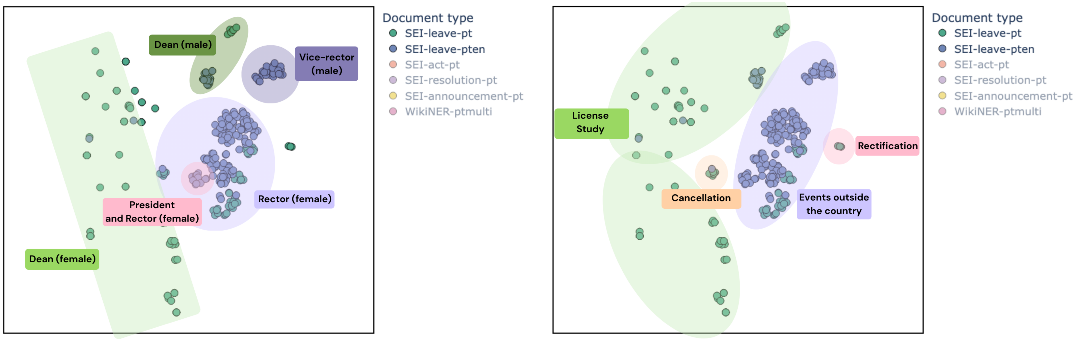
Figure 6.9 presents visualizations based on t-SNE, in which each point represents a document, and its position reflects local relationships among documents. The embeddings were generated with Sentence Transformers (“paraphrase-multilingual-MiniLM-L12-v2”) and the 2D projection was generated with t-SNE. The visualization suggests that Brazilian Portuguese and code-mixed leave documents are similar to each other and that the resolutions, announcements, and some acts may have common characteristics due to their proximity.

The analysis of documents within the SEI corpora uncovered insights beyond the language distinction. It was noticed that the SEI documents have distinct destinations: the Dean of People Management, the Vice-rector, the Rector, and the President and the Rector. In Brazilian Portuguese, the gender of the person occupying a position is indicated by the article of the noun and also by a suffix on the noun, thereby adding semantic information. For example, "Decano" and "Decana" are translated as Dean, and that is the reason why the genders were explicitly indicated in Figure 6.10a. This destination not only implies different titles but also indicates a slightly different information arrangement, reflecting the generation of visual subgroups. Besides, some SEI documents were leave cancellation requests or rectifications and it's documents were also visually separated from the leave requests. The cited patterns were visually indicated in Figure 6.10.



(a) Documents colored by corpus. (b) Documents delimited by polygons.

Figure (6.9) Corpora visualization using t-SNE visualizations.



(a) Main destinations. (b) Main motives for leave request.

Figure (6.10) Patterns discovered in SEI-leave-pt and SEI-leave-pten documents.

Experiments were conducted on the flat corpora, considering only 1-level entities, to quantitatively evaluate the results of the NER models of the proposed methodology and analyze the results on code-mixed multilingual texts. An overview of the models performance on each document type is presented in Table 6.3. In these experiments, four NER models were compared: BiLSTM, CNN-BiLSTM, CRF, and BERT, selected based on their successful prior use in the literature. Two BERT variants was used: bert-base-portuguese-cased (BERTimbau) and bert-base-multilingual-cased (mBERT). Both were applied to the multilingual corpora WikiNER-ptmulti and SEI-leave, while only BERTimbau was used for the Portuguese corpora.

It is interesting to note that leave documents achieved the highest results among all SEI Corpora, followed by resolution and act documents. The leave documents have a greater number of documents, and, besides that, these three document types contain fewer tokens per document, as shown in Table 6.2. In other words, these documents typically present concise and direct information, yielding better entity recognition.

In contrast, 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.

Table (6.3) Macro F1-score results for flat NER task in the corpora using Stratified K-Fold Cross Validation (**K = 5**): average and standard deviation of Macro F1-Score across all folds.

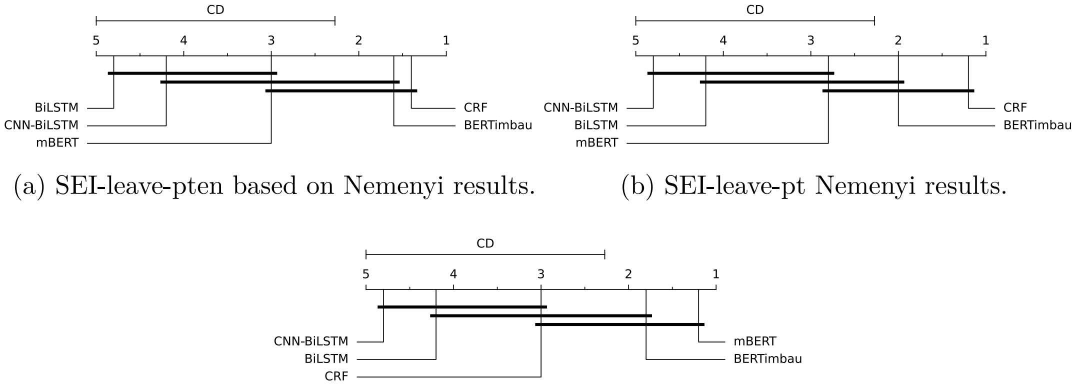
Corpus	**CRF**	**BiLSTM**	**CNN-BiLSTM**	**BERTimbau**	**mBERT**
WikiNER-ptmulti	0.62±0.03	0.25±0.11	0.22±0.12	0.75±0.05	**0.80±0.05**
SEI-leave-pten	**0.97±0.01**	0.64±0.04	0.67±0.02	**0.97±0.01**	0.88±0.01
SEI-leave-pt	**0.97±0.01**	0.74±0.03	0.60±0.19	**0.97±0.01**	0.96±0.01
SEI-act-pt	0.69±0.07	0.29±0.14	0.45±0.07	**0.74±0.04**	–
SEI-announcement-pt	**0.54±0.06**	0.30±0.07	0.09±0.07	0.23±0.02	–
SEI-resolution-pt	0.71±0.09	0.44±0.06	0.22±0.14	**0.77±0.04**	–

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 SEI-leave-pten, SEI-leave-pt and WikiNER-ptmulti documents were 0.00559, 0.00447, 0.00285, respectively. Thus, the null hypothesis is rejected, and a Nemenyi pairwise comparison was performed with a significance level of $\alpha = 5\%$. The models' significant differences are visually exemplified by the Critical Difference Diagram (Herbold (2020)) in Figure 6.11 and by Nemenyi results presented in Table 6.4.



(a) SEI-leave-pten based on Nemenyi results.

(b) SEI-leave-pt Nemenyi results.

(c) WikiNER-ptmulti Nemenyi results.

Figure (6.11) Critical Difference diagram.

Table (6.4) Nemenyi post-hoc test significant results.

Corpus	Pairwise Model Comparison	p-value
WikiNER-ptmulti	mBERT vs BiLSTM	0.023
WikiNER-ptmulti	mBERT vs CNN-BiLSTM	0.003
WikiNER-ptmulti	BERTimbau vs CNN-BiLSTM	0.023
SEI-leave-pten	CRF vs BiLSTM	0.006
SEI-leave-pten	CRF vs CNN-BiLSTM	0.041
SEI-leave-pten	BERTimbau vs BiLSTM	0.012
SEI-leave-pt	CRF vs BiLSTM	0.023
SEI-leave-pt	CRF vs CNN-BiLSTM	0.003
SEI-leave-pt	BERTimbau vs CNN-BiLSTM	0.041
SEI-act-pt	CRF vs BiLSTM	0.036
SEI-act-pt	BERTimbau vs BiLSTM	0.003
SEI-announcement-pt	CRF vs CNN-BiLSTM	0.001
SEI-resolution-pt	BERTimbau vs CNN-BiLSTM	0.001

Critical Difference diagrams provide a visual summary of statistical comparisons between multiple models. Models connected by horizontal lines indicate no statistically significant difference in their performance according to the Nemenyi test. Analogueously, models that are not connected suggest a significant difference in their performance outcomes. According to the statistical analysis, mBERT outperformed BiLSTM and CNN-BiLSTM, and BERTimbau outperformed CNN-BiLSTM in WikiNER-ptmulti documents; CRF outperformed BiLSTM and CNN-BiLSTM, and BERTimbau outperformed BiLSTM in SEI-leave-pten documents; CRF outperformed BiLSTM and CNN-BiLSTM, and BERTimbau outperformed CNN-BiLSTM in SEI-leave-pt documents; CRF and BERTimbau outperformed BiLSTM in SEI-act-pt documents; CRF outperformed CNN-BiLSTM in SEI-announcement-pt documents; BERTimbau outperformed CNN-BiLSTM in SEI-resolution-pt documents.

The results indicate that statistically significant differences in F1-scores were found only for the model comparisons listed above, so there is no significant difference between the F1-scores obtained with CRF, BERTimbau and mBERT. Finally, the experiments validate the corpora and demonstrate that CRF and the BERT variations 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.

Given that each corpus has its own categories and particular characteristics, the discussion should also address these details. The results of the WikiNER-ptmulti corpus is exemplified in Table 6.5. The models showed more difficulty in predicting the `ORG` category than the others. However, according to the literature review, reported results for the code-mixed version of WikiNER were not found for comparison.

Table (6.5) Per-category F1-scores (mean and standard deviation) for flat NER on WikiNER-ptmulti documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**CRF**	**BiLSTM**	**CNN-BiLSTM**	**BERTimbau**	**mBERT**
LOC	0.69±0.08	0.29±0.20	0.14±0.18	0.77±0.06	**0.79±0.04**
MISC	0.69±0.11	0.30±0.08	0.39±0.13	0.81±0.07	**0.82±0.06**
ORG	0.56±0.12	0.27±0.20	0.22±0.12	0.58±0.09	**0.69±0.08**
PER	0.53±0.12	0.14±0.08	0.13±0.11	0.83±0.11	**0.89±0.07**
Mean ± Std	0.62±0.07	0.25±0.06	0.22±0.10	0.75±0.10	**0.80±0.07**

Tables 6.6 and 6.7 compares the results of SEI-leave-pten and SEI-leave-pt in flat NER task and is useful for analyzing multilingual NER. In the code-mixed SEI-leave-pten, was noticed that non-Portuguese words and unusual names belong to the following categories: Location (`LOC`), Motive (`MOT`), Organization (`ORG`), and Departments and Programs related to the university (`UNI`), which are highlighted in the result table. In this sense, the most challenging category to find was `LOC` followed by `MOT`, which were better identified in SEI-leave-pt than in SEI-leave-pten documents. It was also noted that Enrollment Number (`REG`) entities were not successfully identified by the BiLSTM and CNN-BiLSTM models,

a possible reason being the small number of such entities, as shown in Figures 6.3 and 6.4.

