A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning

Rafael O. Nunes, Dennis G. Balreira, André S. Spritzer and **Carla M. D. S. Freitas**
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
{ronunes,dgbalreira,spritzer,carla}@inf.ufrgs.br

Abstract

Even if technology has made legislative documents more accessible, they are often written in jargon that makes them hard to understand for ordinary citizens, researchers, journalists, and even lawmakers. However, recent advances in Natural Language Processing can help bridge this gap. In this paper, we present the self-learning fine-tuning of a BERT model designed for Named Entity Recognition (NER) using active sampling. Our study focuses on legislative documents written in Brazilian Portuguese, using the labeled data from the UlyssesNER-Br corpus and the unlabeled data from the bill's summary of the Brazilian Chamber of Deputies. We achieved F1-scores of 86.70 ± 2.28 around the cross-validation and a final result of 90%, establishing the efficacy of BERTimbau with self-learning in performing Named Entity Recognition for legislative texts encompassing various categories. Our findings highlight its significant potential for enhancing legislative text analysis tasks.

1 Introduction

Democratic politics is about much more than elections. It involves constant vigilance by an informed public and an active civil society that holds politicians into account. Transparency, defined as the availability of information to the public about organizational activities and decisions, fosters accountability by letting citizens scrutinize and evaluate the actions of public officials (Heald, 2006). Open government initiatives increase transparency by giving citizens access to documents and data related to official and public activities (Lathrop and Ruma, 2010). Mere access to information, however, is often not enough to allow for proper public oversight, as researchers, journalists, citizens, and even policymakers may find themselves overwhelmed by the myriad of bills, amendments, and documents written in obtuse jargon that make it a daunting task to analyze and understand political activities and legislative work, in particular.

A significant part of Natural Language Processing (NLP), Named Entity Recognition (NER), involves identifying named entities in texts and classifying them into predefined categories such as people, organizations, locations, and more. NER can aid document comprehension by identifying and emphasizing these domain-specific terms. This enables a comprehensive overview of documents by identifying significant terms and specific domain classes, like in the works by Sultanum et al. (2018) and Nunes et al. (2019). Additionally, NER indirectly contributes to comprehension through enrichment processes by facilitating the indexing of dictionary information. In turn, this helps provide contextual explanations and synonyms. NER can also be an initial step in other NLP tasks, such as constructing domain-specific knowledge graphs and coreference resolution (Kalamkar et al., 2022; Cohen and Hersh, 2005).

This paper explores how NER can be improved using semi-supervised techniques. Our approach consists of a self-learning strategy to fine-tune a BERT model designed for Named Entity Recognition (NER), using legislative documents written in Brazilian Portuguese as a case study. For training and evaluation, it relies on UlyssesNER-Br (Albuquerque et al., 2022), a corpus of bills and legislative consultations from the lower house of the Brazilian national legislature (the Chamber of Deputies) that was explicitly designed for NER. Our paper's main contributions are: (i) an approach for the NER task for Brazilian Portuguese text using a self-learning and active sampling strategy, (ii) its resulting BERT NER classifier in Brazilian Portuguese legislative text[1] and (iii) a comprehensive discussion of NER in the legislative domain,

[1] https://huggingface.co/ronunes/bertimbau-base-ulyssesner_br-bcod-self_learning

including how classes handle the incorporation of additional data from self-learning sourced from unlabeled public data.

2 Related Work

This section explores the literature on NER and the use of unlabeled data in the training loop, specifically for NLP models. Subsection 2.1 presents relevant studies on NER over time, focusing in particular on the use of the Portuguese language legal domain. Subsection 2.2 explores unlabeled data as a source of data augmentation, focusing specifically on self-learning and active learning techniques.

2.1 Named Entity Recognition

More than a decade ago, Dozier et al. (Dozier et al., 2010) proposed one of the most well-known legal NER systems with data from United States courts, mainly consisting of depositions, pleadings, and case law. The authors used three methods for the NER task: *lookup*, *pattern rules*, and *statistical models*, which could also be combined into hybrid systems. They also introduced five taggers, including *jurisdiction*, *court*, *title*, *doctype*, and *judge*. Using a similar approach, other works have explored NER in legal domains for other languages, including German (Darji et al., 2023; Glaser et al., 2018; Leitner et al., 2019), Spanish (Badji, 2018), Greek (Angelidis et al., 2018), and Romanian (Păiș et al., 2021). Regarding the Portuguese language, Dos Santos and Guimarães (Santos and Guimaraes, 2015) proposed the first NER system using the CharWNN architecture, which employs a multilayer perceptron network (Santos and Guimaraes, 2015). Most works on NER for general domain Portuguese text evaluate their models using the HAREM corpus (Santos et al., 2006), which comprises documents from several fields.

