

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE INFORMÁTICA  
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GIOVANI GHENO BOMBARDIERI

**Automatic Generation of a Named Entity  
Set for Analysis of Political Speeches**

Work presented in partial fulfillment of the  
requirements for the degree of Bachelor in  
Computer Science

Advisor: Prof. Dr. Dennis Giovanni Balreira  
Co-advisor: BSc. Rafael Oleques Nunes

Porto Alegre  
August 2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Carlos André Bulhões

Vice-Reitora: Prof<sup>a</sup>. Patricia Pranke

Pró-Reitora de Graduação: Prof<sup>a</sup>. Cíntia Inês Boll

Diretora do Instituto de Informática: Prof<sup>a</sup>. Carla Maria Dal Sasso Freitas

Coordenador do Curso de Ciência de Computação: Prof. Marcelo Walter

Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

*“Taking a new step, uttering a new word,  
is what people fear most.”*  
— FYODOR DOSTOEVSKY

## **AGRADECIMENTOS**

Antes de tudo, agradeço imensamente ao Professor Dennis e ao Rafael pela orientação. Sempre disponíveis, dispostos, prestativos e compreensíveis com minhas limitações. Me apresentaram uma nova área de estudo e me guiaram desde a concepção da ideia do trabalho até o último ponto final. Em cada obstáculo atingido, me mostraram os métodos de contorná-lo. Sem eles, este trabalho não existiria.

Agradeço à UFRGS, e em especial o Instituto de Informática, pela excelência no ensino e pelas oportunidades que me concederam ao longo desta imensa jornada até aqui. Menciono também o Professor Lucas Schnnor e todo o PCAD, já que alguns experimentos deste trabalho utilizaram os recursos da infraestrutura PCAD, <http://gppd-hpc.inf.ufrgs.br>, no INF/UFRGS.

Dedico este trabalho à minha mãe, uma mulher como nenhuma outra que jamais desistiu de mim. Que a conclusão desta etapa nos abra uma porta para uma nova, melhor fase de nossas vidas. E ao meu pai, cuja influência, mesmo não estando mais aqui, ainda pesa em todas as minhas decisões. Aos dois, muito obrigado por tudo.

## ABSTRACT

Extracting meaningful information from unstructured text is essential in fields like political discourse, where language nuances significantly impact public opinion and policy development. In this work, we develop and evaluate a Named Entity Recognition (NER) model tailored to analyze Brazilian political discourse. Leveraging a comprehensive corpus of political speeches from the Brazilian Chamber of Deputies, we automatically generated a set of NER categories using a hybrid methodology that combines distant supervision techniques with a domain-specific Thesaurus. Next, we annotate the resulting corpus to train BERTimbau, a Bidirectional Encoder Representations from Transformers (BERT) based model optimized for Brazilian Portuguese. Our approach also included implementing a proof-of-concept web tool to visualize and interact with the extracted entities, offering valuable insights into political language trends. The results demonstrate that while the model does not achieve state-of-the-art performance, it effectively recognizes key entities, making it a valuable tool for specific applications in political discourse analysis. This work highlights the potential of automated NER systems in understudied languages and domains, providing a foundation for future research and improvements.

**Keywords:** Named Entity Recognition (NER). Political Discourse. Corpus Annotation. Language Models.

# Geração Automática de um Conjunto de Entidades Nomeadas para Análise de Discursos Políticos

## RESUMO

Extrair informações significativas de textos não estruturados é fundamental em áreas como o discurso político, onde as nuances da linguagem impactam significativamente a opinião pública e o desenvolvimento de políticas. Neste trabalho, desenvolvemos e avaliamos um modelo de Reconhecimento de Entidades Nomeadas (NER) especificamente adaptado para a análise do discurso político brasileiro. Utilizando um corpus abrangente de discursos políticos da Câmara dos Deputados, geramos automaticamente um conjunto de categorias de NER por meio de uma metodologia híbrida que combina técnicas de supervisão distante com um Tesouro específico do domínio. O corpus resultante foi então anotado e utilizado para treinar o modelo BERTimbau, baseado no modelo *Bidirectional Encoder Representations from Transformers* (BERT) otimizado para o português do Brasil. Nossa abordagem também incluiu a implementação de uma ferramenta web, em formato de prova de conceito, projetada para visualizar e interagir com as entidades extraídas, oferecendo insights valiosos sobre as tendências da linguagem política. Os resultados demonstram que, embora o modelo não atinja desempenho de ponta, ele reconhece efetivamente entidades-chave, tornando-se uma ferramenta útil para aplicações específicas na análise do discurso político. Este trabalho destaca o potencial de sistemas automatizados de NER em línguas e domínios pouco estudados, fornecendo uma base para futuras pesquisas e melhorias.

**Palavras-chave:** Reconhecimento de Entidades Nomeadas (NER). Discurso Político. Anotação de Corpus. Modelos de Linguagem.

## LIST OF FIGURES

Figure 1.1 End-to-end pipeline of the NER model .....	13
Figure 4.1 Distribution of speeches over time .....	24
Figure 4.2 BCoD speakers with the most speeches .....	24
Figure 5.1 Levenshtein distance impact over 100 samples .....	29
Figure 6.1 Frequency of Number of Tokens per Sentence (Up to 100) .....	37
Figure 7.1 Speech Viewer filtering interface.....	42
Figure 7.2 First result from a Speech Viewer search after filtering by party(PMDB) and UF(RS) .....	43
Figure 7.3 NER Highlight tool.....	43
Figure 7.4 NER Highlight tool with a processed input and its output; the cursor hovers over a named entity exposing its category.....	44

## LIST OF TABLES

Table 4.1	Sample Entries from the Corpus .....	22
Table 5.1	Sample Relations from the Corpus.....	26
Table 5.2	Thesaurus Entry: Partido Democrático Trabalhista (PDT) .....	27
Table 5.3	Representative Terms from the Excluded "Modificador" Category .....	28
Table 5.4	Most common categories.....	32
Table 5.5	Sample Entities for Each NER Category .....	33
Table 5.6	Sample Entries from the Annotated Corpus in CoNLL Format.....	34
Table 6.1	Overall Model Performance by Epoch - Validation Set .....	38
Table 6.2	Final Model Performance - Test Set.....	38
Table 6.3	Performance by NER Category (5th Epoch).....	40



## **LIST OF ABBREVIATIONS AND ACRONYMS**

AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
BCoD	Brazilian Chamber of Deputies
LLM	Large language model
ML	Machine learning
NE	Named entity
NER	Named-entity recognition
NLP	Natural language processing
PLM	Pre-trained language model

## CONTENTS

<b>1 INTRODUCTION</b> .....	<b>11</b>
<b>2 BACKGROUND</b> .....	<b>14</b>
<b>2.1 Brazilian Political Domain</b> .....	<b>14</b>
<b>2.2 Natural Language Processing</b> .....	<b>15</b>
<b>2.3 Named Entity Recognition</b> .....	<b>16</b>
<b>2.4 Large Language Models</b> .....	<b>17</b>
<b>2.5 Model evaluation</b> .....	<b>17</b>
<b>3 RELATED WORK</b> .....	<b>19</b>
<b>4 CORPUS</b> .....	<b>21</b>
<b>4.1 Data selection</b> .....	<b>21</b>
<b>4.2 Data extraction</b> .....	<b>21</b>
<b>4.3 Data preprocessing</b> .....	<b>22</b>
<b>4.4 Statistics</b> .....	<b>23</b>
<b>5 NAMED ENTITY RECOGNITION SET CREATION AND CORPUS AN-   NOTATION</b> .....	<b>25</b>
<b>5.1 Category set</b> .....	<b>25</b>
5.1.1 Thesaurus .....	26
5.1.2 Creating the named entity set.....	27
<b>5.2 Annotation</b> .....	<b>28</b>
<b>6 LANGUAGE MODEL</b> .....	<b>35</b>
<b>6.1 Training</b> .....	<b>35</b>
6.1.1 Setup .....	35
6.1.2 Dataset preparation .....	36
<b>6.2 Results</b> .....	<b>37</b>
<b>7 PROOF-OF-CONCEPT VISUALIZATION TOOL</b> .....	<b>41</b>
<b>7.1 Speech Viewer</b> .....	<b>41</b>
<b>7.2 NER highlight tool</b> .....	<b>41</b>
<b>8 CONCLUSION</b> .....	<b>45</b>
<b>REFERENCES</b> .....	<b>46</b>

## 1 INTRODUCTION

In recent years, the application of Natural Language Processing (NLP) has become increasingly vital for analyzing and understanding large-scale textual data across various domains, including politics and law. The ability to extract meaningful insights from vast amounts of unstructured text is critical, particularly in political discourse, where the subtleties of language and context play a significant role in shaping public opinion and policy. Despite advancements in NLP, there is a considerable gap in the development of tools specifically tailored to handle the unique linguistic characteristics of Brazilian Portuguese, especially within the political and legal contexts (JURAFSKY; MARTIN, 2009; NADEAU; SEKINE, 2007).

Addressing this gap, the primary objective of this work is to develop and evaluate a Named Entity Recognition (NER) model designed for Brazilian political speeches. NER is a crucial component in NLP tasks, allowing for identifying and classifying entities such as persons, organizations, locations, and dates within a text. However, existing NER systems often fall short when applied to languages or domains that differ significantly from the ones originally trained on (LAMPLE et al., 2016). This study aims to bridge this gap by creating a specialized NER model for Brazilian Portuguese, which can accurately recognize and categorize entities within political discourse.

In order to accomplish this task, we compiled a comprehensive corpus of Brazilian political speeches, covering a substantial time frame and a diverse range of speakers and topics. This corpus was the foundation for training a BERT-based NER model, leveraging the BERTimbau variant pre-trained on Brazilian Portuguese. The annotation process employed a hybrid approach, combining distant supervision techniques with dictionary-based methods to ensure the corpus was accurately and comprehensively labeled. This approach not only facilitates the identification of entities but also allows for the incorporation of domain-specific knowledge, making the model more robust and applicable to real-world scenarios (MINTZ et al., 2009).