Table (6.6) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-leave-pten documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	CRF	BiLSTM	CNN-BiLSTM	BERTimbau	mBERT
ART	0.99±0.01	0.99±0.02	0.97±0.01	**1.00±0.00**	**1.00±0.00**
BDT	**0.99±0.02**	0.58±0.21	0.88±0.07	0.99±0.01	**1.00±0.00**
DOU	**1.00±0.00**	0.77±0.15	0.81±0.16	**1.00±0.00**	**1.00±0.00**
EDT	0.99±0.01	0.50±0.09	0.86±0.05	**1.00±0.00**	**1.00±0.00**
LOC	**0.89±0.05**	0.36±0.10	0.33±0.10	0.79±0.09	0.71±0.11
MOT	0.86±0.09	0.23±0.11	0.17±0.05	**0.92±0.04**	0.87±0.06
ONU	**0.99±0.02**	0.93±0.04	0.86±0.05	0.98±0.03	0.95±0.05
ORG	**0.99±0.01**	0.97±0.02	0.90±0.08	0.98±0.01	0.98±0.01
PER	**0.99±0.01**	0.65±0.06	0.66±0.05	**0.99±0.01**	**0.99±0.00**
POS	0.97±0.01	0.95±0.03	0.89±0.05	**0.99±0.01**	0.97±0.01
REG	**1.00±0.00**	0.00±0.00	0.00±0.00	0.93±0.13	0.00±0.00
SEI	0.99±0.01	0.19±0.25	0.24±0.10	**1.00±0.01**	0.99±0.01
SUB	**0.99±0.01**	0.95±0.02	0.91±0.04	**0.99±0.01**	**0.99±0.01**
UNI	**0.97±0.02**	0.96±0.03	0.86±0.16	**0.97±0.05**	0.84±0.18
Mean ± Std	**0.97±0.04**	0.64±0.33	0.67±0.32	**0.97±0.05**	0.88±0.26

Table (6.7) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-leave-pt documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	CRF	BiLSTM	CNN-BiLSTM	BERTimbau	mBERT
ART	0.98±0.01	0.97±0.01	0.66±0.37	**1.00±0.00**	**1.00±0.01**
BDT	**0.98±0.02**	0.86±0.09	0.80±0.13	0.97±0.03	0.96±0.03
DOU	0.99±0.01	0.95±0.04	0.67±0.31	**1.00±0.01**	**1.00±0.01**
EDT	**0.98±0.02**	0.66±0.04	0.71±0.13	**0.98±0.01**	0.97±0.02
LOC	**0.92±0.06**	0.59±0.10	0.17±0.23	0.82±0.13	0.78±0.15
MOT	**0.95±0.03**	0.48±0.09	0.26±0.17	0.93±0.02	0.94±0.03
ONU	**0.96±0.04**	0.91±0.04	0.61±0.23	**0.96±0.01**	0.97±0.01
ORG	**1.00±0.00**	0.98±0.02	0.72±0.28	**1.00±0.00**	**1.00±0.00**
PER	0.99±0.01	0.55±0.02	0.43±0.21	**1.00±0.00**	**1.00±0.00**
POS	0.98±0.02	0.93±0.02	0.61±0.24	**0.99±0.01**	0.99±0.01
REG	**0.94±0.08**	0.00±0.00	0.38±0.47	**0.94±0.08**	0.88±0.11
SEI	**0.99±0.01**	0.48±0.26	0.55±0.32	**1.00±0.00**	0.99±0.01
SUB	**0.99±0.01**	0.96±0.01	0.86±0.07	**0.99±0.01**	0.99±0.01
UNI	**0.99±0.02**	0.98±0.02	0.97±0.03	0.98±0.03	0.98±0.02
Mean ± Std	**0.97±0.02**	0.74±0.28	0.60±0.22	**0.97±0.05**	0.96±0.06

Tables 6.8 and 6.9 compare the results of fine-tuned LLMs with LoRA on SEI's leave documents in the flat NER task. In the code-mixed scenario, both LLaMA and Distilled

DeepSeek had a zero F1-score on domain-specific categories of the corpus, such as Begin Date (BDT), End Date (EDT), Location (LOC), Motive (MOT), Cost (ONU), Enrollment number (REG), SEI number (SEI), Departments and Programs related to the university (UNI), predicted entities in two domain-specific categories (Article (ART), DOU number (DOU), Subject (SUB)), and in the general categories Organization (ORG), Person (PER), Position (POS). In the Portuguese-only scenario, the models could predict more entities of domain-specific categories than in code-mixed scenario, but with low F1-scores compared to state-of-the-art NER models presented above. This result suggests that the fine-tuning was unable to learn domain-specific characteristics of the SEI's leave documents.

Table (6.8) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-leave-pten documents with Fine-tuned LLMs using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**Fine-tuned LLaMA with LoRA**	**Fine-tuned Distilled Deepseek with LoRA**
ART	**0.73±0.06**	0.52±0.02
BDT	0.00±0.00	0.00±0.00
DOU	0.84±0.07	**0.86±0.03**
EDT	0.00±0.00	0.00±0.00
LOC	0.00±0.00	0.00±0.00
MOT	0.00±0.00	0.00±0.00
ONU	0.00±0.00	0.00±0.00
ORG	**0.43±0.05**	0.37±0.08
PER	**0.16±0.02**	0.11±0.04
POS	0.34±0.03	**0.35±0.02**
REG	0.00±0.00	0.00±0.00
SEI	0.00±0.00	0.00±0.00
SUB	**0.61±0.06**	0.59±0.07
UNI	0.00±0.00	0.00±0.00
Mean ± Std	**0.22±0.30**	0.20±0.28

Tables 6.10 and 6.11 compares the results of SEI-leave-pten and SEI-leave-pt in nested NER task. In this experiment, three methods were evaluated: the OR operation between Model-1 and Model-n; the OR operation between Model-1 and the Model-n applied to filtered 1-level entities whose category may have nested entities; and Model-T that consists on retraining the Model-1 on last-level entities. All three strategies had a satisfactory result, achieving a mean F1-score macro of 0.99. In these corpora, the nested entities to be identified belonged to the following categories: EVE, ORG, UNI, ART and LOC; while the 1-level entities which contains the nested ones are: MOT, ART. The small number of entities and the fact that they appeared in only 2 categories were relevant to the great F1-score result.

Table (6.9) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-leave-pt documents with Fine-tuned LLMs using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**Fine-tuned LLaMA with LoRA**	**Fine-tuned Distilled Deepseek with LoRA**
ART	**0.88±0.03**	0.85±0.06
BDT	0.01±0.02	0.00±0.00
DOU	0.73±0.03	**0.80±0.06**
EDT	0.00±0.00	0.00±0.00
LOC	0.00±0.00	0.00±0.00
MOT	**0.05±0.06**	**0.05±0.06**
ONU	**0.38±0.11**	0.02±0.02
ORG	**0.53±0.04**	0.50±0.06
PER	**0.32±0.05**	0.23±0.03
POS	**0.41±0.04**	0.38±0.03
REG	0.00±0.00	0.14±0.14
SEI	**0.41±0.08**	0.05±0.04
SUB	0.74±0.03	**0.82±0.03**
UNI	**0.66±0.09**	0.47±0.05
Mean ± Std	**0.37±0.30**	0.31±0.32

Table (6.10) Per-category F1-scores (mean and standard deviation) for nested NER on SEI-leave-pten with CRF model using Holdout validation.

Category	**Model-1 OR Model-n**	**Model-1 OR Model-n on filtered 1-level entities**	**Model-T**	**Support**
POS	1.00	1.00	1.00	35
PER	1.00	1.00	1.00	35
ORG	1.00	1.00	1.00	35
SEI	0.99	0.99	0.99	35
MOT	1.00	1.00	1.00	35
EVE	1.00	1.00	1.00	24
UNI	0.93	0.88	0.93	14
ART	1.00	1.00	1.00	35
BDT	1.00	1.00	1.00	35
REG	1.00	1.00	1.00	2
ONU	1.00	1.00	1.00	35
DOU	1.00	1.00	1.00	35
SUB	0.99	0.99	0.99	34
LOC	0.99	0.99	1.00	35
EDT	1.00	1.00	1.00	35
Mean ± Std	0.99 ± 0.02	0.99 ± 0.03	0.99 ± 0.02	–

Table (6.11) Per-category F1-scores (mean and standard deviation) for nested NER on SEI-leave-pt with CRF model using Holdout validation.

Category	**Model-1 OR Model-n**	**Model-1 OR Model-n on filtered 1-level entities**	**Model-T**	**Support**
POS	1.00	1.00	1.00	35
PER	1.00	1.00	1.00	35
ORG	1.00	1.00	0.99	34
SEI	1.00	1.00	1.00	35
MOT	0.99	0.97	1.00	33
EVE	0.96	0.92	0.96	11
UNI	1.00	1.00	1.00	26
ART	1.00	0.99	1.00	33
BDT	0.99	0.99	0.99	34
REG	1.00	1.00	1.00	3
ONU	1.00	1.00	1.00	34
DOU	1.00	1.00	1.00	35
SUB	0.97	0.97	1.00	35
LOC	1.00	0.92	1.00	20
EDT	0.99	0.97	0.97	34
Mean ± Std	0.99 ± 0.01	0.98 ± 0.03	0.99 ± 0.01	–

In act documents, shown in Table 6.12, 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.

In announcement and resolution documents, described in Tables 6.13 and 6.14, 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 6.7), 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.

In general, it was observed that BERT has competitive results comparing to CRF, and the pre-training in Portuguese documents plays an important role in this performance. BERT's pre-training endows it with a generalist capability, as it acquires linguistic knowledge from large corpora, allowing it to learn rich representations even with less annotated data.

Table (6.12) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-act-pt documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**CRF**	**BiLSTM**	**CNN-BiLSTM**	**BERTimbau**
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**

Table (6.13) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-announcement-pt documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**CRF**	**BiLSTM**	**CNN-BiLSTM**	**BERTimbau**
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

Regarding the experiments with code-mixed and Portuguese-only corpora, it is interesting to note that, despite the BERT model used being adapted for Portuguese (BERTimbau), the results on SEI-leave-pten were quite similar to those on SEI-leave-pt. This suggests that, even with the presence of English words, the Portuguese context was essential for entity recognition, since they are the majority of the words.

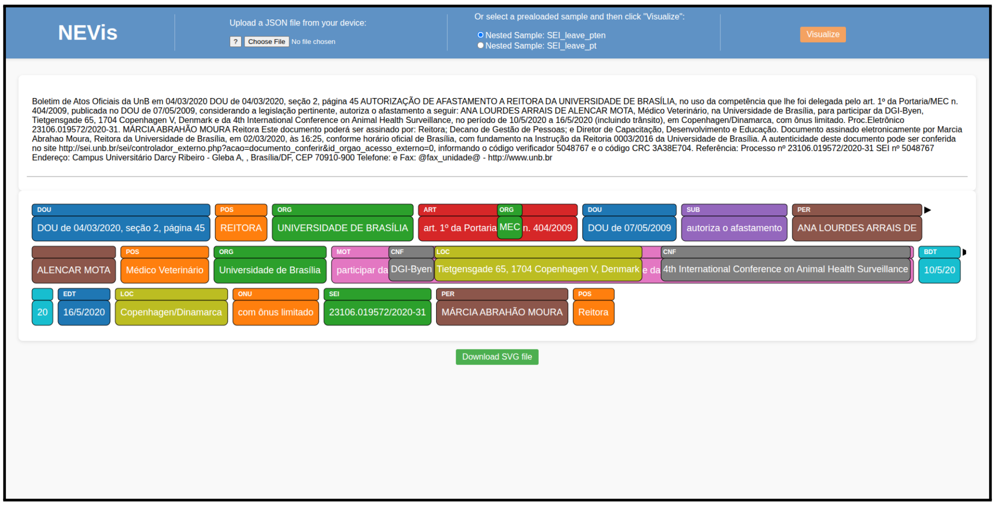
Table (6.14) Per-category F1-scores (mean and standard deviation) for flat NER on SEI-resolution-pt documents using Stratified K-Fold Cross-Validation ($K = 5$).

Category	**CRF**	**BiLSTM**	**CNN-BiLSTM**	**BERTimbau**
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**

NER models based on fine-tuned LLMs presented poor performance, likely due to the limitations of the quantization process and the fact that these models are best suited for text generation rather than information extraction. Although quantization is common for reducing model size and enabling efficient fine-tuning and inference, it can decrease the performance of LLMs. Another contributing factor is the type of tokenization adopted by some LLMs. For instance, LLaMA and DeepSeek use Byte Pair Encoding (BPE), which may negatively affect the representation of fine-grained entities in Portuguese, as it splits words into subword units in inconsistent ways. As alternative approaches, future experiments could explore improved subtoken-to-entity alignment strategies, as well as the use of other LLMs with different parameter sizes and tokenization schemes.