Concerning Portuguese language legal corpora, two recent works introduced Portuguese language datasets for NER in legislative texts (Luz de Araujo et al., 2018; Albuquerque et al., 2022). Araujo et al. (Luz de Araujo et al., 2018) created the first dataset for NER in Brazilian legal text, called LeNER-Br, by gathering 66 legal documents from Brazilian courts and training a long short-term memory (LSTM) conditional random field (CRF) (LSTM-CRF) model (Lample et al., 2016), which resulted in a total F1-score of around 92% for token classification and 86% for entity classification. Albuquerque and colleagues (Albuquerque et al., 2022), in turn, proposed a corpus for NER called UlyssesNER-Br consisting of bills and legislative consultations from the Brazilian Chamber of Deputies (BCoD), with 18 types of entities distributed over seven categories. To validate the corpus, the authors implemented CRF and Hidden Markov Model models, achieving an F1-score of around 80% in the analysis by categories and 81% in the analysis by types.

Similarly, three recent works explore specific legal contexts. Collovini et al. (Collovini et al., 2019) manually annotated a police dataset using testimony, statement, and interrogatory texts, with 916 named entities of the "Person" category achieving an F1-Score of 89% using BiLSTM-CRF-ELMo. Brito et al. (Brito et al., 2023) developed the CDJUR-BR, a Brazilian Judiciary corpus with specific domain entities: *prova* (i.e., evidence), *pena* (i.e., punishment), *sentença* (i.e., sentence), and *norma* (i.e., norm). They achieved an F1-Macro of 0.58 using a BERT model (Devlin et al., 2018). Finally, Correia et al. (Correia et al., 2022) developed a corpus with fine-grained and coarse-grained legal entities from Brazilian Supreme Court (STF) documents annotated by 76 law students. With a BiLSTM-CRF, they obtained a 93% F1-Weighted Score for coarse-grained entities. For fine-grained entities, their results were generally around 70% to 90%.

Building upon advances in research on fine-tuning methodologies and model enhancements, the work proposed by Bonifacio et al. (Bonifacio et al., 2020) investigated the impact of fine-tuning language models on a large intradomain corpus of unlabeled text for NER. Experimental findings revealed that fine-tuning the models on intradomain text significantly improved NER performance, particularly for the BERT model, which achieved state-of-the-art results on the LeNER-Br corpus of Brazilian legal text. Zanus and Rigo (Zanuz and Rigo, 2022) introduced the first fine-tuned BERT models exclusively trained on Brazilian Portuguese for legal NER, achieving new state-of-the-art results on the LeNER-Br dataset.

2.2 Self-Learning and Active Learning in Training

Data augmentation has been widely employed in various NLP tasks to enhance model performance (Li et al., 2022; Feng et al., 2021; Anaby-Tavor et al., 2020). An alternative method to improve the quality of a training corpus is utilizing unla-

beled data. In cases where a substantial amount of unlabeled data is available, semi-supervised techniques are used (Li et al., 2022; Feng et al., 2021), including self-learning, active learning, and their variations. These techniques have been shown to improve the results of models in the tasks such as classification (Sha et al., 2022; Alves-Pinto et al., 2021; Mekala and Shang, 2020; Meng et al., 2020; Dong and de Melo, 2019; Dupre et al., 2019) and NER (Gao et al., 2021; Neto and Faleiros, 2021; Helwe and Elbassuoni, 2019; Clark et al., 2018; Tran et al., 2017; Chen et al., 2015).

The self-learning approach involved leveraging a labeled corpus to train a *professor* model that was employed to predict the classes of unlabeled data, which were subsequently used to train a *student* model (Dupre et al., 2019). In some self-learning strategies, the student model could serve as a *professor* for the next iteration (Dupre et al., 2019). Although this represented a conventional self-learning approach, alternative methods, such as weak labels, ensemble models, and modifications to the loss function, could also be employed. Active learning followed a similar philosophy but incorporated querying methods to select instances of interest for manual annotation. These annotated instances were then used to retrain the model iteratively.

To the best of our knowledge, our work is the first to propose an NER method for Brazilian Portuguese legislative text that used a legislative corpus and improved results through a self-learning strategy.

3 Methodology

This section describes the methodology used in this study. We begin with a brief overview of the Ulysses-NER-Br corpus, followed by a description of the unlabeled corpus, consisting of summaries of bills from the Brazilian Chamber of Deputies from 1991 to 2022. This unlabeled corpus was leveraged in the active learning phase to augment the training data. We also detail the pre-processing steps applied to the data. Subsequently, we explain our approach to corpus division for training and validation, introduce the transformer models utilized, and elucidate the adopted self-learning strategy.