Although other similar works have collected information from Congress legal documents (NUNES, 2023), our tool focuses on Congress members' thoughts and statements, making it essential to collect data from their official speeches. Fortunately, transparency is a significant value within the federal government today. Their sessions are widely documented: every statement made by each congressman is recorded and preserved through the government's Open Data initiative, a program aimed at promoting transparency and

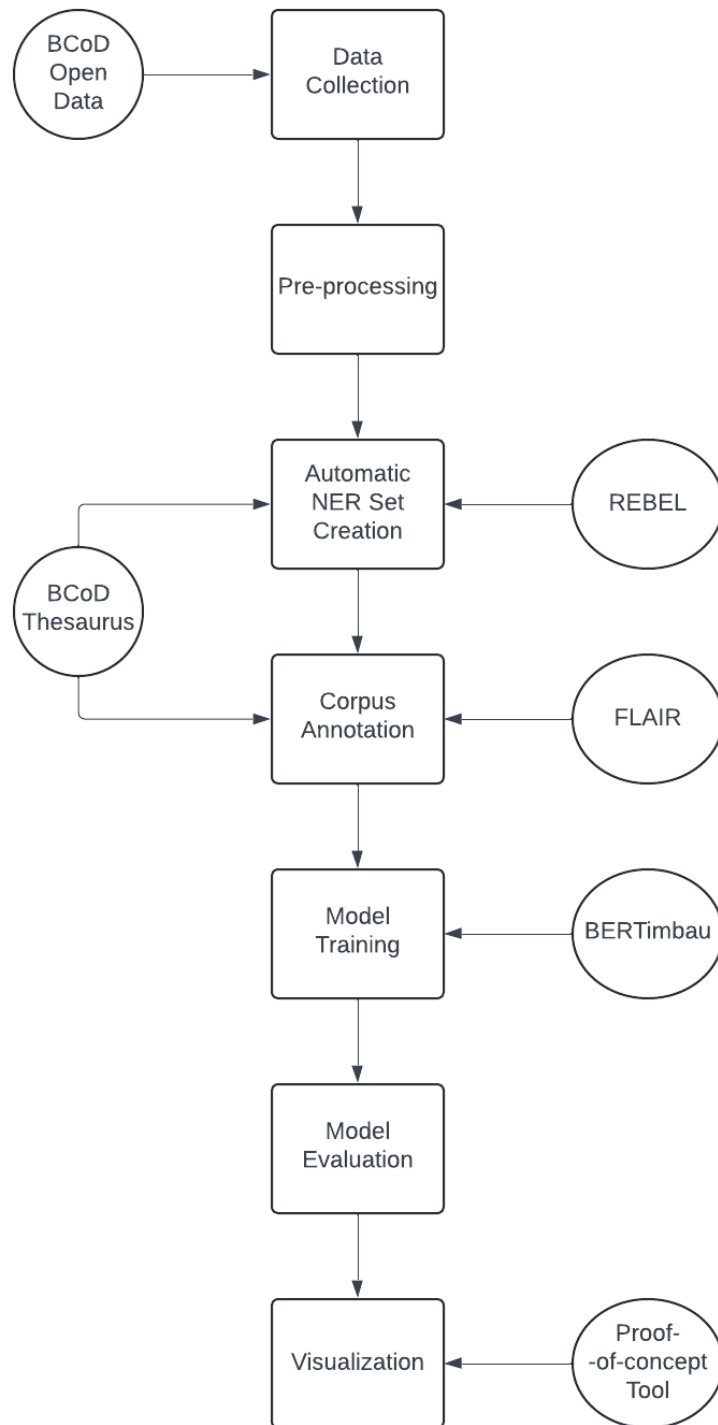
accountability in government.

In addition to developing the NER model, we implement a proof-of-concept web tool to visualize and interact with the extracted entities. This tool enables users to explore political discourse through the lens of NER, providing a practical application of the model and demonstrating its potential utility in academic and professional settings. We have designed this tool to be user-friendly and accessible, allowing for exploring relationships between entities and analyzing political language trends over time.

The contributions of this work are as follows: (i) the creation of a large-scale, annotated corpus of Brazilian political speeches, (ii) the development of a domain-specific NER model, and (iii) the implementation of a web tool for data visualization. These elements help investigate the state of NLP in Brazilian Portuguese, offering a foundation for future research and applications in political and legal domains. Figure 1.1 depicts a flowchart illustrating the end-to-end pipeline of the NER model, from corpus collection to visualization, highlighting key processes such as category creation, annotation, model training, and evaluation.

The remainder of this work is organized as follows: Chapter 2 provides the necessary background relevant to this study. Chapter 3 reviews related work, examining previous research in political discourse analysis, NLP applications in political contexts, and the development of tools for legislative data analysis. Chapter 4 details the corpus construction methodology, including data collection, preprocessing, and the criteria for selecting the texts analyzed in this study. Chapter 5 discusses the methodology for creating the entity set, as well as the automatic annotation, describing the implementation and the process of entity extraction and categorization. Chapter 6 describes the NER model training and evaluation process. Chapter 7 focuses on the visualization component, presenting the design and development of the web interface created to allow users to interact with and explore the annotated corpus. Finally, Chapter 8 concludes the work, summarizing the key findings, discussing their implications, and suggesting directions for future research.

Figure 1.1 – End-to-end pipeline of the NER model



## **2 BACKGROUND**

This chapter briefly touches on some fundamental concepts required to understand our methodology better. Section 2.1 presents an overview of how the Brazilian Chamber of Deputies operates. Section 2.2 introduces the concept of Natural Language Processing. Sections 2.3 and 2.4 address Named Entity Recognition and Large Language Models, respectively, two NLP technologies used in our work. Lastly, section 2.5 approaches metrics for evaluating language models.

### **2.1 Brazilian Political Domain**

The Brazilian political system is a federal presidential republic characterized by a clear separation of powers among the executive, legislative, and judicial branches. The bicameral legislative branch comprises the Federal Senate and the Chamber of Deputies. The latter is the lower house of the National Congress and plays a pivotal role in the creation and revision of federal laws, as well as in the oversight of the executive branch.

Diving into the structural intricacies of the Chamber of Deputies, it comprises 513 individuals, designated as deputies, who secure their positions through a system of proportional representation for four years. These deputies are the political proxies for Brazil's 26 states and the Federal District. The Chamber's operational mandate encompasses the formulation, scrutiny, and ratification of legislation that influences a broad spectrum of Brazilian society, including but not limited to economic policy, social welfare initiatives, and civil liberties. Beyond legislative duties, the Chamber wields the authority to commence impeachment procedures, conduct comprehensive reviews and approvals of the federal budget, and exert supervisory control over the executive branch through various investigative committees.

The legislative mechanism within the Chamber of Deputies consists of a multifaceted and multitiered procedural architecture designed to impose rigorous scrutiny on legislative propositions. Several actors may introduce a bill, including, but not limited to, deputies, the President, members of the judiciary, and public prosecutors' offices. Upon introduction, the bill consists of pertinent committees specializing in specific policy domains, ranging from finance and education to health, where it undergoes extensive deliberation and analysis.

Discussions within the Chamber usually present intense, multi-dimensional de-

bates encapsulating its member's and guests' diverse political, regional, and ideological spectra. Such debates are a cornerstone of the democratic process, enabling the articulation of divergent perspectives, fostering compromise, and facilitating consensus-building. The outputs of these debates are comprehensively documented, creating an exhaustive repository of legislative proceedings.

## **2.2 Natural Language Processing**

Natural Language Processing (NLP) operates at the intersection of computer science, artificial intelligence, and linguistics, with the primary aim of enabling machines to understand, interpret, and generate human language meaningfully and practically. This interdisciplinary field encompasses many tasks, from basic operations like text segmentation and tokenization to more sophisticated challenges such as machine translation, sentiment analysis, and information extraction. As highlighted by Jurafsky and Martin (2009), one of the core objectives of NLP is to create systems that can "extract meaning from the sea of language," facilitating human-computer interaction at a more intuitive level.

A fundamental aspect of NLP involves the transformation of raw text into a computationally manageable format. This process typically begins with tokenization, which breaks down text into constituent words or phrases, followed by preprocessing steps such as removing stop words, normalizing text, and managing punctuation. Manning, Raghavan and Schütze (2008) emphasize that these preprocessing steps are critical for reducing the complexity of the data before it is fed into machine learning models, ensuring that the models can focus on the most relevant features.

Moreover, NLP significantly emphasizes extracting meaningful information from text, a concept particularly relevant to this work. Techniques, like Named Entity Recognition (NER), part-of-speech tagging, and dependency parsing, allow machines to identify entities, classify words into syntactic categories, and understand the structural relationships within sentences. According to Collobert et al. (2011), these techniques form the backbone of tasks such as information retrieval and text summarization, enabling machines to "parse and understand the underlying structure of language" in a way that was previously unattainable with traditional methods.

## 2.3 Named Entity Recognition

NER is a crucial subtask within the broader domain of Natural Language Processing (NLP), focused on identifying and classifying proper nouns and other specific entities within unstructured text. The primary goal of NER is to automatically detect entities such as person names, organizations, locations, dates, and other predefined categories, facilitating a structured understanding of text data. The significance of NER lies in its ability to transform raw text into structured data, which is essential for various downstream applications such as information retrieval, question answering, and machine translation.

NER systems have evolved significantly since their inception, beginning with rule-based approaches that relied heavily on handcrafted features and lexicons. Early work in NER, such as the one presented by Grishman and Sundheim (GRISHMAN; SUNDHEIM, 1996), laid the groundwork by introducing rule-based and statistical methods to identify entities within text. These methods often used regular expressions and manually curated dictionaries, but their reliance on rigid rules made them less adaptable to new or unseen data.

The field of NER has further evolved with the advent of deep learning techniques, particularly using neural networks and pre-trained language models. The introduction of Conditional Random Fields (CRFs) combined with neural network architectures, as explored by Lample et al. (2016), marked a significant breakthrough in NER, enabling models to capture both local and global dependencies within the text. Furthermore, these advancements have been further propelled by the development of contextualized word embeddings, such as those provided by Bidirectional Encoder Representations from Transformers (BERT), which have set new benchmarks in NER performance (DEVLIN et al., 2018).

In recent years, the adoption of transfer learning has become increasingly prevalent in NER, with pre-trained models being fine-tuned on specific NER tasks. This approach has been efficient in low-resource languages or specialized domains where annotated data is scarce. The combination of pre-trained models and fine-tuning has enabled NER systems to achieve state-of-the-art results, significantly enhancing the accuracy and robustness of entity recognition across varied text corpora.