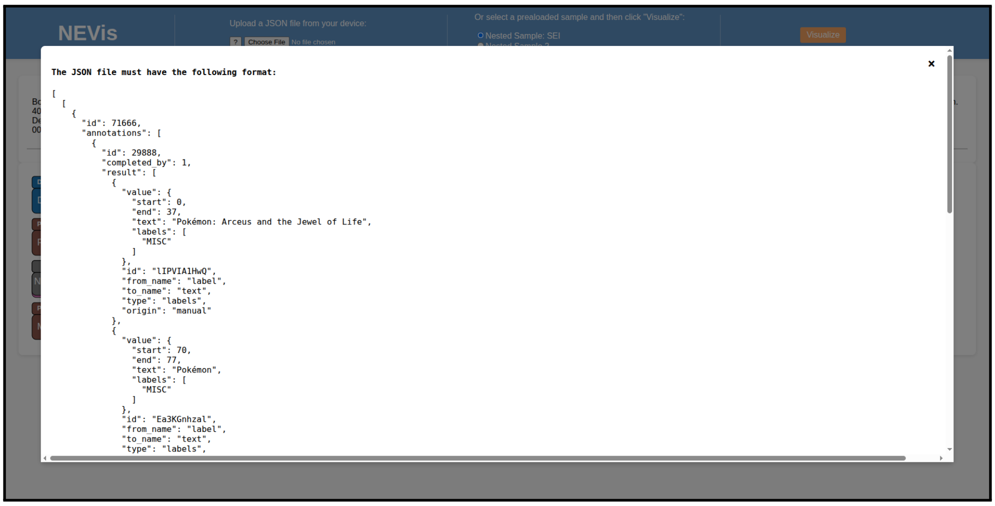
6.2 NEVis

NEVis' main panel is presented in Figure 6.12. The NEVis' main panel consists of the sidebar, which has options to choose the file in .json format or sample data. The tool also shows a popup with a format example of the input by clicking in help button ("?"), as presented in Figure 6.13. The first section has the document's raw text, and the second has the entities colored according to their category and a button to download the visualization in SVG format. Figure 6.14 illustrates the hover interaction and Figures 6.15 and 6.16 illustrate the downloaded visualization.


NEVis
Upload a JSON file from your device:
?
Choose File
No file chosen
Or select a prealoaded sample and then click "Visualize":
○ Nested Sample: SEI_leave_pten
● Nested Sample: SEI_leave_pt
Visualize

Boletim de Atos Oficiais da UnB em 04/03/2020 DOU de 04/03/2020, seção 2, página 45 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA, no uso da competência que lhe foi delegada pelo art. 1º da Portaria/MEC n. 404/2009, publicada no DOU de 07/05/2009, considerando a legislação pertinente, autoriza o afastamento a seguir: ANA LOURDES ARRAIS DE ALENCAR MOTA, Médico Veterinário, na Universidade de Brasília, para participar da DGI-Byen, Tietgensgade 65, 1704 Copenhagen V, Denmark e da 4th International Conference on Animal Health Surveillance, no período de 10/5/2020 a 16/5/2020 (incluindo trânsito), em Copenhagen/Dinamarca, com ônus limitado. Proc.Eletrônico 23106.019572/2020-31. MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação, Desenvolvimento e Educação. Documento assinado eletronicamente por Marcia Abrahao Moura, Reitora da Universidade de Brasília, em 02/03/2020, às 16:25, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília. A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 5048767 e o código CRC 3A38E704. Referência: Processo nº 23106.019572/2020-31 SEI nº 5048767 Endereço: Campus Universitário Darcy Ribeiro - Gleba A, , Brasília/DF, CEP 70910-900 Telefone: e Fax: @fax_unidade@ - http://www.unb.br

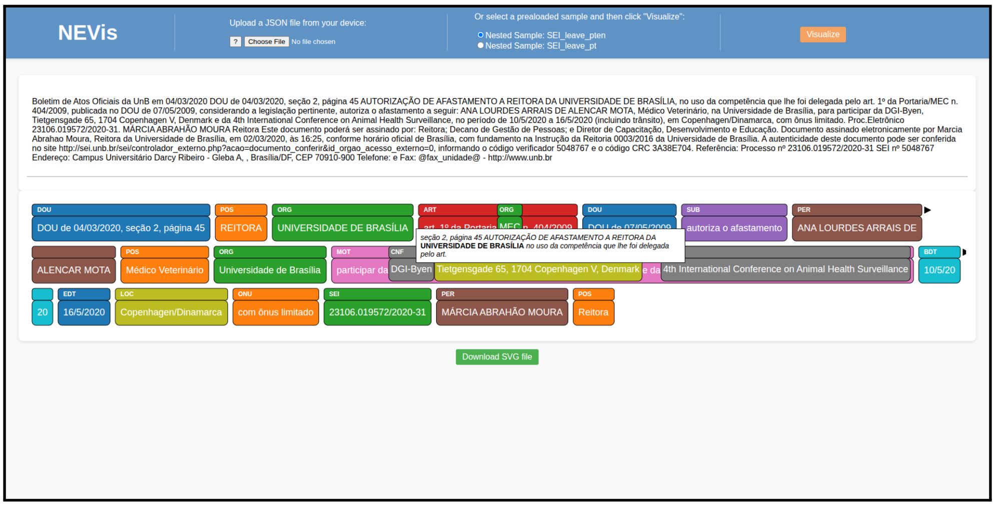
Figure (6.12) NEVis' main panel for nested NER with a sentence from SEI-leave-pten.


NEVis
Upload a JSON file from your device:
?
Choose File
No file chosen
Or select a prealoaded sample and then click "Visualize":
Nested Sample: SEI
Visualize
×

The JSON file must have the following format:

```
[
  [
    {
      "id": 71666,
      "annotations": [
        {
          "id": 29888,
          "completed_by": 1,
          "result": [
            {
              "value": {
                "start": 0,
                "end": 37,
                "text": "Pokémon: Arceus and the Jewel of Life",
                "labels": [
                  "MISC"
                ]
              },
              "id": "lIPVIA1HwQ",
              "from_name": "label",
              "to_name": "text",
              "type": "labels",
              "origin": "manual"
            },
            {
              "value": {
                "start": 70,
                "end": 77,
                "text": "Pokémon",
                "labels": [
                  "MISC"
                ]
              },
              "id": "Ea3KGnhzal",
              "from_name": "label",
              "to_name": "text",
              "type": "labels",
```

Figure (6.13) Help button ("?") with the input format.


NEVis
Upload a JSON file from your device:
? Choose File No file chosen
Or select a preloaded sample and then click "Visualize":
● Nested Sample: SEI_leave_pten
○ Nested Sample: SEI_leave_pt
Visualize

Boletim de Atos Oficiais da UnB em 04/03/2020 DOU de 04/03/2020, seção 2, página 45 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA, no uso da competência que lhe foi delegada pelo art. 1º da Portaria/MEC n. 404/2009, publicada no DOU de 07/05/2009, considerando a legislação pertinente, autoriza o afastamento a seguir: ANA LOURDES ARRAIS DE ALENCAR MOTA, Médico Veterinário, na Universidade de Brasília, para participar da DGI-Byen, Tietgensgade 65, 1704 Copenhagen V, Denmark e da 4th International Conference on Animal Health Surveillance, no período de 10/5/2020 a 16/5/2020 (incluindo trânsito), em Copenhagen/Dinamarca, com ônus limitado. Proc.Eletrônico 23106.019572/2020-31. MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação, Desenvolvimento e Educação. Documento assinado eletronicamente por Marcia Abrahao Moura, Reitora da Universidade de Brasília, em 02/03/2020, às 16:25, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília. A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 5048767 e o código CRC 3A38E704. Referência: Processo nº 23106.019572/2020-31 SEI nº 5048767 Endereço: Campus Universitário Darcy Ribeiro - Gleba A, , Brasília/DF, CEP 70910-900 Telefone: e Fax: @fax_unidade@ - http://www.unb.br

seção 2, página 45 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA **UNIVERSIDADE DE BRASÍLIA** *no uso da competência que lhe foi delegada pelo art.*

Figure (6.14) Tooltip interaction.

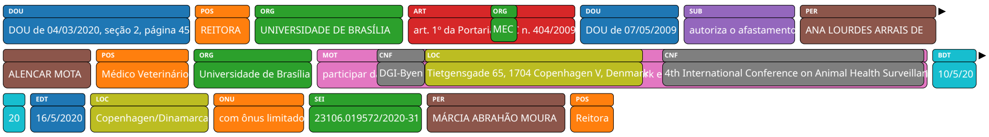


Figure (6.15) Downloaded SVG of the nested NER with a sentence from SEI-leave-pten.

The proposed Interactive Web Application described was developed using JavaScript[2], Python 3.11.11, Flask[3] framework and D3[4] library for visualization purposes.

NEVis was designed to increase the explainability of NLP models' predictions. In this context, it can be used to compare token classification predictions generated by distinct models, thereby extending the analysis beyond traditional quantitative metrics such as F1-score, accuracy, precision, and recall. In the next paragraphs, predictions generated by models CRF, BERTimbau, and LLaMA will be analyzed using NEVis. The texts were extracted from the publicly available SEI-leave-pten corpus; however, for presentation purposes, the professors' names were modified.

The Figures 6.17, 6.19, and 6.21 present real annotations and Figures 6.18, 6.20, and 6.22 show the predicted categories, respectively. In sentence 1, the CRF model correctly predicted all categories, while BERTimbau had one mistake: it does not separate the two positions "*PRESIDENTE DA FUNDAÇÃO*" (Foundation President) and "*REITORA*" (Rector); LLaMA misclassified Position (`POS`) as Organization (`ORG`), failed to recognize temporal expressions and did not associate numerical and date-related information with the Article (`ART`) and `DOU` categories. In sentence 2, BERTimbau misclassified some words that belong to an organization name as Motive (`MOT`), and LLaMA presented similar results comparing to sentence 1.

[2] `https://developer.mozilla.org/en-US/docs/Web/JavaScript`
[3] `https://flask.palletsprojects.com/en/stable/`
[4] `https://d3js.org/`

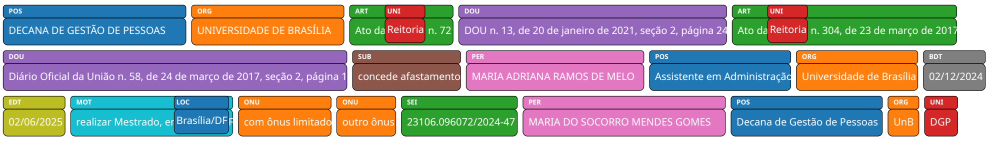
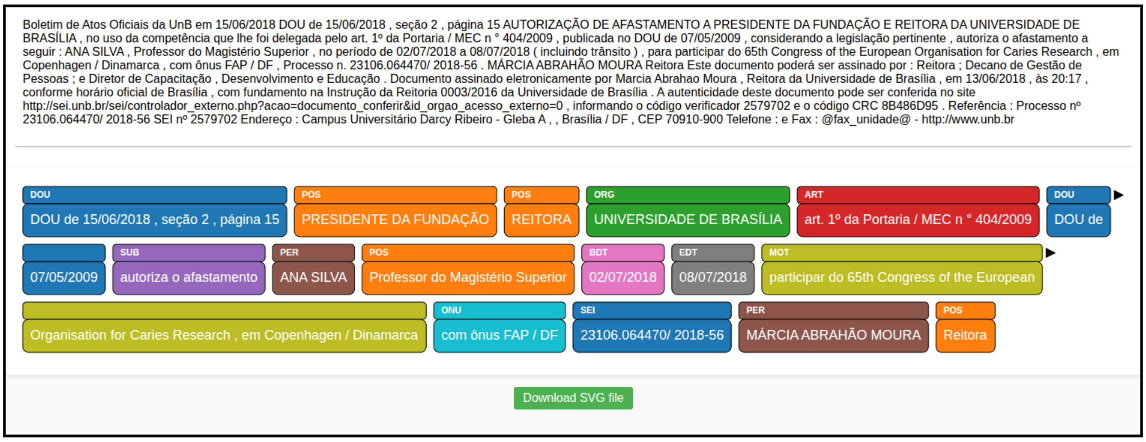

POS
DECANA DE GESTÃO DE PESSOAS
ORG
UNIVERSIDADE DE BRASÍLIA
ART
UNI
Ato da Reitoria n. 72
DOU
DOU n. 13, de 20 de janeiro de 2021, seção 2, página 24
ART
UNI
Ato da Reitoria n. 304, de 23 de março de 2017
DOU
Diário Oficial da União n. 58, de 24 de março de 2017, seção 2, página 1
SUB
concede afastamento
PER
MARIA ADRIANA RAMOS DE MELO
POS
Assistente em Administraçã
ORG
Universidade de Brasília
BDT
02/12/2024
EDT
02/06/2025
MOT
LOC
realizar Mestrado, er Brasília/DF
ONU
com ônus limitado
ONU
outro ônus
SEI
23106.096072/2024-47
PER
MARIA DO SOCORRO MENDES GOMES
POS
Decana de Gestão de Pessoas
ORG
UnB
UNI
DGP

Figure (6.16) Downloaded SVG of the nested NER with a sentence from SEI-leave-pt.