3.1 Legislative NER Corpus

UlyssesNER-Br (Albuquerque et al., 2022) is a Brazilian Portuguese corpus that contains two sources of information and is divided into two corpora for each reference source. The first corpus contained 9,526 sentences from 150 bills (*Projetos de Lei - PL*) of the Brazilian Chamber of Deputies, and the second had 790 sentences from legislative consultations (*solicitações de trabalho - ST*).

The UlyssesNER-Br corpus is divided into two types of entities: *category* and *type*. Categories comprise five traditional entities (Albuquerque et al., 2022): "PESSOA" (*person*), "DATA" (*date*), "ORGANIZAÇÃO" (*organization*), "EVENTO" (*event*), and "LOCALIZAÇÃO" (*location*). Additionally, they include "FUNDAMENTO" (*grounds*) and "PRODUTODELEI" (*legal product*) as references to legislative entities. Types, in turn, are particularizations of categories, e.g., "PRODUTOsistema" (*system product*), "PRODUTOprograma" (*program product*), and "PRODUTOoutros" (*other product*) as particularizations of the "PRODUTODELEI" category.

Unfortunately, the corpus with legislative consultations is not publicly available, as it consists of internal information from the Chamber of Deputies[2], which UlyssesNER-Br's authors were not allowed to share. Therefore, we used only the corpus with the bills information in this study.

In our work, we used only category entities for the self-learning process since the authors pointed out that their results with categories and types did not show significant differences. Thus, categories showed a more straightforward and robust solution to the model's learning (Albuquerque et al., 2022). Table 1 indicates the number of examples of any category in the corpus for training, validation, and testing. We point out that the frequency calculated in the tables does not refer to token frequency but to the frequency of the complete entity, which we use to calculate the metrics in Section 4.

3.2 BCoD Bills Summary

Summary bills are obtained from the BCoD API[3] spanning from 1991 to 2022. These summaries are then segmented into sentences using the regular expression ". (?=[A-Za-z])" to identify periods followed by letters, splitting the text into sentences. We chose this regular expression because the legislative text domain includes constructions like "Art. 123", where the period is part of the article's name and not indicative of the end of a

[2]https://github.com/Convenio-Camara-dos-Deputados/ulyssesner-br-propor/tree/main/Corpora

[3]https://dadosabertos.camara.leg.br/swagger/api.html

Entity type	Train	Validation	Test
DATA	433	72	98
PESSOA	628	114	119
ORGANIZACAO	435	81	94
FUNDAMENTO	490	107	124
LOCAL	369	145	101
PRODUTODELEI	230	46	54
EVENTO	9	5	9
Total	2,594	570	599

Table 1: Frequency of named entities in UlyssesNER-Br for each category.

sentence.

This process produced 428,573 sentences, with an average word count of 32.52 and a standard deviation of 42.64. While obtaining sentences, we excluded the ones already present in the UlyssesNER-Br corpus, making sure to include only distinct sentences to prevent overfitting. All these sentences lacked NER information.

3.3 Data Preparation

The UlyssesNER-Br corpus is available in text format (.txt) on github[4]. The corpus is divided into separate files for training, validation, and testing. Each token in a sentence is split into different lines, with sentences separated by a line containing a "\n". Each token within a sentence also has an entity tag in the format of "B-TAG" or "I-TAG."

To use the corpus for training, we preprocessed the TXT files into JSON files and converted them into a Hugging Face Dataset[5]. We started this process by identifying tokens belonging to the same sentence and organizing them into lists along with their corresponding tags. Instead of storing string-based tags, we converted them to decimal values for training. We obtained decimal values using a dictionary of entity tags and indices.

Then, we iterated the preprocessing step for all sentences, obtaining two lists: (i) one containing all sentences and (ii) another with all sentence tags. We saved them both into a unique JSON file with "sentences" and "ner_tags" keys. We concatenated these training, validation, and test files into the same TXT file to generate a unique corpus and obtain a unique JSON with the preprocessed corpus.

[4]https://github.com/ulysses-camara/ulysses-ner-br/tree/main/annotated-corpora/PL_corpus_conll

[5]https://huggingface.co/docs/datasets/index

3.4 Corpus Division

To train and validate our approach, we used a handout 5-fold cross-validation division method inspired by the original UlyssesNER-Br paper (Albuquerque et al., 2022). The key distinction is that we used stratified division in the handout and cross-validation phases, a modification influenced by Sechidis et al. (2011)'s approach. We also introduced an additional preprocessing step that generates a list equivalent in size to the number of possible distinct entities. Within this list, each position is assigned a flag with a value of one if the corresponding entity is present in the sentence and zero otherwise. This modification enabled us to stratify the division based on the presence of each entity.