## 2.4 Large Language Models

Large Language Models (LLMs) represent a significant advancement in the field of NLP, characterized by their ability to understand and generate human-like text by leveraging vast amounts of data and computational power. These models, often based on deep learning architectures, have dramatically improved the performance of various NLP tasks such as text classification, translation, and summarization.

One of the foundational breakthroughs in LLMs was the introduction of the Transformer architecture by Vaswani et al. (2017), which enabled the efficient training of models on extensive datasets by capturing long-range dependencies in text through self-attention mechanisms. This architecture laid the groundwork for subsequent models that have set new benchmarks in NLP.

Among the most influential LLMs is BERT, introduced by Devlin et al. (2018). BERT differs from its predecessors by applying bidirectional training, allowing it to consider the context from both the left and right of a given word, significantly enhancing its ability to understand nuanced language. This bidirectional approach has made BERT particularly effective in various tasks, from NER to question answering, solidifying its role as a cornerstone in modern NLP.

The advent of even larger models, such as GPT-3 (Brown et al., 2020), has pushed the limits of what LLMs can achieve, demonstrating the ability to perform complex tasks with minimal task-specific training, a phenomenon known as few-shot learning. These models are trained on diverse datasets encompassing vast quantities of text, enabling them to generate coherent and contextually relevant language across various domains.

## 2.5 Model evaluation

Evaluating the performance of machine learning models is critical to ensuring that they perform well on the tasks for which they are designed. Model evaluation involves using various metrics that provide insights into different aspects of model performance, allowing researchers to make informed decisions about model selection and optimization.

One of the most commonly used metrics in machine learning (ML), and consequently, in modern NLP, is accuracy, which measures the proportion of correct predictions out of the total number of predictions. While accuracy is straightforward to interpret, it

can be misleading in cases where the data is imbalanced. In such scenarios, accuracy may overestimate the model's performance by ignoring the importance of minority classes (POWERS, 2011).

To address the limitations of accuracy, precision and recall are often used, particularly in tasks like NER, where the focus is on correctly identifying specific entities within a text. Precision is the ratio of correctly predicted positive instances to the total predicted positives, measuring how many of the identified entities are relevant. Conversely, recall is the ratio of correctly predicted positive instances to all actual positives, reflecting the model's ability to capture all relevant instances (MANNING; RAGHAVAN; SCHÜTZE, 2008).

The F1 score is a harmonic mean of precision and recall, offering a single metric that balances the trade-off between these two measures. The F1 score is particularly useful when the distribution of classes is uneven or when the cost of false positives and false negatives is different. It provides a more nuanced view of model performance than accuracy alone (SASAKI, 2007).

### 3 RELATED WORK

The legal domain - which is intrinsically connected to the BCoD's political domain - has received much attention from the AI research community. There are many reasons for this interest: the immense amount of data produced, the richness and complexity of its texts, and the need for better ways to understand it. The field of NLP, through AI techniques, is perfect for handling most of the issues optimally presented within this domain. Zhong et al. (2020) thoroughly reviews many possible NLP tasks that can greatly benefit the area. Polo et al. (2021) does something similar while providing tools to execute these tasks specifically in the Brazilian Portuguese domain. Garcia et al. (2024) contributes to improving NLP solutions in the Brazilian legal context by providing enhanced models, a specialized corpus, and a benchmark dataset.

In the modern landscape where machine learning is being successfully employed to handle NLP tasks, the first step for any interested researcher is choosing or building a comprehensive corpus of their domain of interest. When it comes to the legal domain, specifically the Brazilian Portuguese domain, there are few but remarkable results available. MultiLegalPile (NIKLAUS et al., 2024), Ulysses-Tesemõ (SIQUEIRA et al., 2024), Iudicium Textum (SOUSA; FABRO, 2019), Acordãos TCU<sup>1</sup> and DataSTF<sup>2</sup> are examples of relevant and extensive corpora that could easily be used for domain studies.

Concerning political speeches, however, the amount of available data falls drastically. Rodrigues et al. (2023) built a dataset with speeches from the Portuguese parliament to help with their work on a transformer model, but it was not the focus of the study. To the best of our knowledge, there are no similar datasets from the Brazilian Congress.

As for the NLP subtask that is the target of our work - NER - there are some relevant studies centered around the Brazilian Portuguese subdomain, most of which were gathered in a survey by Albuquerque et al. (2023). Regarding datasets, LeNER-BR is composed entirely of Brazilian legal documents and contains specific tags for law and legal case entities (ARAUJO et al., 2018). Ulysses-BR consists of bills and legislative consultations from the BCoD (ALBUQUERQUE et al., 2022). On the other hand, Zanuz and Rigo (2022) presents the first fine-tuned BERT models trained exclusively on Brazilian Portuguese for Legal NER. Further studies centered on applying NER techniques include using distant supervision to extract entities (NAVAREZI et al., 2022), as well as NER approaches with self-learning (NUNES et al., 2024) and in-context learning (NUNES et

---

<sup>1</sup><<https://www.kaggle.com/datasets/ferraz/acordaos-tcu>>

<sup>2</sup><<https://legalhackersnatal.wordpress.com/2019/05/09/mais-dados-juridicos/>>

al., 2024).

Ultimately, our proof-of-concept for visualizing the results of the NLP pipeline drew influence from works such as Assogba et al. (2011), responsible for a web-based set of visualization tools that reveals the underlying semantics of a legislative bill.

As close as political speeches are to the legal domain, there are still intricacies to the human vocal language that require proper study, and as far as we know, no such study exists for this particular case, which served as a strong motivation for this study.

## 4 CORPUS

Our process builds an extensive corpus as our first step to achieving our final goal of providing users with a comprehensive, up-to-date, and coherent tool. The corpus is the foundation of our NER system, providing the essential data required for training and evaluation of the model that would serve as the core of this tool. This chapter outlines the process of constructing the corpus, from initial data extraction to the final annotated dataset.

### 4.1 Data selection

Our first major choice was to select which subset of the total amount of speeches would be used in the corpus. A broader view, encompassing data from the Chamber of Deputies and the Federal Senate, thus containing the entire National Congress context, was initially considered. As often happens, that approach had some clear benefits, like following a politician's career when shifting from one institution to another. However, this method posed significant challenges, including doubling the extraction workload and substantially increasing the volume of data to process and later train the language model. Ultimately, we chose to tighten our context and retrieve the already considerable quantity of speeches exclusive to the Chamber of Deputies.

Additionally, we faced a constraint related to the time range of the data. Our data extraction technique, explored in the next section, was only effective for speeches delivered from the year 2000 onward up to the end of 2023. This time range limitation was an external constraint imposed by the available data and the extraction method's compatibility with the data format.

### 4.2 Data extraction

The Chamber of Deputies website offered a tool for visualizing most of the information covered by the Open Data program, but when it came to collecting data the process was not as seamless. The newest version of the available API, which conforms to modern RESTful standards, was made partially available in 2017<sup>1</sup>. However, the data

---

<sup>1</sup><<https://github.com/CamaraDosDeputados/dados-abertos>>

containing the Congressmen’s speeches had not yet been ported to this new API, remaining accessible only through traditional Web Services.

We developed Python scripts to identify and retrieve speeches using the Legislative Open Data (DADOS ABERTOS, 2024) service. It provides a collection of features that allow direct access to legislative data produced by the Chamber of Deputies, including information on deputies, legislative bodies, propositions, plenary sessions, and committee meetings. The obtained data was available in the XML format, which we converted into JSON for better readability, compatibility, and performance. It was important to retrieve the speeches and key metadata that our visualization interface would later leverage. Table 4.1 shows the metadata we chose to preserve and excerpts from manually selected sample speeches.

Table 4.1 – Sample Entries from the Corpus

<i>Name</i>	<i>Date</i>	<i>Party</i>	<i>UF</i>	<i>Speech Excerpt</i>
Francisco Rodrigues	2003-08-25	PFL	RR	"Sr. Presidente, Sras. e Srs. Deputados, neste 25 de agosto..."
Aldo Rebelo	2006-03-08	PCdoB	SP	"O voto 'sim' é pela aprovação do parecer..."
Onofre Santo Agostini	2013-09-04	PSD	SC	"Acho que está havendo um equívoco..."
Fábio Sousa	2017-11-29	PSDB	GO	"Eu considero que esses dois projetos são de suma importância..."
André Figueiredo	2022-04-26	PDT	CE	"Portanto, em defesa do esporte brasileiro..."

### 4.3 Data preprocessing

While transformer models can handle low levels of data cleaning (VIANNA; MOURA, 2022), we opted to perform some preprocessing to arrive at a clean corpus that could be better leveraged to a more varied set of tasks. The first and mandatory step was removing HTML and XML tags, as our extraction method pulled complete web documents. We focused on the speeches after the relevant metadata was cleaned and properly stored as JSON key-value pairs. They were formatted in RTF and encoded as Base64, two undesirable characteristics that we reversed, resulting in speeches in plain text.

The original data also possessed much textual noise that could interfere with the quality of possible visualization features. By applying proper encoding using UTF-8

standard and removing special formatting characters such as `\n` and `\t`, we arrived at a format that could be easily understood by language models, NLP tools, or even the naked eye.

Thinking ahead, one defining feature of these speeches had to be dealt with: their length. Some were very long (up to 86936 words), which would cause troubles later on as language models have token limits. Therefore, when training the models, we split every speech into complete sentences to avoid likely truncation issues.

#### **4.4 Statistics**

The final corpus we arrived at contained 463,085 speeches from the BCoD, ranging from December 2000 to December 2023. They were delivered by 8479 different people, 5167 of whom were members of the BCoD at some point. The remaining speakers consisted of guests from outside the BCoD and included fellow politicians and representatives from various sectors of Brazilian society. The total word count was 209,502,881, with an average speech length of 452.41 words and a standard deviation of 600.09. The size of the vocabulary used in these speeches was 306,210 words. To help visualize the contents of the corpus, figure 4.1 shows a distribution of the speeches over time, while figure 4.2 presents a sample with the BCoD's most prolific speakers.