In sentence 3, CRF failed to identify the subject (SUB), whereas BERTimbau captured only the most semantically relevant part of the subject - the words "*Retificar*" (rectify) and "*afastamento*" (leave). Although all models detected the presence of leave-related motives in the document, this constitutes a misclassification because the text does not present a motive for leave but rather serves as a rectification of a leave document. This behavior suggests that the models have not learned the patterns associated with rectification leave documents.



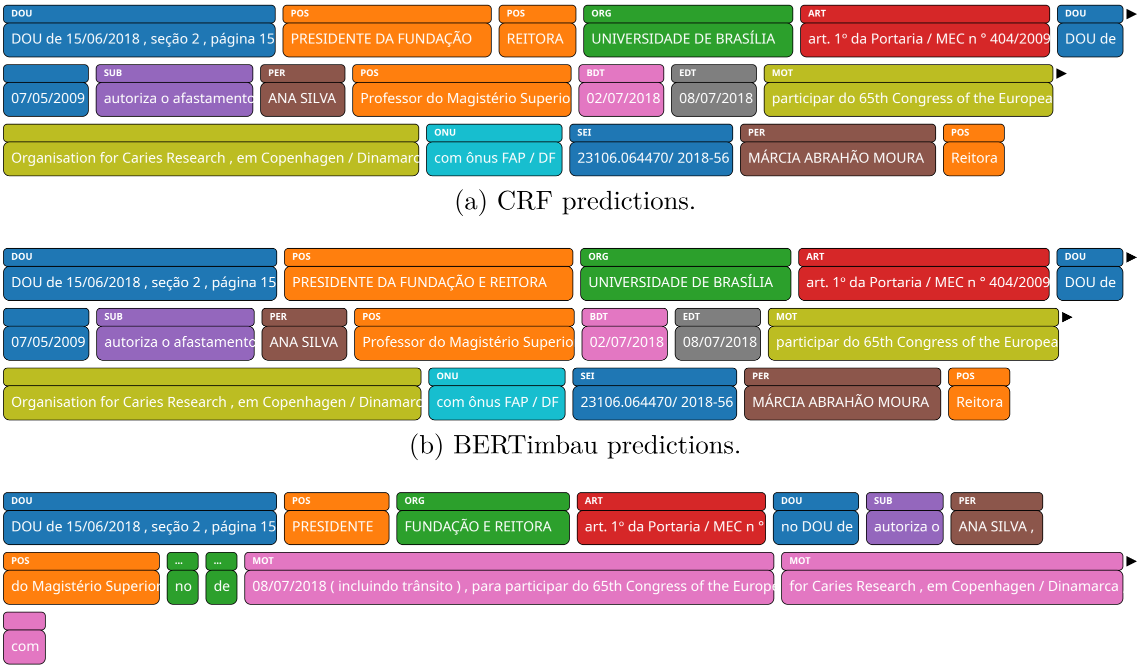
Boletim de Atos Oficiais da UnB em 15/06/2018 DOU de 15/06/2018 , seção 2 , página 15 AUTORIZAÇÃO DE AFASTAMENTO A PRESIDENTE DA FUNDAÇÃO E REITORA DA UNIVERSIDADE DE BRASÍLIA , no uso da competência que lhe foi delegada pelo art. 1º da Portaria / MEC n ° 404/2009 , publicada no DOU de 07/05/2009 , considerando a legislação pertinente , autoriza o afastamento a seguir : ANA SILVA , Professor do Magistério Superior , no período de 02/07/2018 a 08/07/2018 (incluindo trânsito) , para participar do 65th Congress of the European Organisation for Caries Research , em Copenhagen / Dinamarca , com ônus FAP / DF , Processo n. 23106.064470/ 2018-56 . MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por : Reitora ; Decano de Gestão de Pessoas ; e Diretor de Capacitação , Desenvolvimento e Educação . Documento assinado eletronicamente por Marcia Abrahao Moura , Reitora da Universidade de Brasília , em 13/06/2018 , às 20:17 , conforme horário oficial de Brasília , com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília . A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador 2579702 e o código CRC 8B486D95 . Referência : Processo nº 23106.064470/ 2018-56 SEI nº 2579702 Endereço : Campus Universitário Darcy Ribeiro - Gleba A , , Brasília / DF , CEP 70910-900 Telefone : e Fax : @fax_unidade@ - http://www.unb.br

Figure (6.17) Sentence 1 - extracted from SEI-leave-pten corpus.

6.3 Final Considerations

Experiments were conducted to analyze state-of-the-art NER model performance **(RQ2)** on WikiNER-ptmulti and SEI-leave-pten code-mixed corpora and the SEI documents in Portuguese language, showing better F1-score results with CRF and BERT models. However, it was noted that some multilingual entities have a lower F1-score compared to the Portuguese ones in both flat and nested versions of the analyzed SEI corpora.

As highlighted in Chapter 3, some studies compare classical and LLMs models, and the classical ones still have better results on NER tasks. However, with fast advancements in the NLP field, new LLMs were proposed and are worth trying for multilingual NER. The results obtained with fine-tuned LLMs with LoRA suggested that the model was not able to learn the specific patterns of the corpora, so classical NER methods have the best results in this case **(RQ3)**.

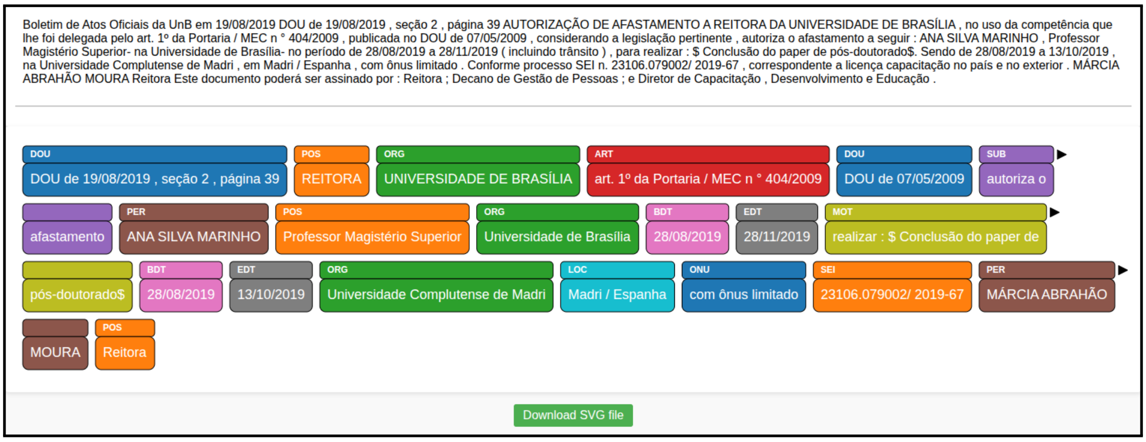

DOU
DOU de 15/06/2018 , seção 2 , página 15
POS
PRESIDENTE DA FUNDAÇÃO
POS
REITORA
ORG
UNIVERSIDADE DE BRASÍLIA
ART
art. 1º da Portaria / MEC n ° 404/2009
DOU
DOU de
07/05/2009
SUB
autoriza o afastamento
PER
ANA SILVA
POS
Professor do Magistério Superio
BDT
02/07/2018
EDT
08/07/2018
MOT
participar do 65th Congress of the Europea
Organisation for Caries Research , em Copenhagen / Dinamarc
ONU
com ônus FAP / DF
SEI
23106.064470/ 2018-56
PER
MÁRCIA ABRAHÃO MOURA
POS
Reitora
DOU
DOU de 15/06/2018 , seção 2 , página 15
POS
PRESIDENTE DA FUNDAÇÃO E REITORA
ORG
UNIVERSIDADE DE BRASÍLIA
ART
art. 1º da Portaria / MEC n ° 404/2009
DOU
DOU de
07/05/2009
SUB
autoriza o afastamento
PER
ANA SILVA
POS
Professor do Magistério Superio
BDT
02/07/2018
EDT
08/07/2018
MOT
participar do 65th Congress of the Europea
Organisation for Caries Research , em Copenhagen / Dinamarc
ONU
com ônus FAP / DF
SEI
23106.064470/ 2018-56
PER
MÁRCIA ABRAHÃO MOURA
POS
Reitora
DOU
DOU de 15/06/2018 , seção 2 , página 15
POS
PRESIDENTE
ORG
FUNDAÇÃO E REITORA
ART
art. 1º da Portaria / MEC n °
DOU
no DOU de
SUB
autoriza o
PER
ANA SILVA ,
POS
do Magistério Superior
...
no
...
de
MOT
08/07/2018 (incluindo trânsito) , para participar do 65th Congress of the Europ
MOT
for Caries Research , em Copenhagen / Dinamarca
com

(a) CRF predictions.

(b) BERTimbau predictions.

(c) LLaMA predictions.

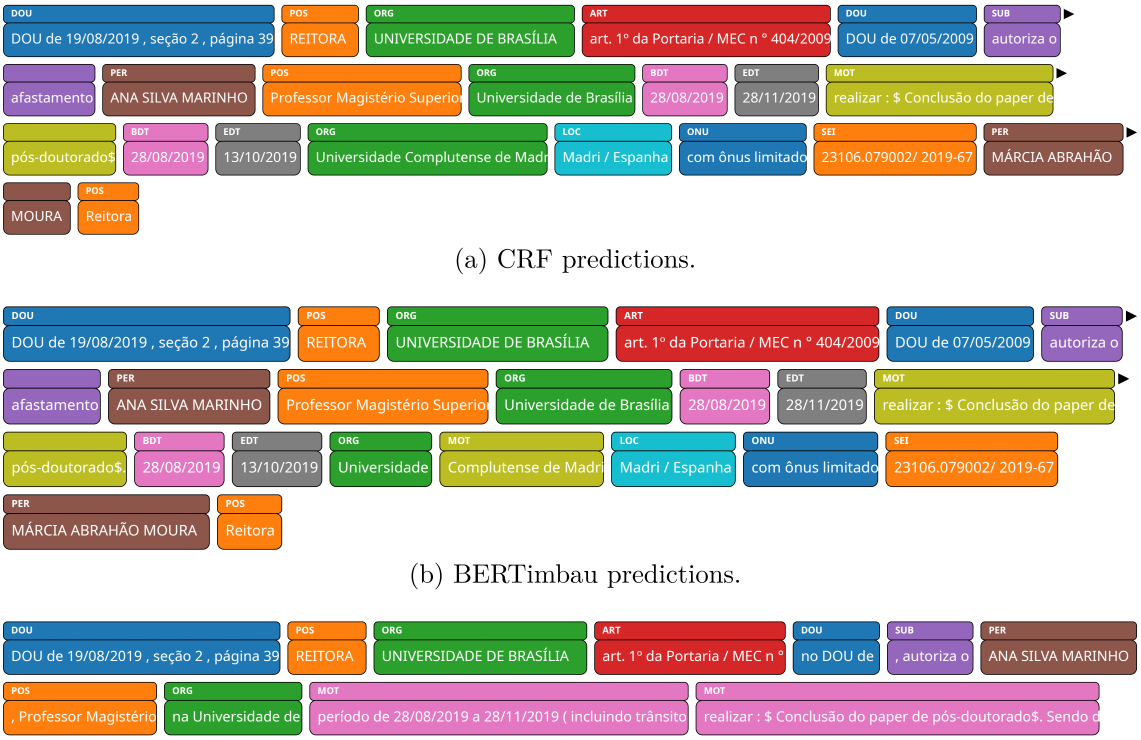
Figure (6.18) Sentence 1 with categories predicted by NER models.