However, it is essential to highlight the significance of this stratification step, mainly because of the substantial class imbalance in the original corpus. This imbalance is evident when examining examples from minority and majority classes, such as "Eventos" with only 23 instances, and "Pessoa" with 861 instances.

3.5 Models

We briefly describe two of the most prominent existing transformer models for the Portuguese language: (i) BERTimbau and (ii) SBERT. We use the BERTimbau model for the NER task and the SBERT model for the active sampling.

BERTimbau is a pre-trained BERT model fine-tuned to Brazilian Portuguese (Souza et al., 2020). To the best of our knowledge, BERTimbau is the state-of-the-art in Named Entity Recognition, sentence textual similarity, and recognition of textual entailment in Brazilian Portuguese. This work uses the base version available at Hugging Face Hub[6] to train the classifier models.

SBERT (Reimers and Gurevych, 2019) is a modification of BERT models that uses siamese and triplet networks to obtain contextual embeddings relative to a whole sentence. To generate embeddings to Portuguese text, we used the multilingual version of SBERT (Reimers and Gurevych, 2020) that is available at Hugging Face Hub[7] in the active sampling technique.

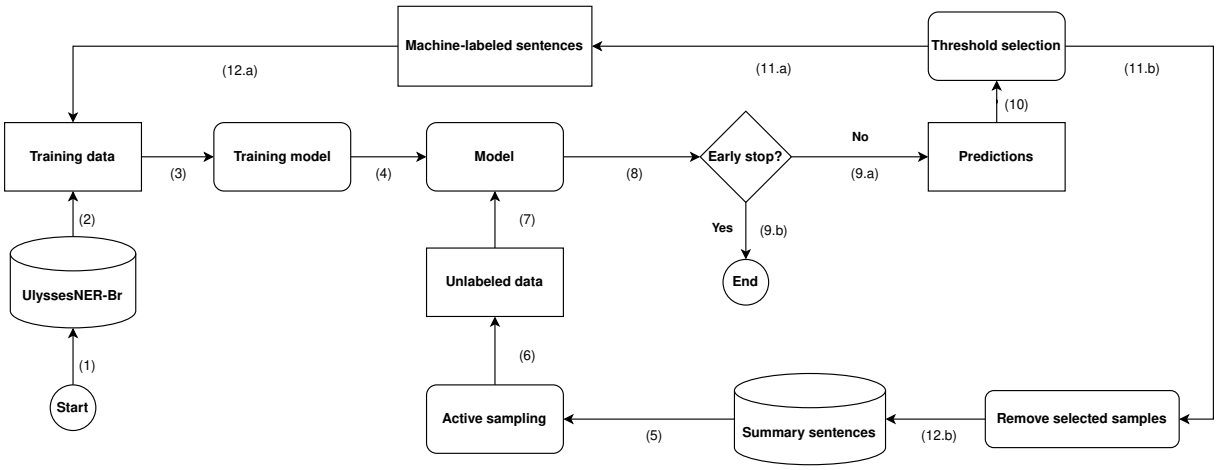


Figure 1: Self-learning pipeline.

3.6 Self-learning

Our self-learning pipeline is shown in Figure 1. We followed the pipeline in each iteration of the cross-validation and handout. Our pipeline starts with the trainer of the first classifier using training data (1 to 4). Once our summary corpus contains a large amount of data, the sampling technique is used to optimize the training time throughout the self-learning iterations (5 to 6).

Inspired by Sha et al. (2022)'s dynamic sampling, our sampling technique begins with random sampling from the summary corpus, producing $N\%$ of the unlabeled data with a minimum of 2,000 samples. Next, we use diversity-based sampling (Tran et al., 2017; Chen et al., 2015) by applying cosine similarity to the sampled data relative to the training data. Subsequently, we obtain the most dissimilar samples in a total of $K\%$ of the total data sampled, with a minimum of 1,000, which is used in the model to predict the NER tags for each sentence. The embeddings used to calculate cosine similarity are generated by SBERT because of its ability to recognize important sentence features.

After the active sampling step, we apply the NER classifier to each sentence (7 to 9.a). To determine the sentences to be added to the training data, we measure the average prediction confidence of the predicted entities (Gao et al., 2021) (10). If the average confidence is equal to or higher than a threshold, it is used in the training data (11.a to 12.a) and removed from the summary corpus (11.b to 12.b); otherwise, it is retained in the summary corpus and is not used for training. Subsequently, the pipeline restarted using the new training set to train a new model and repeat the entire pipeline.