Figure 4.1 – Distribution of speeches over time

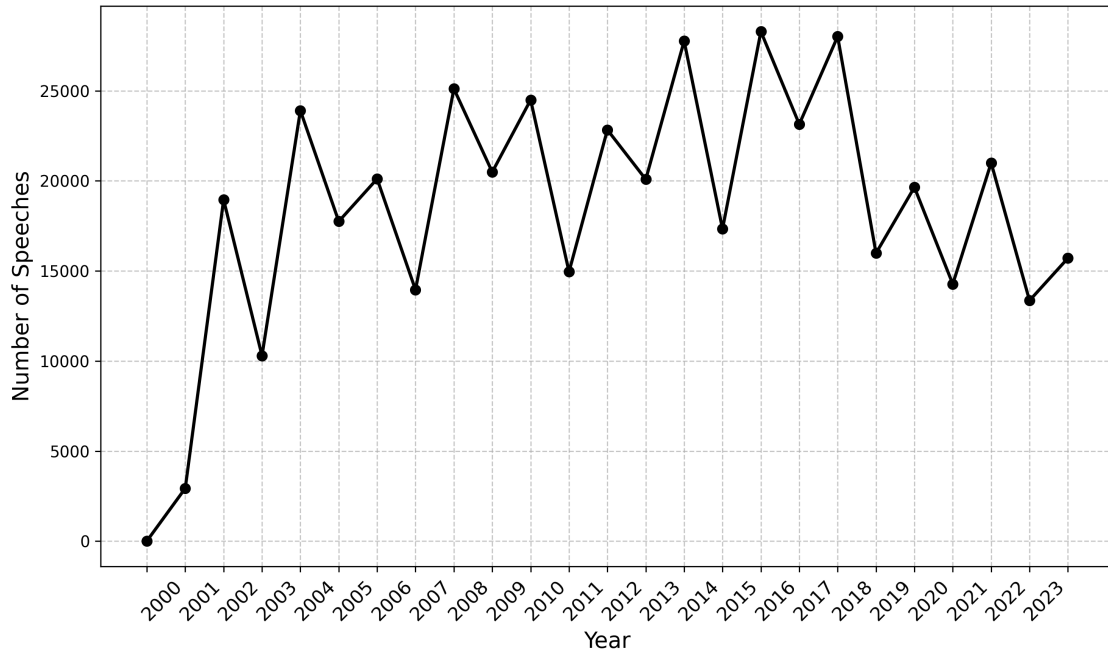
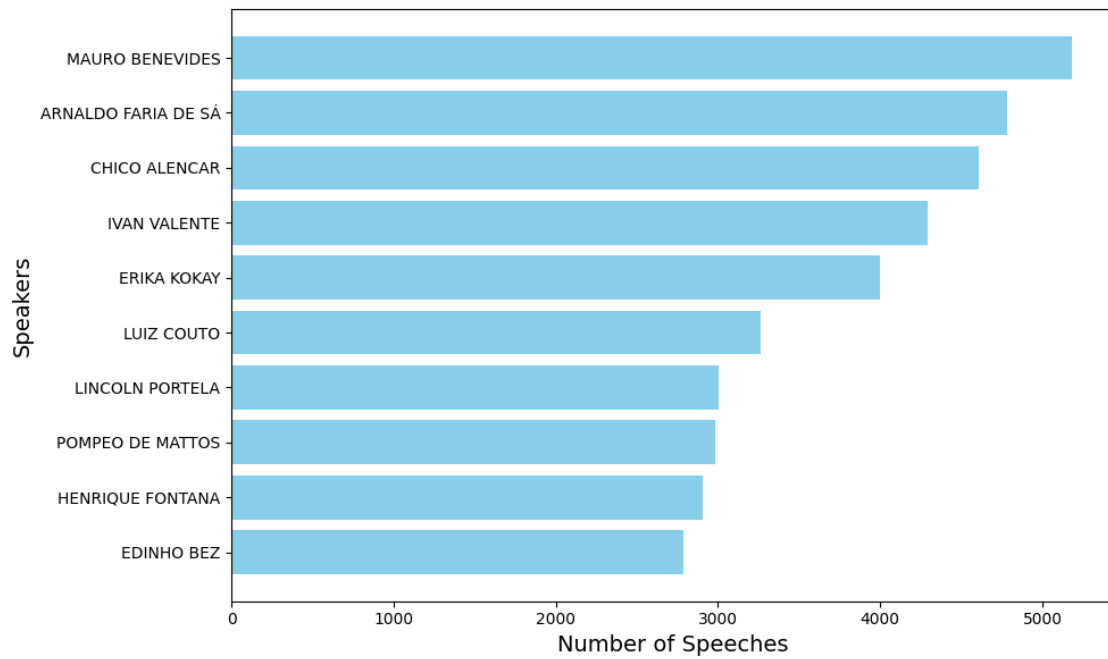


Figure 4.2 – BCoD speakers with the most speeches





## 5 NAMED ENTITY RECOGNITION SET CREATION AND CORPUS ANNOTATION

After finalizing our corpus, the next step was to train a language model. However, we faced with two issues. The first issue was generating a set of categories to be used for the task of NER. The next issue that is very common when collecting data for NER training is the lack of annotations. To properly train, validate, and test the language model, we needed to annotate our corpus or at least part of it. In this chapter, we'll discuss how we handle each issue.

### 5.1 Category set

Named entities (NEs) are usually divided into two types: generic NEs (e.g., person and location) and domain-specific NEs (e.g., proteins, enzymes, and genes) (LI et al., 2022). We focused on generating a set of domain-specific NEs automatically, but to improve both our visualization tool and the language model's performance, we also identified generic NEs. This section will go over the former.

The approach for automatically generating a NER category set in this study follows the methodology described by Matos, Rodrigues and Teixeira (2024). Their relevant and innovative approach serves as the foundation for the steps outlined in this section.

The first step was to extract relations from the sentences of our corpus. According to Manning and Schütze (1999), relation extraction aims to identify and classify relationships between pairs of entities within a text. We achieve this by using the state-of-the-art language model called Relation Extraction By End-to-end Language Generation (REBEL), an autoregressive approach that frames Relation Extraction as a seq2seq task (CABOT; NAVIGLI, 2021). The resulting leads to triplets composed of a head, tail, and type by applying our corpus speech to this model. "Heads" and "tails" are the entities identified by REBEL, and "type" is the type of the relation. Table 5.1 illustrates a few examples of these triplets. Only the entities are relevant in this method, so we stored them in a unique set.

After collecting these "potential entities", the next step consists of obtaining fitting categories for them. Matos, Rodrigues and Teixeira (2024) uses the Portuguese subset of the WikiNER dataset (NOTHMAN et al., 2013), which contains 7,200 manually-labelled

Table 5.1 – Sample Relations from the Corpus

<i>Head</i>	<i>Type</i>	<i>Tail</i>
Diego	Occupation	Policial
Amazonas Energia	Headquarters Location	Manaus
ANEEL	Has Part	Conforme prevê solução da ANEEL
JORGE SOLLA	Member of Political Party	PT
Ideal Clube	Inception	1985

Wikipedia articles. They map these entities to their Wikidata entries using the Wikimap-  
per tool <sup>1</sup> and extract that entry’s category, using it as a potential category for NER. We  
tried to replicate their approach, but due to the lack of public availability of a precom-  
puted index for the Portuguese Wikipedia dump and the effort it would take to create it,  
we went in a different direction and started to drift away from the method proposed by  
Matos, Rodrigues and Teixeira (2024), instead following our own which we describe in  
the next section.

### 5.1.1 Thesaurus

The BCoD keeps a daily updated Thesaurus file with all relevant terms to help  
users understand the legislative domain as part of the Open Data initiative. Table 5.2  
presents a term detailed in this dictionary, and the complete version is available online at  
the BCoD Open Data portal<sup>2</sup>. Each entry in this file has categories, much like Wikidata  
entries. It contains 70,577 indexed terms spread over 79 categories. We decided to use  
this Thesaurus to map our entities because of its clarity, availability, and strong relevance  
to the domain of our corpus.

The most obvious choice to mimic our referenced paper’s methodology was to di-  
rectly extract the category of each entry as a possible category for NER. However, using  
the Thesaurus brought two key issues with this proposal. Some categories are very broad,  
as seen in table 5.2 where "Identificador" encompasses identifiers for political parties,  
organizations, commemorative dates, and animals. We solved this issue by using subcat-  
egories instead, as they include most of the actual categories and specifying ones like the  
aforementioned "Identificador" into "Identificador de Partidos Políticos," for instance. Is-  
sue number two was that a dictionary term could have multiple subcategories. Given that  
there was no recognizable pattern in the order these subcategories were applied to each

<sup>1</sup><<https://github.com/jcklie/wikimapper>>

<sup>2</sup><<https://dadosabertos.camara.leg.br/arquivos/tecadTermos/json/tecadTermos.json>>

Table 5.2 – Thesaurus Entry: Partido Democrático Trabalhista (PDT)

<i>Field</i>	<i>Details</i>
Term Code	31762
Term	Partido Democrático Trabalhista (PDT)
Category	Identificador
Subcategory	Identificador de Partidos Políticos
Explanatory Notes	Partido político. Registro provisório: Resolução -TSE nº 10.899, de 16 de setembro de 1980.
Historical Notes	
Applicative Notes	
Sources	
Use	
Used For	PDT
Specific Terms	
Generic Terms	
Related Terms	Partido Comunitário Nacional (PCN) (1988)

term by the developers of the Thesaurus, we also chose what we believed was as close to an impartial solution as possible: selecting one randomly. We recognize this might have influenced our results, and analyzing the impact should be a priority in future works.

### 5.1.2 Creating the named entity set

With our entities extracted by REBEL now properly matched to categories from the Thesaurus, we needed to assemble the proper categories for NER. In this context, matching refers to the process of comparing strings (entities) from our extracted data with the terms in the Thesaurus to identify exact or close alignments between them. This ensures that each entity is correctly associated with a relevant category from the Thesaurus. Matos, Rodrigues and Teixeira (2024) chose to use the N most common categories from their Wikidata mapping, varying N from 5 up to 20 and comparing the results. Given that this variance did not significantly affect the final results of their model, we chose a set big enough to encompass the categories we judged relevant, settling on the size of 30.

We had an outlier case with a specific subcategory called "Modificador." This category was by far the most common in the corpus because it encompassed common words, mostly nouns, that did not correlate with specific subjects or ideas. Table 5.3 provides a sample of these words. We ignored this subcategory to avoid cluttering our data with potentially useless information.