Boletim de Atos Oficiais da UnB em 19/08/2019 DOU de 19/08/2019 , seção 2 , página 39 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA , no uso da competência que lhe foi delegada pelo art. 1º da Portaria / MEC n ° 404/2009 , publicada no DOU de 07/05/2009 , considerando a legislação pertinente , autoriza o afastamento a seguir : ANA SILVA MARINHO , Professor Magistério Superior- na Universidade de Brasília- no período de 28/08/2019 a 28/11/2019 (incluindo trânsito) , para realizar : $ Conclusão do paper de pós-doutorado$. Sendo de 28/08/2019 a 13/10/2019 , na Universidade Complutense de Madri , em Madri / Espanha , com ônus limitado . Conforme processo SEI n. 23106.079002/ 2019-67 , correspondente a licença capacitação no país e no exterior . MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por : Reitora ; Decano de Gestão de Pessoas ; e Diretor de Capacitação , Desenvolvimento e Educação .

Figure (6.19) Sentence 2 - extracted from SEI-leave-pten corpus.

The NEVis was developed to visualize flat and nested NER outputs based on an algorithm that creates highlighted blocks for each category, allowing the visualization of long documents with multiple overlapping and nested entities **(RQ4)**. This visual analysis supports the interpretation of F1-score results. In next versions, other visualization approaches should be implemented and tested by user studies. The hierarchical graph can be explored to visualize the hierarchy levels between the main entity and the nested ones while maintaining the visual relation between main entities. Multidimensional projections created with t-SNE and visualized by point-placement visualizations can also be explored to understand the similarity between entities by the distances in the visualization.



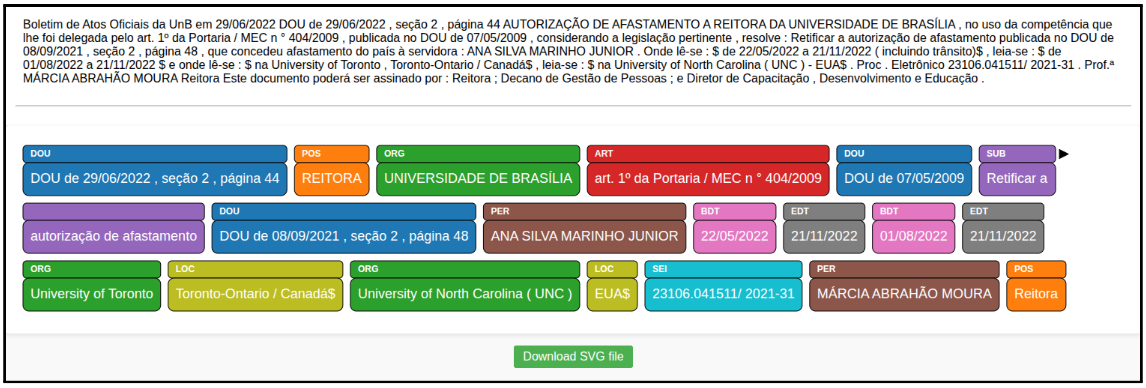
(a) CRF predictions.

(b) BERTimbau predictions.

DOU
DOU de 19/08/2019 , seção 2 , página 39
POS
REITORA
ORG
UNIVERSIDADE DE BRASÍLIA
ART
art. 1º da Portaria / MEC n °
DOU
no DOU de
SUB
, autoriza o
PER
ANA SILVA MARINHO
POS
, Professor Magistério
ORG
na Universidade de
MOT
período de 28/08/2019 a 28/11/2019 (incluindo trânsito
MOT
realizar : $ Conclusão do paper de pós-doutorado$. Sendo d

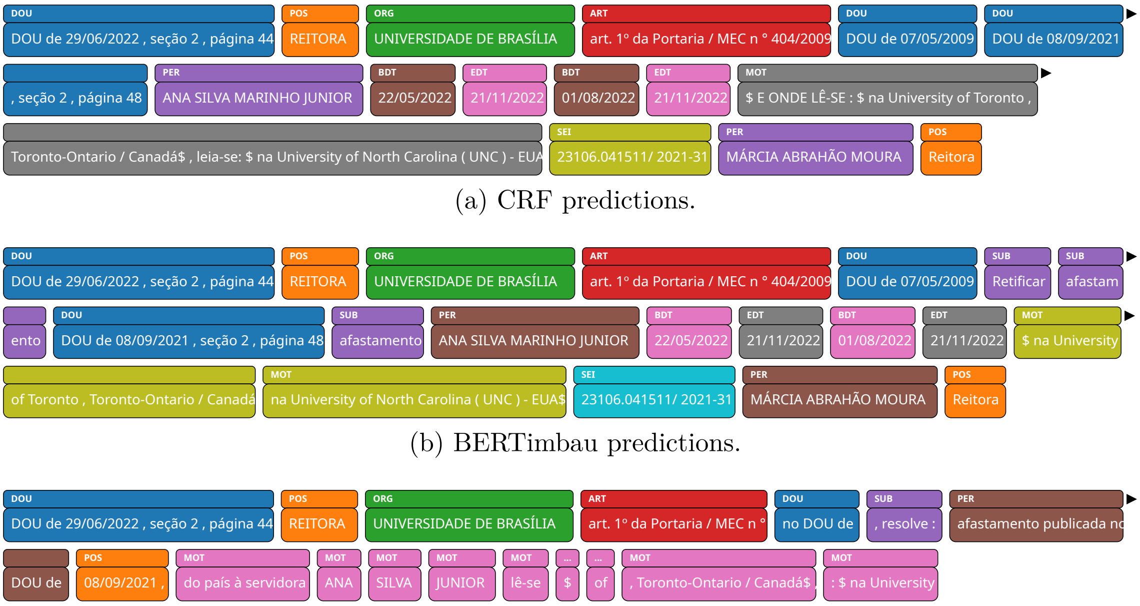
(c) LLaMA predictions.

Figure (6.20) Sentence 2 with categories predicted by NER models.



Boletim de Atos Oficiais da UnB em 29/06/2022 DOU de 29/06/2022 , seção 2 , página 44 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA , no uso da competência que lhe foi delegada pelo art. 1º da Portaria / MEC n ° 404/2009 , publicada no DOU de 07/05/2009 , considerando a legislação pertinente , resolve : Retificar a autorização de afastamento publicada no DOU de 08/09/2021 , seção 2 , página 48 , que concedeu afastamento do país à servidora : ANA SILVA MARINHO JUNIOR . Onde lê-se : $ de 22/05/2022 a 21/11/2022 (incluindo trânsito)$, leia-se : $ de 01/08/2022 a 21/11/2022 $ e onde lê-se : $ na University of Toronto , Toronto-Ontario / Canadá$, leia-se : $ na University of North Carolina (UNC) - EUA$. Proc . Eletrônico 23106.041511/ 2021-31 . Prof.ª MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por : Reitora ; Decano de Gestão de Pessoas ; e Diretor de Capacitação , Desenvolvimento e Educação .

Figure (6.21) Sentence 3 - extracted from SEI-leave-pten corpus.



(a) CRF predictions.

(b) BERTimbau predictions.

(c) LLaMA predictions.

Figure (6.22) Sentence 3 with categories predicted by NER models.

Chapter 7

Conclusion

This master's thesis investigated multilingual and nested NER in institutional and academic documents. In Brazilian public higher education institutions, SEI is widely used to manage administrative processes, encompassing a wide diversity of institutional documents such as administrative orders, circulars, resolutions, and official statements. NER techniques can support the structuring of such documents, enabling the development of efficient search and information retrieval strategies. As NER is a supervised learning task, the construction of a domain-specific corpus was required to train and fine-tune NER models. For this purpose, an annotation process was developed to construct corpora by using public SEI documents from the Universidade de Brasília and the manual labeling by trained volunteer annotators, which also considered nested entities that can be in, at least, in Portuguese or English.

The annotation process for constructing SEI corpora was validated using inter-annotator disagreement metrics (Cohen's Kappa and Krippendorff's Alpha) as well as text-level quality measures based on Levenshtein distance. To enable this, a greedy algorithm was developed to align and compare entity annotations made by two distinct annotators, allowing the identification of which entity spans were actually being compared. The results showed that Krippendorff's alpha and Cohen's kappa scores in the initial annotation round were higher than 0.8 in both considered corpora for flat NER, indicating the reliability of the entity labels and the accuracy of the tagged entity texts. However, for entities at deeper levels of nesting, the agreement metrics decreased, given the increased complexity for the annotators in identifying such entities.

Experiments were also conducted to compare the performance of state-of-the-art NER techniques under flat and nested NER settings. For flat NER, the models CRF, BERTimbau, mBERT, CNN-BiLSTM, BiLSTM, Fine-tuned LLaMA with LoRA, and Fine-tuned DeepSeek with LoRA were implemented, and their performances were evaluated using a K-Fold Cross-Validation strategy followed by statistical analysis with the Friedman test and the Nemenyi pairwise comparison. The results showed that, in the multilingual and Portuguese SEI corpus for leaves, CRF and BERT-based models achieved the highest macro F1-scores. CRF present the most dominant performances regarding the SEI corpora on other types containing only documents in Portuguese. In the multilingual WikiNER-ptmulti corpus, BERT-based models have the higher F1-scores, likely because the entity labels are similar to those found in the well-know CoNLL-2003 corpus, suggesting that BERT may have been contacted to similar patterns during pre-training.

Finally, NEVis, a technique for visualizing named entities was introduced aiming at supporting tasks related to analyzing and interpreting the outputs of NER models. NEVis can handle flat and nested entities, unlike existing approaches that address only flat entities. Moreover, it handles longer texts with multiple entities, which are displayed as blocks in a compact multi-row layout, effectively addressing spatial limitations not tackled by current approaches. The method also incorporates a tooltip interaction for showing surrounding text when the mouse hovers over the blocks, to provide contextual information and enhance the understanding of entities.

7.1 Contributions

This master's dissertation presented several contributions to multilingual NER, including Portuguese, mainly applied to institutional and academic documents. First, it introduced a Portuguese-English corpus composed of long documents and multiple entities related to the domain of academic texts, specifically designed to support nested named entity recognition. Second, it provided baseline experiments using these corpora with both state-of-the-art NER models in flat and nested approaches. Third, the work proposed a new visualization technique for named entities capable of handling long texts containing multiple and nested entities.

7.2 Future Work

As future work, additional approaches for nested NER found in the literature, particularly graph-based methods that can be implemented for a better representation of hierarchical structure. Another possibility involves expanding the corpora by including other types of institutional and academic documents available in SEI, which would increase the diversity of textual structures. Finally, with the continuous advancement of LLMs, experiments with alternative NER strategies, such as zero-shot models or models fine-tuned on different corpora and entity types, can be considered to assess their applicability in this domain.

In relation to NEVis, a next step involves conducting a formal user evaluation with participants of varying levels of NLP expertise to assess its effectiveness in accurately conveying entities and their associated contextual information.

7.3 Publications

This research resulted in direct and indirect publications during the master's degree. The paper "Could you tell me the process ID? Structuring Text Documents from the Brazilian Electronic Information System Using a Named Entity Recognition Approach" was accepted for direct publication in the research track of the XXI Simpósio Brasileiro de Sistemas de Informação (SBSI 2025).

Indirect contributions related to this research were also published, resulting in two conference papers. The paper "Natural Language Processing Approaches for Accrediting Students on Extracurricular Activities" was published at the XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024) (Cavalcante et al. (2024)) and the paper "Automatic Classification of Access Levels in Documents from the Brazilian Electronic Information System" was published in the XIII Latin American Symposium on Digital Government (LASDiGov 2025) an event from the XLV Congresso da Sociedade Brasileira de Computação (CSBC 2025) (Borges et al. (2025)).