The pipeline halts when an early stop condition is found based on overall F1. We describe the hyperparameters in Section 4.3. We also implemented an early stop criterion in which no data were added to the training or if the unlabeled set became empty (i.e., all available data were utilized). Occasionally, owing to the random sampling approach, it is possible that randomly selected data may not contain suitable examples for training, resulting in no additions to the training set. In such cases, we implement a waiting criterion that allows for a maximum of W new samplings before terminating the self-learning process. Each of these new samplings uses different random seeds to generate distinct sets, aiming to address potential issues with the initial random selection.

4 Experimental Evaluation

In this section, we present an experimental assessment of the proposed approach. We describe the setup, including the hardware used and the development environment. We also describe our model's hyperparameters, the self-learning training, and the metrics used for evaluation.

4.1 Setup

We used a computer with an Nvidia GeForce RTX 3060 GPU and 32.0 GB of RAM for the training and evaluation of the models and for obtaining the data from the BCoD API[8]. We chose the Python

[6]https://huggingface.co/neuralmind/bert-base-portuguese-cased

[7]https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[8]https://dadosabertos.camara.leg.br/swagger/api.html

3.7.6 programming language because of its variety of libraries for Machine Learning and Natural Language Processing.

4.2 Model Hyperparameters

We calibrated the model based on previous studies (Zanuz and Rigo, 2022; Bonifacio et al., 2020). We used the bert-base-portuguese-cased BERT model (BERTimbau Base) trained with the UlyssesNER-Br PL-corpus for the NER task. We used the HuggingFace Trainer API, which has a maximum sentence length of 512, as well as padding and truncation.

We used the following hyperparameters to build our model: evaluation_strategy = epochs, save_total_limit = 5, learning_rate = 2e-05, weight_decay = 0.01, optimizer = Adam with betas = (0.9,0.999), and epsilon= 1e-08. The remaining non-specified parameters follow the model's default parameters. We also set the save_strategy to epoch and used the overall F1 as the metric chosen for the best model.

4.3 Self-learning Hyperparameters

For active sampling, we used N = 0.05 for the percentage of random samples, K = 0.6 for the percentage of dissimilar samples, and 42 as the first seed for random sampling. We used W = 5 as the maximum number of new random samplings to increase the amount of training data. We also used patience equal to 3 to wait for an increase in the overall F1 around the self-learning iterations. We used a range of values from 0.9 to 0.999 for the average prediction confidence threshold, with intermediary thresholds in between, following a similar approach to Gao et al. (2021).

4.4 Metrics

We used the seqval[9] library to compute metrics. An interesting aspect of this library is that it calculates the results based on the sequence of tags to each entity (starting with a "B-TAG" and the following "I-TAG"); for a complete sentence, it is important to clearly recognize an entity rather than just a specific token.

We computed the following metrics: F1-score, precision, recall, and accuracy (only for the overall case). We chose the F1-score as the main metric for our analyses since we wanted to balance the correct prediction of the positive class and how well this class is predicted.

[9] https://github.com/chakki-works/seqeval

4.5 Handout Stratified K-Fold Cross-Validation

Under a previously outlined approach, we conducted a benchmark study by fine-tuning the BERT model for an NER task using a Portuguese legislative corpus (Albuquerque et al., 2022). Our training started with the initial fine-tuning of the model through stratified 5-fold cross-validation to ascertain the optimal self-learning threshold value, as described in Subsection 4.3. Then, we applied the threshold within each fold, as stated in Section 3.6.

It is noteworthy that self-learning yields metrics for classifiers in each iteration. Instead of relying solely on the final classifier in the pipeline, we adopted a more robust approach by selecting the metrics from the classifier that exhibited the best overall F1-score. Within each fold, the threshold for optimal performance was determined based on the highest F1-score. To establish the best F1-score for the final model, we selected the best F1-score around 5-fold using the average and standard deviation to identify the threshold that consistently produced superior results. We used this threshold during the handout-training phase.

5 Results and Discussion

The selection of the threshold value is a critical aspect of our approach because it plays a pivotal role in determining the overall performance of the NER system. In the cross-validation phase, our evaluation revealed that the threshold value of 0.99 consistently demonstrated superior performance throughout the cross-validation phase, resulting in an F1-score within the 86.70 ± 2.28 range across all the folds. It is important to highlight that thresholds of 0.95 and 0.975 present similar results, as shown in Table 2. Therefore, to choose the threshold between them, we select what has a greater increase in most entities. We based this threshold selection on the best F1-score in the cross-validation, as elaborated in Section 4.5, which proved to be the most effective choice across the five folds. Consequently, the results presented in this section encompass the conclusive metrics acquired with a threshold set at 0.99.