We ran a mapping function to select these 30 categories, matching our extracted

Table 5.3 – Representative Terms from the Excluded "Modificador" Category

---

agrupamento  
convocação  
luta  
programação  
titularidade

---

entities with the Thesaurus and verifying which of the dictionary's subcategories were the most frequent in our corpus. The final selection and the number of times each of them was matched can be seen in Table 5.4. Table 5.5 helps visualize the selected categories, providing a sample entry of the Thesaurus for each one of them.

## 5.2 Annotation

We already had our automatically generated categories to use in NER tasks; we only needed to train a language model. The next major issue, however, presented itself: we had no annotated data to feed the model with, unlike Matos, Rodrigues and Teixeira (2024).

To address this, we employed a custom hybrid technique inspired by methods such as distant supervision and the traditional gazetteer approach. Distant supervision is a technique in which the labeled data is automatically generated by aligning the unannotated text with an external knowledge base (MINTZ et al., 2009). Gazetteer methods use predefined lists of entities to annotate text, typically in tasks requiring high precision (CUNNINGHAM et al., 2002).

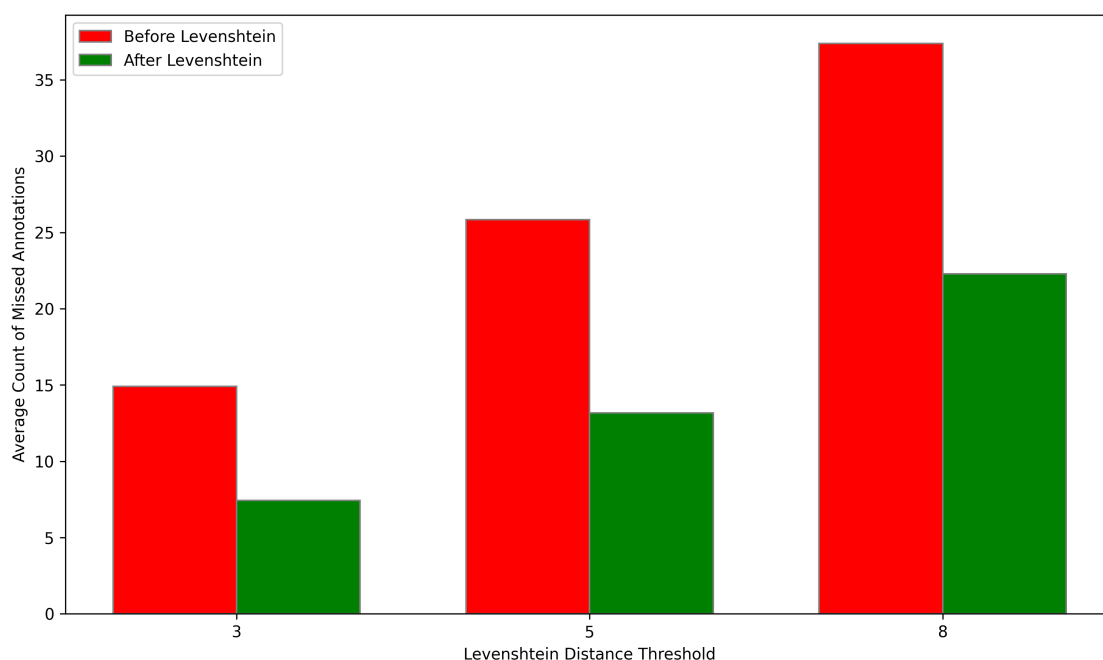
We leveraged the domain-specific dictionary the BCoD's Thesaurus provided, applying a reverse matching strategy. Specifically, we took each dictionary term within our 30 selected subcategories and compared it against the list of potential entities extracted from REBEL. When a match was identified, we annotated the corresponding entries in the corpus with the term's subcategory, effectively labeling the data.

During initial testing, we observed that this approach showed promise but required refinement. One main issue was identified: missed annotations. Some dictionary terms failed to match precisely with the entities, leading to their omission by the algorithm. To address this issue, we incorporated the Levenshtein Distance (LEVENSHTEIN et al., 1966), a commonly used metric in NLP, to allow for approximate matching. This method evaluates the similarity between strings by measuring the number of single-character edits

required to transform one string into another. By fine-tuning the distance thresholds, we significantly improved the accuracy of the annotation process while avoiding mistakenly matching with incorrect entities. Figure 5.1 shows the results after testing the Levenshtein Distance application against 100 entries with the different established thresholds, that were set based on a string's length. The longer the string, the more leeway we allowed. The red bar represents, out of this 100 entity sample, how many that should've been tagged were missed by our algorithm before the implementation of the Levenshtein distance. This is done for each of our chosen thresholds (3,5 and 8). The green bar, on the other hand, shows the improved results after applying this method, with fewer missed entities.

Some additional text processing was also employed to avoid mismatches. Separators like "-" were ignored when matching. The content inside the parenthesis was treated as a different term since many dictionary entries had the full name and possible abbreviations as a single entry, with both forms possibly appearing on the corpus. When there were cases of nested entities, we chose the more specific, larger one for the annotation.

Figure 5.1 – Levenshtein distance impact over 100 samples



While we were aware of the limitations of distant supervision, particularly its tendency to introduce noise through automatic labeling, we determined that it was the most viable option given our constraints. In political discourse, where context and nuanced language play a significant role, distant supervision often results in incorrect or incomplete annotations due to ambiguities in aligning text with knowledge bases. However,

the alternative—manual annotation—was impractical given the vast size of our corpus and the specialized nature of the political lexicon we were working with. Although this method’s reliance on existing structured data means that rare or emerging entities may be overlooked, the ability to quickly generate large amounts of labeled data was essential for our project. As Mintz et al. (2009) note, distant supervision can compromise data quality, but additional refinement and filtering provided us with a workable solution under the circumstances.

Given our very human domain of politics and speeches, our annotated corpus would still be severely lacking by omitting two specific NER categories: persons and dates. Neither were present in the Thesaurus and, therefore, were ignored by our method. Thinking ahead, we felt that our model and possible continued developments of this work could be hindered without including two of the most common and important categorizations (NADEAU; SEKINE, 2007). We decided to add another step to our process to mitigate that and possibly complement the shortcomings of a distant supervision-like approach.

We applied a different, state-of-the-art FLAIR model to our partially annotated corpus. We used its NER capabilities to tag precisely the two missing categories: persons (PER) and dates (DATE). FLAIR is an NLP framework designed to facilitate the training and distribution of state-of-the-art sequence labeling, text classification, and language models (AKBIK et al., 2019). We chose it because of its capabilities and because it had also been used by Matos, Rodrigues and Teixeira (2024). We selected its largest variant for NER, "ner-large," which obtained an F1 score of 94,36 and filtered it to tag persons and dates in our corpus specifically.

We finally arrived at an adequately annotated corpus, and our final step was to transform it into a proper format for model integration. The most common method of formatting a dataset for NER involves converting the data into a CoNLL (Conference on Natural Language Learning) format, which is widely used in the NLP community. In this format, each word in the text is placed on a new line, followed by its corresponding entity tag. A blank line separates sentences, and special tokens such as "O" are used to indicate words that do not belong to any entity class. The IOB2 tagging scheme, which stands for Inside-Outside-Beginning, is used in our dataset due to its simplicity and widespread adoption in NER tasks. In this scheme, each word in a sentence is tagged as either the beginning of a named entity (B), inside a named entity (I), or outside a named entity (O). This approach ensures compatibility with established models and tools, making it

a standard choice in this field (RAMSHAW; MARCUS, 1995). Table 5.6 presents one manually selected partial sentence from the corpus after going through the annotation process.

At this point, we had finally arrived at the desired dataset and could move on to training the model.

Table 5.4 – Most common categories

<i>Category</i>	<i>Frequency</i>	<i>Proportion</i>
Direito Constitucional	3,248,145	
Processo Legislativo e Atuação Parla- mentar	3,092,265	
Administração Pública	1,789,031	
Política, Partidos e Eleições	1,724,894	
Identificador de Acidentes, Localidades e Nomes Geográficos	1,567,095	
Modificador	1,066,120	
Identificador de Conselhos, Entidades e Organizações	987,194	
Defesa e Segurança	486,632	
Direito Civil e Processual Civil	466,155	
Identificador de Programas, Planos, Pro- jetos e Sistemas	437,470	
Saúde	421,168	
Cidades e Desenvolvimento Urbano	413,085	
Trabalho e Emprego	410,883	
Economia	406,107	
Comunicações	332,202	
Educação	319,217	
Ciência, Tecnologia e Inovação	298,677	
Relações Internacionais e Comércio Ex- terior	285,634	
Finanças Públicas e Orçamento	259,894	
Indústria, Comércio e Serviços	258,126	
Direito Penal e Processual Penal	246,419	
Meio Ambiente e Desenvolvimento Sus- tentável	238,267	
Identificador de Animais	228,183	
Identificador de Cargos, Ocupações e Categoria Profissional	223,629	
Viação, Transporte e Mobilidade	207,476	
Identificador de Partidos Políticos	206,484	
Identificador de Materiais e Produtos	201,310	
Arquitetura e Urbanismo	180,636	
Arte, Cultura e Religião	174,716	



Table 5.5 – Sample Entities for Each NER Category

<i>Category</i>	<i>Sample Entity</i>
Direito Constitucional	Ação afirmativa
Processo Legislativo e Atuação Parlamentar	Matéria aduaneira
Administração Pública	CCE (Cargo Comissionado Executivo)
Política, Partidos e Eleições	Quebra de sigilo
Identificador de Acidentes, Localidades e Nomes Geográficos	Puerto Rico
Identificador de Conselhos, Entidades e Organizações	Zoológico de São Paulo
Defesa e Segurança	Autoridade Marítima Brasileira (AMB)
Direito Civil e Processual Civil	Massa falida
Identificador de Programas, Planos, Projetos e Sistemas	Regiões Especiais de Turismo (RET)
Saúde	Código genético
Cidades e Desenvolvimento Urbano	Licença de ampliação
Trabalho e Emprego	Promoção profissional
Economia	Sistema de Financiamento Imobiliário (SFI)
Comunicações	Destinatário
Educação	Plataforma Lattes
Ciência, Tecnologia e Inovação	Conta de usuário
Relações Internacionais e Comércio Exterior	Tratado comercial
Finanças Públicas e Orçamento	Alíquota
Indústria, Comércio e Serviços	Cooperação empresarial
Direito Penal e Processual Penal	Medida socioeducativa
Meio Ambiente e Desenvolvimento Sustentável	Desmatamento ilegal
Identificador de Animais	Elefante
Identificador de Cargos, Ocupações e Categoria Profissional	Reitor
Viação, Transporte e Mobilidade	Acostamento
Identificador de Partidos Políticos	MDB
Identificador de Materiais e Produtos	Catalisador automotivo
Arquitetura e Urbanismo	Habitação flutuante
Arte, Cultura e Religião	Revista em quadrinhos