7.4 Other Activities

The master's student also completed a teaching internship in the course Algorithms and Computer Programming at UnB, contributing to lectures and actively collaborating on the course's GitHub repository[1]. Furthermore, the preliminary stage of this research was presented in the "Seminários" course at UnB, an academic space where master's and doctoral students are invited to share their ongoing research with peers in the graduate program.

[1] https://github.com/viniciusrpb/cic0004_apc_engcomp

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D., Zhang, C., and Zhang, Y. (2024). Phi-4 technical report. 18, 29

Abilio, R., Coelho, G. P., and da Silva, A. E. A. (2024). Evaluating named entity recognition: A comparative analysis of mono-and multilingual transformer models on a novel brazilian corporate earnings call transcripts dataset. *Applied Soft Computing*, 166:112158. 2

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning*. Springer. x, 12, 13, 14, 16, 53

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 33

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., et al. (2022). Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *International Conference on Computational Processing of the Portuguese Language*, pages 3–14. Springer. 32, 66

Alencar, A. B., de Oliveira, M. C. F., and Paulovich, F. V. (2012). Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):476–492. 24

Atzberger, D., Cech, T., Trapp, M., Richter, R., Scheibel, W., Döllner, J., and Schreck, T. (2023). Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):902–912. 25

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*, abs/1409.0473. 16

Bakhashwain, N. and Sagheer, A. E. (2021). Online tuning of hyperparameters in deep lstm for time series applications. *International Journal of Intelligent Engineering and Systems*. 48

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171. 28

Beck, C., Booth, H., El-Assady, M., and Butt, M. (2020). Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In Dipper, S. and Zeldes, A., editors, *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics. 8

Benato, B. C., Gomes, J. F., Telea, A. C., and Falcão, A. X. (2021). Semi-automatic data annotation guided by feature space projection. *Pattern Recognition*, 109:107612. 25

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166. 28

Benikova, D., Biemann, C., and Reznicek, M. (2014). Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531. 36

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305. 49

Bezerra, L. F., Gonçalves, C. P., da Cunha, D. d. O., and Zouain, D. (2022). Os efeitos da capacitação do sistema eletrônico de informação em uma instituição pública federal. *Navus: Revista de Gestão e Tecnologia*, (12):27. 1

Borges, A., Marinho, M., Nogueira, R., Bordim, J., and Borges, V. (2025). Automatic classification of access levels in documents from the brazilian electronic information system. In *Anais do XIII Latin American Symposium on Digital Government*, pages 203–214, Porto Alegre, RS, Brasil. SBC. 88

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. e. a. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc. 18, 29

Carpendale, S. (2008). Evaluating information visualizations. In *Information visualization: Human-centered issues and perspectives*, pages 19–45. Springer. 64

Cavalcante, J., Marinho, M., and Borges, V. (2024). Natural language processing approaches for accrediting students on extracurricular activities. pages 1796–1809. 88

Chatzimparmpas, A., Martins, R. M., Jusufi, I., and Kerren, A. (2020). A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233. 3

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791. 3

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., and Gehrmann, S. e. a. (2024). Palm: scaling language modeling with pathways. *The Journal of Machine Learning Research*, 24(1). 18, 29

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. 8

Conneau, A., Baevski, A., Collobert, R., rahman Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *ArXiv*, abs/2006.13979. 28

Coria, J. M., Veron, M., Ghannay, S., Bernard, G., Bredin, H., Galibert, O., and Rosset, S. (2022). Analyzing BERT cross-lingual transfer capabilities in continual sequence labeling. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 15–25, Virtual. International Conference on Computational Linguistics. 34

Coscia, A. and Endert, A. (2024). Knowledgevis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6520–6532. 26

Darji, H., Mitrović, J., and Granitzer, M. (2023). German bert model for legal named entity recognition. *arXiv preprint arXiv:2303.05388*. 2

Das, S. S. S., Katiyar, A., Passonneau, R., and Zhang, R. (2022). CONTaiNER: Few-shot named entity recognition via contrastive learning. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics. 2

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. (2025a). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. 18, 29, 52

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. (2025b). Deepseek-v3 technical report. 18

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30. 22

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115. 19

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 28, 34, 51

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*. 2, 16, 17

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. 10

Durango, M. C., Torres-Silva, E. A., and Orozco-Duque, A. (2023). Named entity recognition in electronic health records: a methodological review. *Healthcare Informatics Research*, 29(4):286–300. 2

Egashira, K., Vero, M., Staab, R., He, J., and Vechev, M. (2024). Exploiting llm quantization. *Advances in Neural Information Processing Systems*, 37:41709–41732. 19

Fetahu, B., Chen, Z., Kar, S., Rokhlenko, O., and Malmasi, S. (2023a). Multiconer v2: a large multilingual dataset for fine-grained and noisy named entity recognition. 29, 41

Fetahu, B., Fang, A., Rokhlenko, O., and Malmasi, S. (2021). Gazetteer enhanced named entity recognition for code-mixed web queries. pages 1677–1681. 29, 30, 41

Fetahu, B., Kar, S., Chen, Z., Rokhlenko, O., and Malmasi, S. (2023b). SemEval-2023 task 2: Fine-grained multilingual named entity recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics. 35

Garside, R., Leech, G., and McEnery, T. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Pearson Education. Longman. 7, 8

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models. 18, 52

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 2

Gui, T., Ma, R., Zhang, Q., Zhao, L., Jiang, Y.-G., and Huang, X. (2019). Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4982–4988. International Joint Conferences on Artificial Intelligence Organization. 28

Guimarães, G. M., da Silva, F. X., Queiroz, A. L., Marcacini, R. M., Faleiros, T. P., Borges, V. R., and Garcia, L. P. (2024). Dodfminer: An automated tool for named entity recognition from official gazettes. *Neurocomputing*, 568:127064. 33

Hagiwara, M. (2021). *Real-World Natural Language Processing*. Hanning. 17, 21, 53

Hazem, A., Bouhandi, M., Boudin, F., and Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 648–662. 2

Herbold, S. (2020). Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173. 72

Hobson Lane, C. H. and Hapke, H. M. (2019). *Natural Language Processing*. Hanning. 10

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8). 2, 15, 28

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. 19, 51

Hu, Z. and Ma, X. (2023). A novel neural network model fusion approach for improving medical named entity recognition in online health expert question-answering services. *Expert Systems with Applications*, 223:119880. 2

Islam, M. R., Akter, S., Islam, L., Razzak, I., Wang, X., and Xu, G. (2024). Strategies for evaluating visual analytics systems: A systematic review and new perspectives. *Information Visualization*, 23(1):84–101. 64

Jehangir, B., Radhakrishnan, S., and Agarwal, R. (2023). A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017. 33

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. 18, 29

Ju, M., Miwa, M., and Ananiadou, S. (2018a). A neural layered model for nested named entity recognition. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics. 28

Ju, M., Miwa, M., and Ananiadou, S. (2018b). A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459. 36

Keraghel, I., Morbieu, S., and Nadif, M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. 27

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*. 49

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213. 19

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, 2:139. 8

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5:93–112. 8

Kucher, K., Paradis, C., and Kerren, A. (2018). Dosvis: Document stance visualization. In *International Conference on Information Visualization Theory and Applications (IVAPP), Funchal-Madeira, Portugal, 27-29 January, 2018*, volume 3, pages 168–175. SciTePress. 38

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 2, 20, 28

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. 3, 37

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536. 64

Lane, H., Hapke, H., and Howard, C. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications. x, 14

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., and Huang, J. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics. 3, 37

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551. 15, 28

Li, J., Sun, A., Han, J., and Li, C. (2022a). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50 – 70. 2, 6, 8, 9

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., and Li, J. (2020). A unified MRC framework for named entity recognition. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics. 37

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2022b). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019. 15

Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., Jiang, L., and Keim, D. A. (2018). Bridging text visualization and mining: A task-driven survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2482–2504. 24

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. 28

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692. 28

Lo, P.-S., Wu, J.-L., Deng, S.-T., and Wang, K.-C. (2022). Cnervis: a visual diagnosis tool for chinese named entity recognition. *Journal of Visualization*, 25(3):653–669. 38

Lovon-Melgarejo, J., Moreno, J. G., Besançon, R., Ferret, O., and Lechani, L. (2023). MEERQAT-IRIT at SemEval-2023 task 2: Leveraging contextualized tag descriptors for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 878–884, Toronto, Canada. Association for Computational Linguistics. 35

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421. Association for Computational Linguistics. 16

Luz de Araujo, P. H., de Campos, T., Oliveira, R., Stauffer, M., Couto, S., and De Souza Bermejo, P. (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*, pages 313–323. 20, 32

M. C. Guimarães, G., X. B. da Silva, F., A. B. Macedo, L., H. F. Lisboa, V., M. Marcacini, R., L. Queiroz, A., R. P. Borges, V., P. Faleiros, T., and P. F. Garcia, L. (2024). Legal document segmentation and labeling through named entity recognition approaches. *Journal of Information and Data Management*, 15(1):123–131. 28

Ma, J.-Y., Chen, B., Gu, J.-C., Ling, Z., Guo, W., Liu, Q., Chen, Z., and Liu, C. (2022). Wider & closer: Mixture of short-channel distillers for zero-shot cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5171–5183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 33

Malmasi, S., Fang, A., Fetahu, B., Kar, S., and Rokhlenko, O. (2022). SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics. 29, 34, 41

Mayhew, S., Blevins, T., Liu, S., Suppa, M., Gonen, H., Imperial, J. M., Karlsson, B., Lin, P., Ljubešić, N., Miranda, L. J., Plank, B., Riabi, A., and Pinter, Y. (2024). Universal NER: A gold-standard multilingual named entity recognition benchmark. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics. 2

Meyer, M., Sedlmair, M., and Munzner, T. (2015). The pair typology of visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2192–2201. 57

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *ArXiv*, abs/2402.06196. 18

Mo, Y., Yang, J., Liu, J., Wang, Q., Chen, R., Wang, J., and Li, Z. (2024). Mcl-ner: cross-lingual named entity recognition via multi-view contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18789–18797. 2

Moscato, V., Postiglione, M., and Sperlí, G. (2023). Few-shot named entity recognition: Definition, taxonomy and research directions. *ACM Transactions on Intelligent Systems and Technology*, 14(5). 2

Munzner, T. (2014). *Visualization analysis and design.* AK Peters Visualization Series. CRC Press. xi, 56, 57, 64

Nonato, L. G. and Aupetit, M. (2018). Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673. 25

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources. 29, 31, 33, 43

Nunes, R. O., Puttlitz, L. M., Boll, A. O., Spritzer, A., Freitas, C. M. D. S., Balreira, D. G., and Tavares, A. R. (2025). An ensemble of llms finetuned with lora for ner in portuguese legal documents. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 127–140, Cham. Springer Nature Switzerland. 38

Nunes, R. O., Santos, J., Spritzer, A., Balreira, D. G., Freitas, C. M. D. S., Olival, F., Cameron, H. F., and Vieira, R. (2024). Assessing european and brazilian portuguese llms for ner in specialised domains. In *Brazilian Conference on Intelligent Systems*, pages 215–230. Springer. 2

Oliveira, V., Nogueira, G., Faleiros, T., and Marcacini, R. (2024). Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, pages 1–21. 9

OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774. 29

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., and et al., S. A. (2024). Gpt-4 technical report. 18

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics. 29, 41

Parsaeimehr, E., Fartash, M., and Akbari Torkestani, D. J. (2023). Improving feature extraction using a hybrid of cnn and lstm for entity identification. *Neural Processing Letters*, 55:1–16. 28

Paulovich, F. V. and Minghim, R. (2008). Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1229–1236. 25

Rodrigues, F. B., Giozza, W. F., de Oliveira Albuquerque, R., and Villalba, L. J. G. (2022). Natural language processing applied to forensics information extraction with transformers and graph visualization. *IEEE Transactions on Computational Social Systems*. 39

Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., and Agirre, E. (2024). Gollie: Annotation guidelines improve zero-shot information-extraction. 33