Table 2 shows the impact of self-learning on the final results. This table displays the F1-score for the entity classes. Our approach was able to achieve significantly higher results for most classes, as can be seen in entity "LOCAL", which had an

Threshold	DATA	EVENTO	FUNDAMENTO	LOCAL	ORGANIZACAO	PESSOA	PRODUTODELEI	Overall
0.9	94.49 ± 3.13	48.89 ± 33.41	88.01 ± 2.12	85.54 ± 3.50	82.56 ± 5.95	83.98 ± 4.06	75.11 ± 5.72	85.02 ± 2.45
0.925	94.62 ± 2.65	54.53 ± 32.68	89.34 ± 1.61	85.10 ± 4.34	83.16 ± 4.49	84.40 ± 3.55	71.36 ± 4.14	85.12 ± 2.31
0.95	**95.08 ± 1.89**	49.05 ± 33.86	89.99 ± 2.45	87.49 ± 5.03	**84.82 ± 2.57**	85.55 ± 5.24	76.01 ± 7.18	**86.56 ± 1.99**
0.975	95.08 ± 3.41	50.48 ± 34.02	**90.50 ± 2.54**	85.11 ± 4.25	85.30 ± 5.75	85.75 ± 4.83	75.83 ± 5.94	**86.48 ± 2.91**
0.99	94.77 ± 2.65	**58.10 ± 34.16**	88.60 ± 2.29	**86.46 ± 3.73**	84.89 ± 5.77	**87.48 ± 2.79**	**75.42 ± 4.47**	**86.70 ± 2.28**
0.9975	93.35 ± 2.66	3.64 ± 7.27	88.91 ± 2.24	80.78 ± 2.89	79.12 ± 4.33	87.91 ± 3.32	72.81 ± 5.65	84.16 ± 2.28
0.999	93.10 ± 2.21	20.00 ± 24.49	88.37 ± 1.65	80.87 ± 4.06	79.73 ± 2.94	87.22 ± 3.02	70.74 ± 9.06	83.87 ± 2.36
Standard	94.25 ± 2.69	0.00 ± 0.00	88.59 ± 3.86	78.96 ± 3.90	78.33 ± 4.44	87.77 ± 3.19	70.44 ± 7.40	83.53 ± 2.56

Table 2: Cross-validation results to each threshold with self-learning and the result without self-learning.

Model	Accuracy	Precision	Recall	F1-score
HMM	93.07 ± 0.78	60.45 ± 2.18	30.82 ± 1.81	40.74 ± 1.83
CRF	97.27 ± 0.77	83.42 ± 0.91	70.40 ± 1.54	76.28 ± 1.12
BiLSTM-CRF + Glove	97.66 ± 0.47	80.48 ± 2.69	73.63 ± 2.65	76.89 ± 2.49
BERTimbau	98.30 ± 0.32	80.17 ± 3.67	87.63 ± 1.13	83.53 ± 2.56
BERTimbau + Self-learning	**98.45 ± 0.24**	**85.37 ± 2.91**	**89.02 ± 1.45**	**86.70 ± 2.28**

Table 3: Original results and our results with BERT and self-learning using the threshold of 0.99.

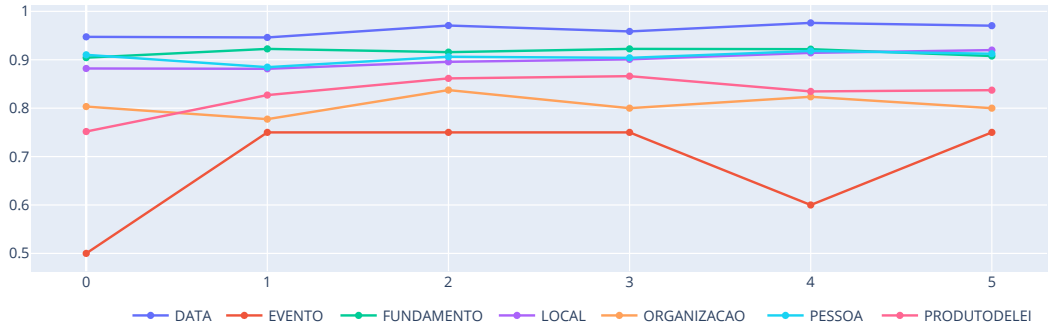
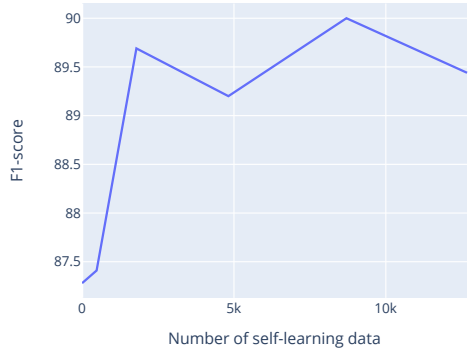
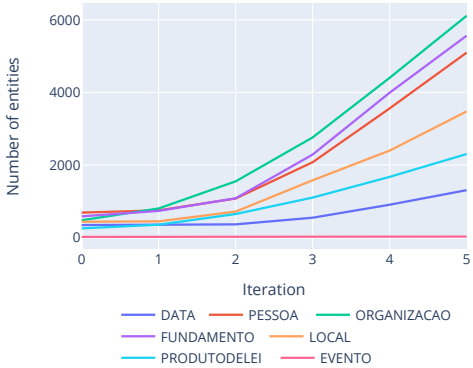