Table 5.6 – Sample Entries from the Annotated Corpus in CoNLL Format

<i>Word</i>	<i>Dictionary Term</i>	<i>Tag</i>
e		O
sempre		O
testemunhei	Testemunha	B-Direito Civil e Processual Civil
sua		O
competência		O
,		O
seriedade		O
e		O
ponderação		O
,		O
o		O
que		O
me		O
faz		O
convicto		O
de		O
que		O
meu		O
partido		O
,		O
o		O
PMDB	PMDB	B-Identificador de Partidos Políticos
,		O
o		O
nosso		O
Líder	Líder	B-Processo Legislativo e Atuação Parlamentar
,		O
o		O
Deputado	Deputado	B-Processo Legislativo e Atuação Parlamentar
Henrique		B-PER
Alves		I-PER
,		O
e		O
o		O
nosso		O
Vice-Presidente	Vice-presidente da República	B-Administração Pública
da	Vice-presidente da República	I-Administração Pública
República	Vice-presidente da República	I-Administração Pública
,		O
Michel		B-PER
Temer		I-PER
,		O
sentirão		O
muito		O
orgulho		O

## 6 LANGUAGE MODEL

In this chapter, we review the choices and process of training an LLM using our dataset. We then proceed to evaluate it using common industry metrics. Although it would be ideal to see how the dataset performed in different models, we stuck to the principle of achieving one usable result to later employ on our visualization tool.

### 6.1 Training

We chose a BERT-based model for training the dataset due to its demonstrated effectiveness in capturing complex contextual relationships within the text, which is critical for NER (DEVLIN et al., 2018). BERT’s bidirectional architecture enables the model to consider both preceding and succeeding words, thereby providing a more comprehensive understanding of context, which is particularly important when dealing with the intricate language and entities in political discourse.

More specifically, BERTimbau was selected as the model of choice because it is pre-trained on a substantial corpus of Brazilian Portuguese text, aligning closely with the linguistic characteristics of the dataset (SOUZA; NOGUEIRA; LOTUFO, 2020). Given that our dataset comprises speeches and texts deeply rooted in the Brazilian political language, BERTimbau’s ability to understand and process this specific linguistic context ensures a higher level of accuracy in entity recognition, which is essential for the success of our NER task. The base version of BERTimbau, which we used for this training, comes from HuggingFace<sup>1</sup>. This version has 12 layers, 768 hidden dimensions, 32 attention heads, and 110M parameters.

#### 6.1.1 Setup

The training was performed remotely in a cluster maintained by the High-Performance Computing Facility of our academic institution’s Department of Informatics. The specific node featured an Intel i9-14900KF (Q4’23), 3.2 GHz, 32 threads, 16 + 8 cores as its CPU, 64 GB of DDR5 RAM, and an NVIDIA RTX 4090 with 16384 CUDA cores.

Regarding software, we used Python 3.11 for the training scripts due to its balance

---

<sup>1</sup><https://huggingface.co/neuralmind/bert-base-portuguese-cased>

between modern features, compatibility with required libraries, and stability. We relied on the Hugging Face Transformers library<sup>2</sup>, which provides an easy-to-use interface for implementing state-of-the-art models like BERTimbau. The training process was managed using the PyTorch deep learning framework<sup>3</sup>, known for its flexibility and efficiency in handling large-scale machine learning tasks. We also utilized the Datasets library from Hugging Face<sup>4</sup> to streamline the preprocessing and management of our dataset, ensuring seamless integration with the model training pipeline. Additionally, we employed Seqeval<sup>5</sup> for evaluating the performance of our NER model, as it provides specialized metrics like precision, recall, and F1 score tailored specifically for sequence labeling tasks. Seqeval's ability to accurately assess the quality of the BIO-tagged sequences in our dataset made it an essential tool for ensuring the reliability of our model's output.

### 6.1.2 Dataset preparation

We split the dataset into training, validation, and test sets. Specifically, 70% of the data was allocated for training, 15% for validation, and the remaining 15% for testing. This split was designed to ensure that the model was trained and evaluated on distinct subsets of data and is a common practice in machine learning to ensure robust model evaluation (BISHOP, 2006).

When choosing the training parameters, we opted to remain as close as possible to the most widely accepted default settings that have been proven effective in various NER tasks. We opted for this decision because default hyperparameters, as recommended in literature and by established frameworks, offer a balanced starting point for training deep learning models. These settings result from extensive empirical research and are often chosen to provide a good trade-off between model performance and training stability. By adhering to these default values, we aimed to minimize the risk of introducing unnecessary complexity or instability into the training process, allowing the model to achieve reliable results on our dataset without extensive hyperparameter tuning (DEVLIN et al., 2018; WOLF et al., 2020).

We employed the default BERTimbau tokenizer, the Hugging Face library provided, which is pre-trained and optimized for Brazilian Portuguese text, ensuring that the

---

<sup>2</sup><<https://huggingface.co/transformers/>>

<sup>3</sup><<https://pytorch.org/>>

<sup>4</sup><<https://huggingface.co/docs/datasets/>>

<sup>5</sup><<https://github.com/chakki-works/seqeval>>

input data was tokenized consistently with the model’s pre-training. The dataset consisted of 294,682,942 total tokens. Figure 6.1 shows a breakdown of the size of sentences in terms of tokens. We also adhered to the default model architecture and optimization settings, including the AdamW optimizer.

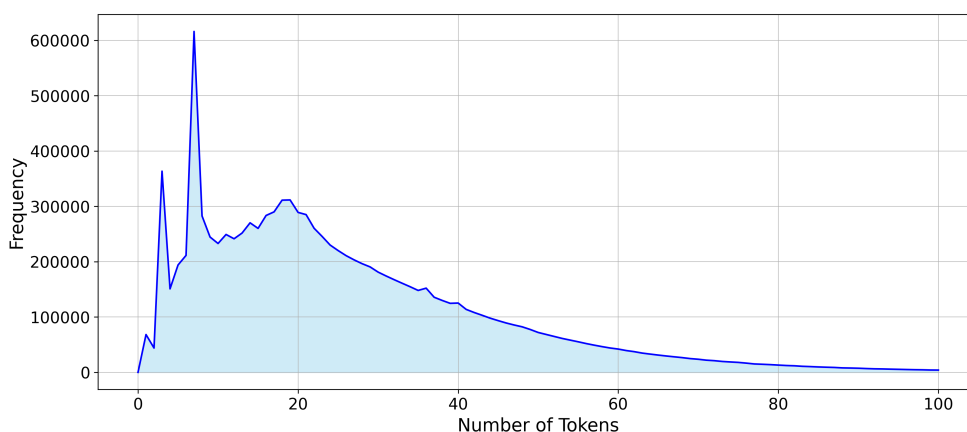


Figure 6.1 – Frequency of Number of Tokens per Sentence (Up to 100)

We used a learning rate of  $5e-5$  to facilitate stable and efficient convergence, while a batch size of 32 was chosen to optimize computational performance without sacrificing training stability. We opted for three epochs to balance sufficient learning and minimize the overfitting risk. Additionally, setting the maximum sequence length to 128 ensured that the model could fit over 95% of the sentences in our corpus without truncation, and any shorter sentences were filled with padding, following default settings.

## 6.2 Results

The primary objective of this section is to evaluate the effectiveness of our BERT-based model in performing NER on the Brazilian political discourse dataset. To this end, we assessed the model’s performance using four key metrics: Precision, Recall, F1 Score, and Accuracy. The F1 Score was calculated using a macro-averaging approach, meaning that the F1 score was computed independently for each category, and then the average was taken across all categories. This approach provides a balanced view of the model’s performance across all entity types, regardless of their frequency in the dataset. The metrics for individual classes are detailed in the corresponding table, where a breakdown by category offers deeper insights into the model’s strengths and areas for improvement.

It is important to note that during the preprocessing of the dataset, duplicate sen-

tences were not removed. This oversight could potentially lead to data leakage, where similar or identical sentences with different annotations appear in both the training and test sets, potentially inflating the model’s performance. Future iterations of this work should carefully ensure that duplicates are identified and removed, and that sentences with similar structures are evenly distributed across training and test folds to avoid any unintended data leakage.

During the training phase, we monitored the model’s performance per epoch using the key metrics mentioned earlier, calculated on the validation set. These per-epoch metrics provided insights into the model’s learning progress and helped in tuning the training process.

Table 6.1 shows these metrics after our training’s completion. We can see that every metric improved in the initial epochs but soon started to stabilize, suggesting the model was close to reaching its maximum performance and further training might not have been beneficial. These metrics were extracted from the validation split results, to show the variations per epoch.

Table 6.1 – Overall Model Performance by Epoch - Validation Set

<i>Epoch</i>	<i>Precision (%)</i>	<i>Recall (%)</i>	<i>F1 Score (%)</i>	<i>Accuracy (%)</i>
1	58.4	49.3	53.5	60.3
2	63.5	58.0	60.6	66.2
3	67.3	61.9	64.5	68.9
4	68.5	62.7	65.5	69.8
5	68.9	63.0	65.8	70.1

The results are summarized in Table 6.2, which presents the overall performance of the model after training, this time on the testing set.

Table 6.2 – Final Model Performance - Test Set

<i>Epoch</i>	<i>Precision (%)</i>	<i>Recall (%)</i>	<i>F1 Score (%)</i>	<i>Accuracy (%)</i>
Final	68.7	62.9	65.7	69.8

Although the scores were lower than the ones found by Matos, Rodrigues and Teixeira (2024) when using a similar approach, this result is unsurprising. The lack of annotated data, which led to experimental methods in annotation, and the adherence to default values when training the model are the likely culprits for the decreased metrics.