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. 17

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: An advanced NER evaluation contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA). 32

Santos, J., Cameron, H. F., Olival, F., Farrica, F., and Vieira, R. (2024). Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 117–126. 3

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al. (2024). The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*. 19

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., and Anandkumar, A. (2017). Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256. Association for Computational Linguistics. 28

Sherkat, E., Milios, E. E., and Minghim, R. (2019). A visual analytics approach for interactive document clustering. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(1):1–33. 25

Silva, P., Franco, A., Santos, T., Brito, M., and Pereira, D. (2023). Cachacaner: a dataset for named entity recognition in texts about the cachaça beverage. *Lang. Resour. Eval.*, 58(4):1315–1333. 32, 66

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. 17, 51

Sui, Y., Bu, F., Hu, Y., Zhang, L., and Yan, W. (2022). Trigger-gnn: a trigger-based graph neural network for nested named entity recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE. 37

Sultanum, N., Singh, D., Brudno, M., and Chevalier, F. (2019). Doccurate: A curation-based approach for clinical text visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):142–151. 38

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373. 20, 28

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. (2021). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics. 33

Tedeschi, S. and Navigli, R. (2022). MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics. 29, 41

Tian, Z., Zhai, X., van Driel, D., van Steenpaal, G., Espadoto, M., and Telea, A. (2021). Using multiple attribute-based explanations of multidimensional projections to explore high-dimensional data. *Computers & Graphics*, 98:93–104. 3

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 29, 41

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. 29, 41

Tkachenko, M., Malyuk, M., Holmanyuk, A., and Liubimov, N. (2020-2025). Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio. 33

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971. 18, 29

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models. 18

van den Elzen, S., Andrienko, G., Andrienko, N., Fisher, B. D., Martins, R. M., Peltonen, J., Telea, A. C., and Verleysen, M. (2023). The flow of trust: A visualization framework to externalize, explore, and explain trust in ml applications. *IEEE computer graphics and applications*, 43(2):78–88. 3

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11). 25

Vasile Păiș, Maria Mitrofan, C. L. G. A. I. C. G. V. S. C. and Onuț, A. (2023). Legalnero: A linked corpus for named entity recognition in the romanian legal domain. *Miscellaneous*, 15(3):831–844. 33

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. 2, 16, 28

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). Ace 2005 multilingual training corpus. *(No Title)*. 36

Wan, J., Ru, D., Zhang, W., and Yu, Y. (2022). Nested named entity recognition with span-level graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 892–903. 36

Wang, D., Feng, X., Liu, Z., and Wang, C. (2024). 2m-ner: contrastive learning for multilingual and multimodal ner with language and modal fusion. *Applied Intelligence*, 54:1–17. 35

Wang, S. Y., Huang, J., Hwang, H., Hu, W., Tao, S., and Hernandez-Boussard, T. (2022a). Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *International Journal of Medical Informatics*, 167:104864. 2

Wang, Y., Shindo, H., Matsumoto, Y., and Watanabe, T. (2022b). Nested named entity recognition via explicitly excluding the influence of the best path. *Journal of Natural Language Processing*, 29(1):23–52. 6

Wang, Y., Tong, H., Zhu, Z., and Li, Y. (2022c). Nested named entity recognition: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29. 35

Ward, M., Grinstein, G., and Keim, D. (2015). *Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition.* 360 Degree Business. CRC Press. xi, 56, 57

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., and Wang, H. (2023). Empirical study of zero-shot ner with chatgpt. pages 7935–7956. 37

Yang, K., Yang, Z., Zhao, S., Yang, Z., Zhang, S., and Chen, H. (2024). Uncertainty-aware contrastive learning for semi-supervised named entity recognition. *Knowledge-Based Systems*, 296:111762. 2

Yu, J., Bohnet, B., and Poesio, M. (2020). Named entity recognition as dependency parsing. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics. 37

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21. 19

Zanuz, L. and Rigo, S. J. (2022). Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 219–229. Springer. 2

Zaratiana, U., Tomeh, N., Holat, P., and Charnois, T. (2023). Gliner: Generalist model for named entity recognition using bidirectional transformer. 37

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. 26

Zhang, J., Zhang, Y., Chen, Y., and Xu, J. (2023a). Structure and label constrained data augmentation for cross-domain few-shot NER. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 518–530, Singapore. Association for Computational Linguistics. 33

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. (2023b). Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*. 19

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. (2023c). Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*. 19

Zhang, X., Li, S., Hauer, B., Shi, N., and Kondrak, G. (2023d). Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics. 3, 37

Zhang, Y., Liu, J., Zhong, X., and Wu, L. (2025). Seclmner: A framework for enhanced named entity recognition in multi-source cybersecurity data using large language models. *Expert Systems with Applications*, 271:126651. 3

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., and et al., Z. D. (2024). A survey of large language models. 18, 28

Zhou, L., Li, J., Gu, Z., Qiu, J., Gupta, B. B., and Tian, Z. (2022). Panner: Pos-aware nested named entity recognition through heterogeneous graph neural network. *IEEE Transactions on Computational Social Systems*, 11(4):4718–4726. 36

Zhou, W., Zhang, S., Gu, Y., Chen, M., and Poon, H. (2024). UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*. 38

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In Li, S., Sun, M., Liu, Y., Wu, H., Liu, K., Che, W., He, S., and Rao, G., editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China. 17

Annex I

Annotation Guideline

SEI Corpus - **Leave pt** and **Leave pt/en** Annotation Guidelines

Tag’s dictionary

Leave

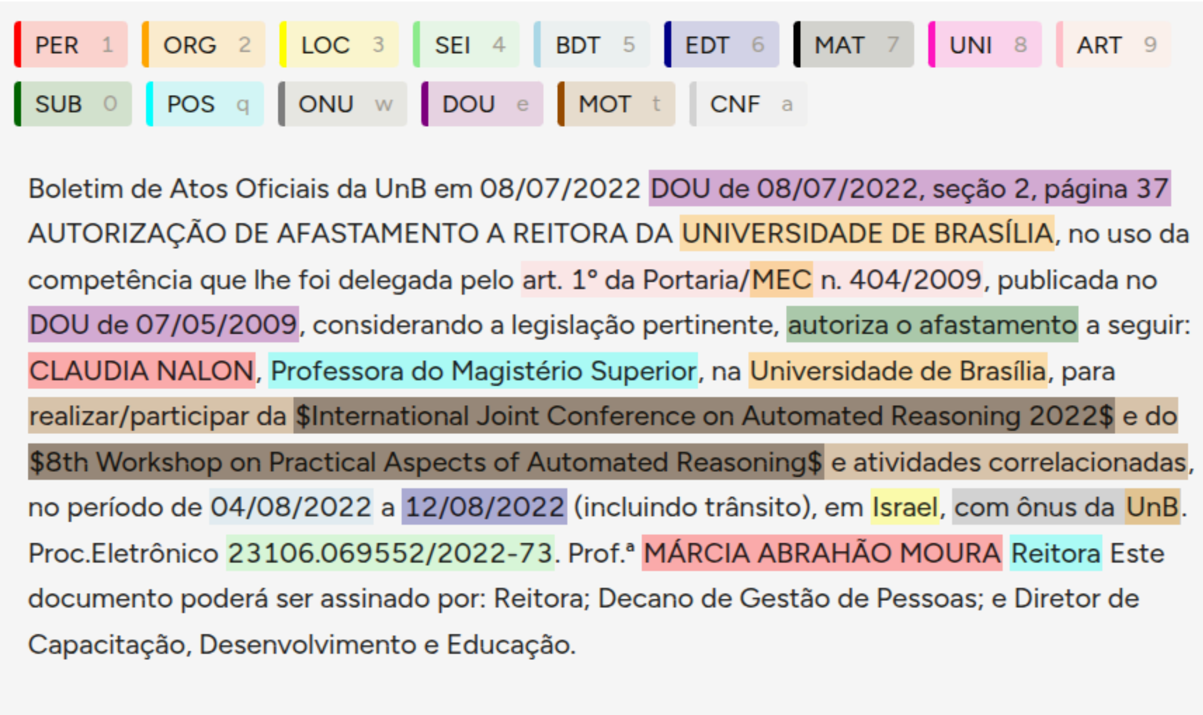
Entity	Tag Name	Description
SEI's Process Number	SEI	The number of SEI process
Location	LOC	Location
Organization	ORG	Organization's name or acronym
Person's Name	PER	Person's full name
Begin Date	BDT	Initial date of leave
End Date	EDT	Final date of leave
Enrollment number	REG	Enrollment number that identifies a civil staff
Subject	SUB	Subject of the publication
University related	UNI	Name and Acronyms of Departments, Secretariats, Divisions, Programs
Cost	ONU	Define if there are costs to the university or organization
Position	POS	Full role of a person
DOU information	DOU	Brazilian Official Gazette information
Justification	MOT	Leave justification
Events	EVE	Name of events, including conferences, workshops, meetings

Regulation or Article number	ART	Regulation or Article number

Annotation instructions and examples

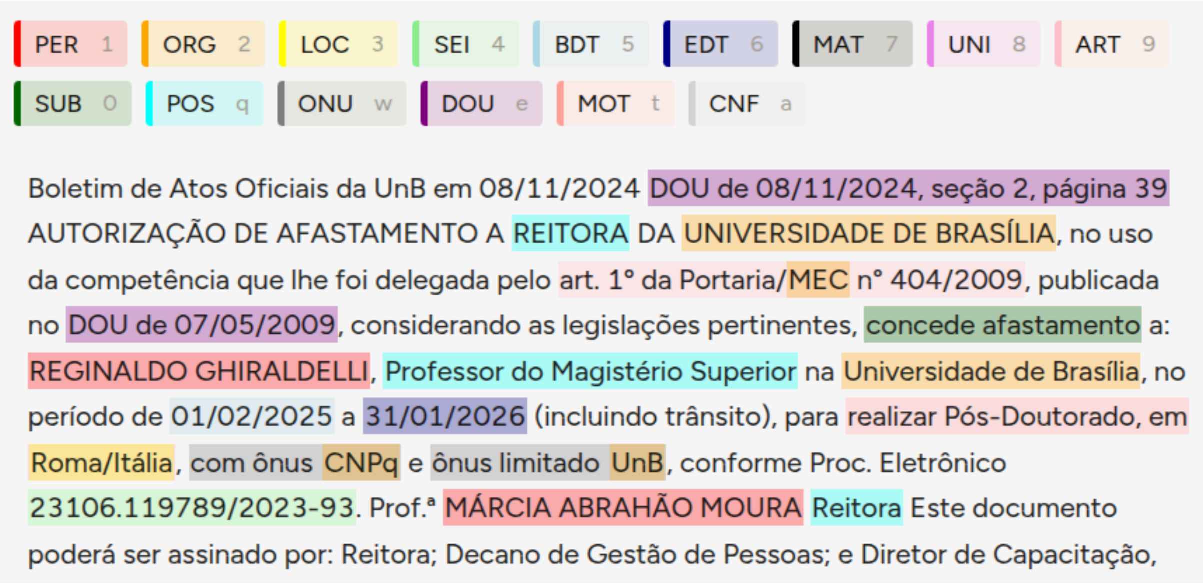
The annotation starts after the date of the first sentence and finish before “Documento assinado eletronicamente por…” or "Este documento poderá ser assinado por...".

LabelStudio’s annotation example of Leave pt/en:



Boletim de Atos Oficiais da UnB em 08/07/2022 DOU de 08/07/2022, seção 2, página 37 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA, no uso da competência que lhe foi delegada pelo art. 1° da Portaria/MEC n. 404/2009, publicada no DOU de 07/05/2009, considerando a legislação pertinente, autoriza o afastamento a seguir: CLAUDIA NALON, Professora do Magistério Superior, na Universidade de Brasília, para realizar/participar da $International Joint Conference on Automated Reasoning 2022$ e do $8th Workshop on Practical Aspects of Automated Reasoning$ e atividades correlacionadas, no período de 04/08/2022 a 12/08/2022 (incluindo trânsito), em Israel, com ônus da UnB. Proc.Eletrônico 23106.069552/2022-73. Prof.ª MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação, Desenvolvimento e Educação.