Figure 2: F1-score for each entity over iterations.



(a) Learning curve with the relation between the cumulative number of added data in each iteration and its respective F1-score.



(b) Cumulative number of examples of each entity added in each iteration.

Figure 3: Impact of self-learning in the training phase.

increase of 8% in its F1-score. Similarly, "ORGANIZACAO" has an increase of 6% and decreases the standard deviation, the same as "PRODUTODELEI" with an increase of 5% and a decrease in the standard deviation. It is interesting to highlight "EVENTO", which was not possible to predict with the original data having 0.0 ± 0.0 of the F1-score, and we managed to achieve 58.10 ± 34.16. "DATA". "FUNDAMENTO" and "PESSOA" did not have significant impacts, having only fewer increases on average or fewer decreases in standard deviation.

In the original UlyssesNER-Br paper (Albuquerque et al., 2022), the authors used the corpus to train a Hidden Markov Model (HMM) and a Conditional Random Field (CRF) model. Furthermore, they also used BiLSTM-CRF and the Glove architecture to compare with the results achieved in the work of Luz de Araujo et al. (2018) with the LeNER-Br corpus. In this way, Table 3 demonstrates the higher results, in which only using a BERT model fine-tuned to Portuguese (Souza et al., 2020) could increase the F1-score by 6.64%. However, introducing self-learning emerged as an important factor in increasing the F1-score of 9.81%.

To conduct a more in-depth analysis of the results, we performed a final training on the role data used within the cross-validation and tested it on a previously unutilized dataset. The handout cross-validation approach served a dual purpose: it not only aided in fine-tuning the threshold hyperparameter but also offered a more comprehensive means of validating results with a predefined number of folds. Furthermore, this approach enabled a detailed analysis of the results in the test set, with particular attention to the impact and consequences of each category.

The learning curve for each entity, as depicted in Figure 2, illustrates the significant impact of self-learning on classes, resulting in a noticeable increase in the metric compared with the standard result at iteration zero. However, it is worth noting that some oscillations were observed at specific iterations, possibly owing to incorrectly annotated examples introduced into the training set. Nevertheless, the overall trend demonstrates the robustness of employing self-learning and highlights the influence of the chosen threshold in filtering out a substantial portion of the noisy data.

Similarly, Figure 3a illustrates the learning curve for the cumulative number of sentences added over iterations, focusing on the overall F1-score. By the fourth iteration, we achieved our highest F1-score of 90%, underscoring the positive impact of augmenting the original corpus. This result holds great promise compared to the F1-score of 87.28% obtained using only the BERT model in iteration zero.

Figure 3b shows the increase in the number of examples for each entity during iterations. It should be noted that both classes with more and less data exhibited a considerable increase in the number of examples. Even so, the classes "DATA" and "EVENTO" had the slightest increase. We believe this fact occurred because dates have specific formats, thus being easier to filter noise, and "EVENTO" being the minority class, slightly increasing over the iterations precisely due to its small number of data.

6 Conclusion

This paper presented an NER method with self-learning and active sampling, using Portuguese legislative text from the UlyssesNER-BR corpus as a case study. Our results show that BERTimbau using self-learning achieved an overall average F1-score of 86.70 ± 2.28 around the cross-validation and a final result of 90%, showing strong performance in entity recognition compared to using only BERTimbau and the previous benchmarks. This finding demonstrates the effectiveness of BERTimbau with self-learning for Named Entity Recognition in the legal/legislative domain, highlighting its potential for legal text analysis tasks.