Following up with more results, Table 6.3 displays the same metrics but showcases every NER category. We shortened some categories’ names for clarity, but their full description can be seen in Table 5.4. This table shows the categories PER and DATE annotated by the pre-trained model, FLAIR, performed above all others. We expected

this result, as it further suggest that our annotation method may have hurt performance. We understand that the use of these two categories inflated the overall performance results of the model, but feel like they were important to include in the tool we aimed to build. We can also note that the "Identificador" categories which comprise shorter, more objective, terms and entities, generally performed better. "Comunicações" and "Direito Penal e Processual Penal" obtained the lowest scores, but still not far from the average. The standard deviation of the F1 score was of approximately 3.84.

Although the model's performance falls short of state-of-the-art benchmarks, the results indicate that it still holds potential for practical applications, particularly in contexts like our web tool.

Table 6.3 – Performance by NER Category (5th Epoch)

<i>NER Category</i>	<i>Precision (%)</i>	<i>Recall (%)</i>	<i>F1 Score (%)</i>	<i>Accuracy (%)</i>
PER	79.3	74.8	76.9	82.1
DATE	77.5	72.1	74.7	80.6
Direito Constitucional	66.1	60.5	63.2	68.7
Processo Legislativo e Atuação Parlamentar	64.8	59.3	62.0	67.9
Administração Pública	65.2	60.0	62.5	68.3
Política, Partidos e Eleições	63.9	58.6	61.1	66.5
Identificador de Acidentes, Localidades e Nomes Geográficos	68.4	63.7	65.9	70.4
Identificador de Conselhos, Entidades e Organizações	67.8	62.9	65.2	70.0
Defesa e Segurança	62.3	56.8	59.4	64.7
Direito Civil e Processual Civil	61.5	55.6	58.4	63.9
Identificador de Programas, Planos, Projetos e Sistemas	67.1	62.3	64.6	69.8
Saúde	63.5	58.4	60.8	66.9
Cidades e Desenvolvimento Urbano	64.1	58.9	61.4	67.4
Trabalho e Emprego	62.9	57.3	60.0	65.8
Economia	63.8	58.1	60.8	67.2
Comunicações	60.3	54.7	57.4	62.9
Educação	62.5	57.4	59.8	65.9
Ciência, Tecnologia e Inovação	61.9	56.6	59.1	64.6
Relações Internacionais e Comércio Exterior	61.3	56.4	58.7	64.1
Finanças Públicas e Orçamento	63.0	58.0	60.4	66.4
Indústria, Comércio e Serviços	62.7	57.2	59.8	65.7
Direito Penal e Processual Penal	60.8	55.3	57.9	63.4
Meio Ambiente e Desenvolvimento Sustentável	61.6	56.3	58.8	64.8
Identificador de Animais	68.7	64.1	66.3	70.7
Identificador de Cargos, Ocupações e Categoria Profissional	69.2	64.5	66.8	71.1
Viação, Transporte e Mobilidade	61.0	55.7	58.3	63.9
Identificador de Partidos Políticos	70.1	65.4	67.7	71.9
Identificador de Materiais e Produtos	68.4	63.5	65.8	70.3
Arquitetura e Urbanismo	62.1	56.9	59.4	64.9
Arte, Cultura e Religião	60.9	55.5	58.0	63.7
Direitos Humanos e Minorias	63.7	58.3	60.9	66.9
Sociologia	61.2	55.9	58.4	64.2



## 7 PROOF-OF-CONCEPT VISUALIZATION TOOL

This chapter describes the proof-of-concept visualization tool developed to interact with the entities identified by the NER model. The tool allows users to explore the political discourse data in a structured and accessible way, focusing on the relationships between entities.

We designed the visualization tool to provide an intuitive interface for exploring the annotated political discourse data. Users can search for specific entities, filter results by NER categories, and visualize the relationships between different entities over time. We built the tool using modern web standards to ensure responsiveness and ease of use. To achieve this, we developed the interface with HTML, CSS, and JavaScript, utilizing the most recent versions where applicable. The Bootstrap<sup>1</sup> framework was also incorporated for consistent styling and layout, enabling us to create a simple yet user-friendly, responsive application that adapts well to different screen sizes and devices. To separate and differentiate the functionalities of our application, we developed two distinct web interfaces, one for visualizing speeches directly from our initial corpus, and another to interact with our model through a NER application.

### 7.1 Speech Viewer

The first, referred to as the "Speech Viewer" and shown in Figure 7.1, is a straightforward corpus navigation tool. It uses the data and metadata extracted from the BCoD after our process of cleansing and organizing it into a JSON file. Users can filter entries from the original, pre-annotation corpus using various criteria, including the speaker's name, political party, state, date, and the content of the speeches. These filters can be applied independently or in combination. The results are displayed below the filters in chronological order and can be seen in Figure 7.2.

### 7.2 NER highlight tool

Our second interface utilizes the results obtained from training and serves as a practical demonstration of the NER model developed in the previous chapter. Users can

---

<sup>1</sup><https://getbootstrap.com/>

# Speech Viewer

## 24 years of Congressional Speeches

Filter Chamber of Deputies speeches by name, party, UF, date or content

**Name**

**Party**                      **State (UF)**

**Day**                              **Month**                              **Year**




**Sentences**

Figure 7.1 – Speech Viewer filtering interface

input political texts or speeches into a text field. Upon submission, the tool processes the text using our model and highlights identified entities directly within the text through NER. Each highlighted entity is color-coded, and when hovered over, it displays its corresponding category. This provides a simple and intuitive way to visualize our results, proving that the model can effectively tag domain-specific entities in unseen text. Figures 7.3 and 7.4 show the interface before any prompt, and after processing an unseen speech from the Federal Senate, respectively.

These proof-of-concept interfaces help us clearly visualize our results, while showcasing potential future applications.



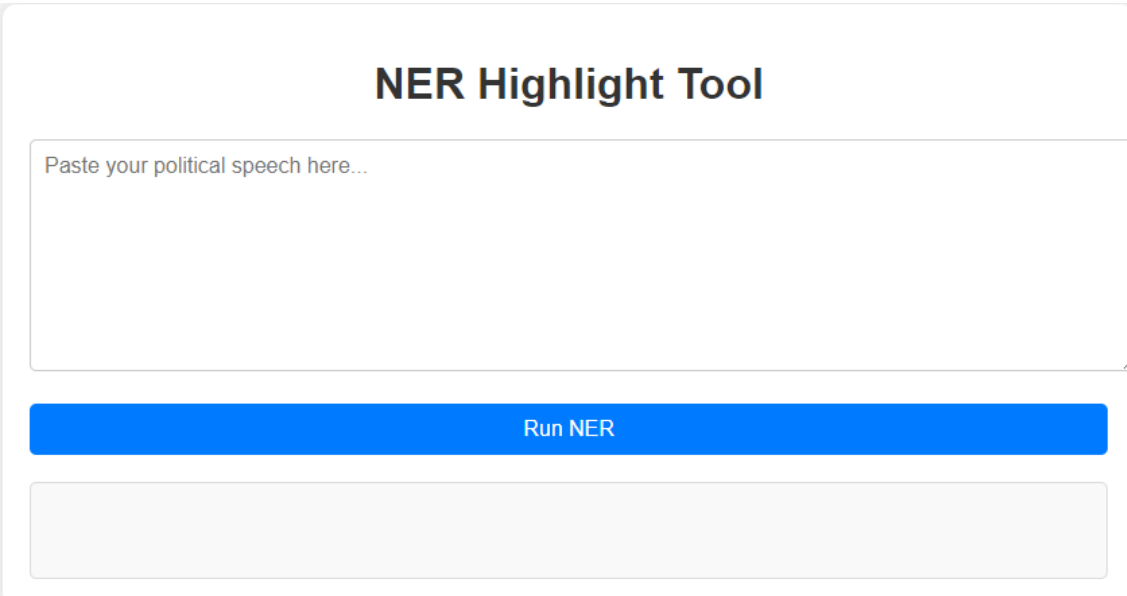
**Search Results**  
4241 results found

**DARCÍSIO PERONDI (PMDB-RS)**  
**Date:** 13/12/2000 19:52:00

**Sentences:**

- O SR. DARCÍSIO PERONDI (Bloco/PMDB-RS.
- Para emitir parecer.
- Sem revisão do orador.)
- - Sr. Presidente, na forma do parecer do Deputado Relator, o Projeto de Lei nº 2.685, de 2000, é constitucional, adequado financeira e orçamentariamente e fará bem ao País.

Figure 7.2 – First result from a Speech Viewer search after filtering by party(PMDB) and UF(RS)



**NER Highlight Tool**

Paste your political speech here...

**Run NER**

Figure 7.3 – NER Highlight tool

## NER Highlight Tool

E, a partir da manifestação de V. Exa., a quem realmente eu agradeço e com a qual fico bastante lisonjeado, de fato, nós temos que ter a compreensão – e é a fala também do Senador Randolfe Rodrigues nesse sentido – do nosso papel e do nosso compromisso em relação a um tema que optamos neste Congresso Nacional para o Brasil e que é de nossa responsabilidade resolver, que é o tema da desoneração da folha de pagamento de 17 setores, prorrogada até 2027, e da desoneração da folha, com a redução de alíquota de municípios até 156 mil habitantes, um projeto de autoria do Senado, do Senador Efraim Filho, relatado pelo Senador Angelo Coronel. Houve toda uma celeuma em torno desse projeto, inclusive com edição de medida provisória, depois com judicialização desse tema. E houve, então, ao final, um acordo entre Legislativo e Executivo em torno da desoneração, que importa na reoneração, a partir de 2025, e com a manutenção, em 2024, da desoneração de 17 setores e da redução de alíquota dos municípios, mas, para isso, o Congresso Nacional precisa contribuir com o Poder Executivo e com o Ministério da Fazenda para se encontrar a compensação financeira e orçamentária para essa desoneração. Isso é algo, realmente, em que, agora, sob a relatoria de V. Exa. do projeto que materializa esse acordo entre Executivo, Legislativo, municípios, 17 setores, nós precisamos encontrar essa fonte de compensação.