LabelStudio’s annotation example of Leave pt:



Boletim de Atos Oficiais da UnB em 08/11/2024 DOU de 08/11/2024, seção 2, página 39 AUTORIZAÇÃO DE AFASTAMENTO A REITORA DA UNIVERSIDADE DE BRASÍLIA, no uso da competência que lhe foi delegada pelo art. 1° da Portaria/MEC n° 404/2009, publicada no DOU de 07/05/2009, considerando as legislações pertinentes, concede afastamento a: REGINALDO GHIRALDELLI, Professor do Magistério Superior na Universidade de Brasília, no período de 01/02/2025 a 31/01/2026 (incluindo trânsito), para realizar Pós-Doutorado, em Roma/Itália, com ônus CNPq e ônus limitado UnB, conforme Proc. Eletrônico 23106.119789/2023-93. Prof.ª MÁRCIA ABRAHÃO MOURA Reitora Este documento poderá ser assinado por: Reitora; Decano de Gestão de Pessoas; e Diretor de Capacitação,

ORG, POS, UNI

1) **Diretora de Gestão de Pessoas da Universidade de Brasília**

POS = Diretora de Gestão de Pessoas
ORG = Universidade de Brasília

- The positions “Diretora”, “Decana” should be labeled with their corresponding areas.

2) **Decana de Gestão de Pessoas UnB/DGP**

POS = Decana de Gestão de Pessoas
ORG = UnB
UNI = DGP

3) **Reitora da Universidade de Brasília**

POS = Reitora
ORG = Universidade de Brasília

- The position “Reitora” should be labeled **without** “Universidade de Brasília” because there is only one Dean in the University of Brasília.

4) **Prof. Enrique Vice-Reitor e Presidente do CEPE**

PER = Enrique
POS = Vice-Reitor
POS = Presidente do CEPE
UNI = CEPE

5) **Maria Lucia, diretora de capacitação, desenvolvimento e educação UnB/DGP/DCADE**

PER = Maria Lucia
POS = diretora de capacitação, desenvolvimento e educação
ORG = UnB
UNI = DGP/DCADE

6) **autoriza afastamento a seguir: Prof. Dr. Enrique, Professor(a) do Magistério Superior, para…**

SUB = autoriza o afastamento
PER = Enrique
POS = Professor(a) do Magistério Superior

- The term “Prof. Dr.” before the name of a person cannot be labeled as POS.

7) **Decanato de Pesquisa e Inovação (DPI)**

UNI = Decanato de Pesquisa e Inovação (DPI)

8) **Reitoria da Universidade de Brasília**

UNI = Reitoria
ORG = Universidade de Brasília

9) **Iniciativas do Campus UnB Gama**

UNI = Campus UnB Gama

- In this context, the term “UnB” is part of the campus name “UnB Gama”, then it **must not** be labeled as ORG.

10) Hospital Universitário da Universidade de Brasília (UnB)

UNI = Hospital Universitário
ORG = Universidade de Brasília (UnB)

11) Laboratório de Registro de Imagens e Interações Sociais-IRIS/DAN/UnB

UNI = Laboratório de Registro de Imagens e Interações Sociais-IRIS/DAN/UnB
ORG = UnB

REG

12) matrícula nº 1047698

REG = 1047698

13) matrícula SIAPE nº 112270-6

REG = 112270-6

- Only the number should be annotated.

SEI

14) Proc. Eletrônico 23106.101108/2024-11

SEI = 23106.101108/2024-11

15) …conforme processo SEI 23106.101108/2024-11

SEI = 23106.101108/2024-11

- Only the number should be annotated.

SUB

16) …considerando as legislações pertinentes, concede afastamento a Enrique, assistente de administração, na Universidade de Brasília, no período de…

SUB = Concede afastamento
PER = Enrique
POS = assistente de administração
ORG = Universidade de Brasília

17) Concede Licença Capacitação ao(a) servidor(a): Enrique, Professor(a) do Magistério Superior na Universidade de Brasília

SUB = Concede Licença Capacitação
PER = Enrique
POS = Professor(a) do Magistério Superior
ORG = Universidade de Brasília

18) resolve: Tornar sem efeito a publicação no DOU setembro/2024, que concedeu afastamento do país ao servidor Enrique

SUB = Tornar sem efeito
DOU = DOU setembro/2024

PER = Enrique

19) AUTORIZAÇÃO DE AFASTAMENTO RETIFICAÇÃO No DOU de 18/09/2024, seção 2, página 34, na publicação referente à autorização de afastamento da servidora Aline, onde se lê…

SUB = RETIFICAÇÃO
DOU = DOU de 18/09/2024, seção 2, página 34
PER = Aline

ART

20) Por meio do Ato da Reitoria n. 304, de 23 de março de 2017

ART = Ato da Reitoria n. 304, de 23 de março de 2017
UNI = Reitoria

21) Art. 2º e 4º dizem que

ART = Art. 2º e 4º

22) Art. 2º do inciso X da Constituição Federal

ART = Art. 2º do inciso X da Constituição Federal

23) Refutada pelo art. 1º da Portaria/MEC nº 404/2009

ART = art. 1º da Portaria/MEC nº 404/2009
ORG = MEC

24) INSTRUÇÃO NORMATIVA SGD/ME N 96

ART = INSTRUÇÃO NORMATIVA SGD/ME N 96
UNI = SGD
ORG = ME

- ME is an abbreviation for “Ministério da Economia”.

DOU

25) DOU n. 236, de 16 de dezembro de 2022, da seção X, página Y

DOU = DOU n. 236, de 16 de dezembro de 2022, da seção X, página Y

26) Diário Oficial da União n. 236, de 16 de dezembro de 2022, da seção X, página Y

DOU = Diário Oficial da União n. 236, de 16 de dezembro de 2022, da seção X, página Y

27) publicada no DOU de 7/5/2009

DOU = DOU de 7/5/2009

ORG

28) Universidade de Brasília (UnB)

ORG = Universidade de Brasília (UnB)

29) Universidade de Brasília - UnB

ORG = Universidade de Brasília - UnB

30) Realizado na UnB

ORG = UnB

31) Realizado na Fundação Universidade de Brasília

ORG = Fundação Universidade de Brasília

- In this case "Universidade de Brasília" **must not** be labeled as ORG, because it is part of the organization's name "Fundação Universidade de Brasília".

LOC

32) Recife/PE

LOC = Recife/PE

33) Recife-PE

LOC = Recife-PE

34) Recife, em Pernambuco, no Brasil

LOC = Recife, em Pernambuco, no Brasil

35) Coimbra e Lisboa/Portugal

LOC = Coimbra e Lisboa/Portugal

MOT

36) para realizar Pós-Doutorado em Roma/Itália

MOT = realizar Pós-Doutorado em Roma/Itália

LOC = Roma/Itália

- The MOT also **includes the location.**
- The MOT **does not include the preposition "para".**

37) para realizar escrita de tese de doutorado no programa de African and African Diaspora Studies

MOT = realizar escrita de tese de doutorado no programa de African and African Diaspora Studies

UNI = African and African Diaspora Studies

- The MOT includes the thesis program, if it exists.

38) para realizar reuniões de trabalho e participar no $Biostimulants World Congress$

MOT = realizar reuniões de trabalho

MOT = participar no $Biostimulants World Congress$

EVE = $Biostimulants World Congress$

- In this case, there are two verbs ("realizar" and "participar"), so there are two MOTs and they should be annotated separately.

39) **para realizar/participar da $International Joint Conference on Automated Reasoning 2022$ e do $8th Workshop on Practical Aspects of Automated Reasoning$ e atividades correlacionadas**

MOT = realizar/participar da $International Joint Conference on Automated Reasoning 2022$ e do $8th Workshop on Practical Aspects of Automated Reasoning$ e atividades correlacionadas
EVE = $International Joint Conference on Automated Reasoning 2022$
EVE = $8th Workshop on Practical Aspects of Automated Reasoning$

- The **“$” must be annotated** alongside the name of the event as EVE.
- In this case, there is one verb “realizar/participar”, so there is just one MOT.

40) **para concluir pesquisa/paper de pós-Doutorado**

MOT = concluir pesquisa/paper de pós-Doutorado

41) **para realizar ação de capacitação conforme previsto no Decreto 9.991/2019, art. 25, I; no País**

MOT = para realizar ação de capacitação conforme previsto no Decreto 9.991/2019, art. 25, I; no País
ART = Decreto 9.991/2019, art. 25, I

42) **para participar de palestra na Clark University, em Worcester, Massachusetts, no Estados Unidos**

MOT = para participar de palestra na Clark University, em Worcester, Massachusetts, no Estados Unidos
ORG = Clark University
LOC = Worcester, Massachusetts, no Estados Unidos

43) **para participar do Symposium Sessions 2024 - MRS Fall Meeting e Exhibit, em Boston, Massachusetts, Estados Unidos da América**

MOT = para participar do Symposium Sessions 2024 - MRS Fall Meeting e Exhibit, em Boston, Massachusetts, Estados Unidos da América
EVE = Symposium Sessions 2024 - MRS Fall Meeting e Exhibit
LOC = Boston, Massachusetts, Estados Unidos da América

44) **realizar visita para estabelecer colaboração internacional no Department of Experimental Biomedicine and Clinical Neuroscience of University of Palermo**

MOT = realizar visita para estabelecer colaboração internacional no Department of Experimental Biomedicine and Clinical Neuroscience of University of Palermo
UNI = Department of Experimental Biomedicine and Clinical Neuroscience
ORG = University of Palermo

45) **para realizar visita técnica à Faculdade de Ciências do Desporto e Educação Física da Universidade de Coimbra e participar da Conferência Anual da EAS Annual Conference 2021 no Instituto Universitário de Lisboa, no período de…**

MOT = para realizar visita técnica à Faculdade de Ciências do Desporto e Educação Física da Universidade de Coimbra
UNI = Faculdade de Ciências do Desporto e Educação Física
ORG = Universidade de Coimbra
MOT = participar da Conferência Anual da EAS Annual Conference 2021 no Instituto Universitário de Lisboa
EVE = Conferência Anual da EAS Annual Conference 2021
ORG = Instituto Universitário de Lisboa

- Note that there are two reasons (motivos) in this sentence.

ONU

46) com ônus CNPq e ônus limitado para UnB

ONU = com ônus CNPq
ORG = CNPq
ONU = ônus limitado para UnB
ORG = UnB

BDT, EDT

47) … solicita afastamento para participar de conferência no período de 07/03/2024 a 14/03/2024 (incluindo trânsito).

MOT = participar de conferência
BDT = 07/03/2024
EDT = 14/03/2024

48) … no período de 20 a 24 de novembro de 2024

BDT = 20
EDT = 24 de novembro de 2024

Annex II

Hierarchical Dictionary Sample

```
{
  "name": "Root",
  "text": "The president of University of California, Los Angeles spoke at the summit at Silicon Valley on September 23, 2023.",
  "children": [
    {
      "name": "president",
      "start": 4,
      "end": 13,
      "label": "MISC",
      "gap_left": 4,
      "gap_right": 4,
      "children": []
    },
    {
      "name": "University of California, Los Angeles",
      "start": 17,
      "end": 57,
      "label": "ORG",
      "gap_left": 4,
      "gap_right": 15,
      "children": [
        {
          "name": "Los Angeles",
          "start": 45,
          "end": 57,
          "label": "LOC",
          "gap_left": 28,
          "gap_right": 0,
          "children": []
        }
      ]
    },
    {
      "name": "Silicon Valley",
      "start": 72,
      "end": 86,
      "label": "LOC",
      "gap_left": 15,
      "gap_right": 5,
      "children": []
    },
    {
      "name": "September 23, 2023",
      "start": 91,
      "end": 110,
      "label": "DATE",
      "gap_left": 5,
      "gap_right": 0,
      "children": []
    }
  ]
}
```

Listing (II.1) Hierarchial dictionary of the input sample.