Despite the positive results, our study has some limitations. We only conducted the experiments at the entity category level and did not evaluate how it would work at the type level. As future work, we plan to conduct experiments at the type level and compare the correlations between the levels. We also plan to adopt an ensemble approach of our model with BERTimbau fine-tuned with LeNER-Br corpus[10] using equivalent entities between corpora in the ensemble. Concerning the fine-tuning of the models, we plan to make experiments with the BERT-CRF and BERT-LSTM-CRF versions of BERTimbau available in their official repository[11]. We also plan to perform experiments using other recent Portuguese BERT-like models, such as Albertina (Rodrigues et al., 2023) and LegalBert-pt

[10] https://huggingface.co/Luciano/bertimbau-large-lener_br

[11] https://github.com/neuralmind-ai/portuguese-bert/tree/master/ner_evaluation

(Silveira et al., 2023). Our findings also emphasize the importance of having a diverse and representative dataset for fine-tuning models in specific domains. Further research should focus on expanding the training data, curating new data with experts so that it can be made available for general use, and exploring other pre-training and fine-tuning techniques to improve the performance of NER models in the legislative domain.

Acknowledgements

This study was partially funded by the Brazilian funding agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Hidelberg O Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia FF da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, et al. 2022. Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 3–14. Springer.

Ana Alves-Pinto, Christoph Demus, Michael Spranger, Dirk Labudde, and Eleanor Hobley. 2021. Iterative named entity recognition with conditional random fields. *Applied Sciences*, 12(1):330.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *JURIX*, pages 1–10.

Ines Badji. 2018. *Legal entity extraction with NER systems*. Ph.D. thesis, ETSI_Informatica.

Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 648–662. Springer.

Maurício Brito, Vládia Pinheiro, Vasco Furtado, Joao Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André Câmara Ferreira da Costa, and Raquel Silveira. 2023. Cdjur-br-uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 177–186. SBC.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. 2019. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In *IberLEF@ SEPLN*, pages 390–410.

Fernando A Correia, Alexandre AA Almeida, José Luiz Nunes, Kaline G Santos, Ivar A Hartmann, Felipe A Silva, and Hélio Lopes. 2022. Fine-grained legal entity annotation: A case study on the brazilian supreme court. *Information Processing & Management*, 59(1):102794.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German bert model for legal named entity recognition. *arXiv preprint arXiv:2303.05388*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xin Luna Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6306–6310.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named entity recognition and resolution in legal text*. Springer.

Robert Dupre, Jiri Fajtl, Vasileios Argyriou, and Paolo Remagnino. 2019. Improving dataset volumes and model accuracy with semi-supervised iterative self-learning. *IEEE Transactions on Image Processing*, 29:4337–4348.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Shang Gao, Olivera Kotevska, Alexandre Sorokine, and J Blair Christian. 2021. A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, 16(2):e0246310.

Ingo Glaser, Bernhard Waltl, and Florian Matthes. 2018. Named entity recognition, extraction, and linking in german legal contracts. In *IRIS: Internationales Rechtsinformatik Symposium*, pages 325–334.

David Heald. 2006. *Varieties of Transparency*, volume 1. Oxford University Press on Demand.

Chadi Helwe and Shady Elbassuoni. 2019. Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, 52:197–215.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Daniel Lathrop and Laurel Ruma. 2010. *Open Government: Collaboration, Transparency, and Participation in Practice*. O'Reilly Media, Inc.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings*, pages 272–287. Springer.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 313–323. Springer.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

José Reinaldo CSAVS Neto and Thiago de Paulo Faleiros. 2021. Deep active-self learning applied to named entity recognition. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 405–418. Springer.

Rafael O Nunes, João E Soares, Henrique DP dos Santos, and Renata Vieira. 2019. Meshx-notes: web-system for clinical notes. In *Artificial Intelligence in Health: First International Workshop, AIH 2018, Stockholm, Sweden, July 13-14, 2018, Revised Selected Papers 1*, pages 5–12. Springer.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*.

Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Lele Sha, Yuheng Li, Dragan Gasevic, and Guanliang Chen. 2022. Bigger data or fairer data? augmenting bert via active sampling for educational text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1275–1285.

Raquel Silveira, Caio Ponte, Vitor Almeida, Vládia Pinheiro, and Vasco Furtado. 2023. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Brazilian Conference on Intelligent Systems*, pages 268–282. Springer.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Nicole Sultanum, Devin Singh, Michael Brudno, and Fanny Chevalier. 2018. Doccurate: A curation-based approach for clinical text visualization. *IEEE transactions on visualization and computer graphics*, 25(1):142–151.

Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita, Dinh Tuyen Hoang, and Dosam Hwang. 2017. A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields. *Knowledge-Based Systems*, 132:179–187.

Luciano Zanuz and Sandro José Rigo. 2022. Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 219–229. Springer.