Run NER

E, a partir da manifestação de **V. Exa.**, a quem realmente eu agradeço e com a qual fico bastante lisonjeado, de fato, nós temos que ter a compreensão – e é a fala também do **Senador Randolfe Rodrigues** nesse sentido – do nosso papel e do nosso compromisso em relação a um tema que optamos neste **Congresso Nacional** para o **Brasil** e que é de nossa responsabilidade resolver, que é o tema da desoneração da folha de pagamento de **17 setores**, prorrogada até **2027**, e da desoneração da folha, com a redução de alíquota de municípios até **156 mil habitantes**, um projeto de autoria do **Senado**, do **Senador Efraim Filho**, relatado pelo **Senador Angelo Coronel**. Houve toda uma celeuma em torno desse projeto, inclusive com edição de **medida provisória**, depois com judicialização desse tema. E houve, então, ao final, um acordo entre **Legislativo** e **Executivo** em torno da desoneração, que importa na reoneração, a partir de **2025**, e com a manutenção, em **2024**, da desoneração de **17 setores** e da redução de alíquota dos municípios, mas, para isso, o **Congresso Nacional** precisa contribuir com o **Poder Executivo** e com o **Ministério da Fazenda** para se encontrar a compensação financeira e orçamentária para essa desoneração. Isso é algo, realmente, em que, agora, sob a relatoria de **V. Exa.** do projeto que materializa esse acordo entre **Executivo**, **Legislativo**, municípios, **17 setores**, nós precisamos encontrar essa fonte de compensação.

Figure 7.4 – NER Highlight tool with a processed input and its output; the cursor hovers over a named entity exposing its category

## 8 CONCLUSION

This study explored applying NLP techniques to analyzing Brazilian political discourse, focusing on developing a comprehensive speech corpus, training a NER model, experimenting with dictionary-based annotation, and creating a proof-of-concept visualization tool.

The speech corpus we built provided a rich dataset, capturing Brazil's linguistic diversity of political communication. The experiment with distant supervision and dictionary-based annotation, while helpful in enhancing the NER process, revealed the limitations of static lexicons in adapting to evolving political language, especially considering more modern alternatives. Nonetheless, it suggested the potential benefits of hybrid approaches that combine machine learning with domain-specific dictionaries.

Training a tailored NER model on this corpus demonstrated moderate success in accurately identifying key entities such as persons, organizations, legislative terms, and potential topic subjects.

We also developed a proof-of-concept tool to visualize the relationships between entities and topics within the corpus, highlighting the practical applications of the NER model. This tool shows how NLP and data visualization can make political discourse more accessible and understandable.

While this study provided a solid starting point for many academic and practical uses, it also stretched itself too thin by only touching the surface of the techniques applied within it. Future work could expand the corpus to include additional political texts, such as the speeches from the Federal Senate. A subset of the dataset could be manually annotated to open up new possibilities in automatic annotation, leveraging newer techniques such as active learning. Self-training and in-context learning techniques could also arrive at better results. We also intend to expand the visualization tool to support a broader range of uses and offer new insights. Exploring hybrid annotation techniques and applying these methods to other domains or languages could provide valuable perspectives as well.

In conclusion, this research has demonstrated the potential of NLP in transforming political discourse analysis, contributing to greater transparency and engagement in legislative processes.

## REFERENCES

AKBIK, A. et al. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: **Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**. [S.l.: s.n.], 2019. p. 54–59.

ALBUQUERQUE, H. et al. Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In: PINHEIRO, V. et al. (Ed.). **Computational Processing of the Portuguese Language, PROPOR 2022**. Springer, Cham, 2022. (Lecture Notes in Computer Science, v. 13208). Available from Internet: <[https://doi.org/10.1007/978-3-030-98305-5\\_1](https://doi.org/10.1007/978-3-030-98305-5_1)>.

ALBUQUERQUE, H. O. et al. Named entity recognition: A survey for the portuguese language. **Procesamiento del Lenguaje Natural**, n. 70, p. 171–185, March 2023. Received 16-12-2022, revised 31-01-2023, accepted 09-02-2023.

ARAÚJO, P. Luz de et al. Lener-br: A dataset for named entity recognition in brazilian legal text. In: VILLAVICENCIO, A. et al. (Ed.). **Computational Processing of the Portuguese Language, PROPOR 2018**. Springer, Cham, 2018. (Lecture Notes in Computer Science, v. 11122). Available from Internet: <[https://doi.org/10.1007/978-3-319-99722-3\\_32](https://doi.org/10.1007/978-3-319-99722-3_32)>.

ASSOGBA, Y. et al. Many bills: engaging citizens through visualizations of congressional legislation. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2011. (CHI '11), p. 433–442. ISBN 9781450302289. Available from Internet: <<https://doi.org/10.1145/1978942.1979004>>.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006.

BROWN, T. et al. Language models are few-shot learners. **Advances in Neural Information Processing Systems**, v. 33, p. 1877–1901, 2020.

CABOT, P.-L. H.; NAVIGLI, R. REBEL: Relation extraction by end-to-end language generation. In: **Findings of the Association for Computational Linguistics: EMNLP 2021**. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 2370–2381. Available from Internet: <<https://aclanthology.org/2021.findings-emnlp.204>>.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. **Journal of Machine Learning Research**, v. 12, p. 2493–2537, 2011.

CUNNINGHAM, H. et al. A framework and graphical development environment for robust nlp tools and applications. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. [S.l.], 2002. p. 168–175.

DADOS ABERTOS. **Dados Abertos da Câmara dos Deputados**. 2024. Accessed: 2024-08-08. Available from Internet: <<https://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo>>.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

GARCIA, E. A. S. et al. RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In: GAMALLO, P. et al. (Ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 374–383. Available from Internet: <<https://aclanthology.org/2024.propor-1.38>>.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**. [S.l.: s.n.], 1996. p. 466–471.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 2nd. ed. [S.l.]: Pearson, 2009.

LAMPLE, G. et al. Neural architectures for named entity recognition. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.: s.n.], 2016. p. 260–270.

LEVENSHTEIN, V. I. et al. Binary codes capable of correcting deletions, insertions, and reversals. In: SOVIET UNION. **Soviet physics doklady**. [S.l.], 1966. v. 10, n. 8, p. 707–710.

LI, J. et al. A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, n. 1, p. 50–70, 2022.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2008.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. [S.l.]: MIT Press, 1999. 575–580 p. ISBN 9780262133609.

MATOS, E.; RODRIGUES, M.; TEIXEIRA, A. Towards the automatic creation of NER systems for new domains. In: GAMALLO, P. et al. (Ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 218–227. Available from Internet: <<https://aclanthology.org/2024.propor-1.22>>.

MINTZ, M. et al. Distant supervision for relation extraction without labeled data. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP**. [S.l.], 2009. p. 1003–1011.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Linguisticae Investigationes**, John Benjamins, v. 30, n. 1, p. 3–26, 2007.

NAVAREZI, L. et al. Entity extraction from portuguese legal documents using distant supervision. In: PINHEIRO, V. et al. (Ed.). **Computational Processing of the Portuguese Language, PROPOR 2022**. Springer, Cham, 2022. (Lecture Notes in Computer Science, v. 13208). Available from Internet: <[https://doi.org/10.1007/978-3-030-98305-5\\_16](https://doi.org/10.1007/978-3-030-98305-5_16)>.

NIKLAUS, J. et al. **MultiLegalPile: A 689GB Multilingual Legal Corpus**. 2024. Available from Internet: <<https://arxiv.org/abs/2306.02069>>.

NOTHMAN, J. et al. Learning multilingual named entity recognition from wikipedia. **Artificial Intelligence**, v. 194, p. 151–175, 2013.

NUNES, R. et al. **Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning**. 2024.

NUNES, R. O. Work completion of graduation, **A classification approach for estimating subjects of bills in the Brazilian Chamber of Deputies**. 2023. Graduation Work, Advisor: Carla Maria Dal Sasso Freitas, Co-advisor: Dennis Giovanni Balreira. Available from Internet: <<http://hdl.handle.net/10183/267612>>.

NUNES, R. O. et al. A named entity recognition approach for Portuguese legislative texts using self-learning. In: GAMALLO, P. et al. (Ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 290–300. Available from Internet: <<https://aclanthology.org/2024.propor-1.30>>.

POLO, F. M. et al. Legalnlp-natural language processing methods for the brazilian legal language. In: SBC. **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2021. p. 763–774.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. In: **Third Workshop on Very Large Corpora**. [S.l.: s.n.], 1995. p. 82–94.

RODRIGUES, J. et al. Advancing neural encoding of portuguese with transformer albertina pt-\*. In: \_\_\_\_\_. **Progress in Artificial Intelligence**. Springer Nature Switzerland, 2023. p. 441–453. ISBN 9783031490088. Available from Internet: <[http://dx.doi.org/10.1007/978-3-031-49008-8\\_35](http://dx.doi.org/10.1007/978-3-031-49008-8_35)>.

SASAKI, Y. The truth of the f-measure. In: **Proceedings of the Workshop on Performance Metrics for Intelligent Systems (PerMIS)**. [S.l.: s.n.], 2007. p. 1–6.

SIQUEIRA, F. et al. Ulysses tesemõ: A new large corpus for brazilian legal and governmental domain. **Language Resources and Evaluation**, 2024. Available from Internet: <<https://doi.org/10.1007/s10579-024-09762-8>>.

SOUSA, A. W.; FABRO, M. **Iudicium Textum Dataset: Uma Base de Textos Jurídicos para NLP**. 2019.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020.

VASWANI, A. et al. Attention is all you need. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. p. 5998–6008.



VIANNA, D.; MOURA, E. S. de. Organizing portuguese legal documents through topic discovery. In: **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2022. (SIGIR '22), p. 3388–3392. ISBN 9781450387323. Available from Internet: <<https://doi.org/10.1145/3477495.3536329>>.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2020.

ZANUZ, L.; RIGO, S. Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In: PINHEIRO, V. et al. (Ed.). **Computational Processing of the Portuguese Language, PROPOR 2022**. Springer, Cham, 2022. (Lecture Notes in Computer Science, v. 13208). Available from Internet: <[https://doi.org/10.1007/978-3-030-98305-5\\_21](https://doi.org/10.1007/978-3-030-98305-5_21)>.

ZHONG, H. et al. How does NLP benefit legal system: A summary of legal artificial intelligence. In: JURAFSKY, D. et al. (Ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 5218–5230. Available from Internet: <<https://aclanthology.org/2020.acl-main.466>>.