Applying Active Learning in Named Entity Recognition Corpora Expansion in Legal Domain

Rafael P. Gouveia[a*], André C.P.L.F. de Carvalho[b], Ellen Souza[c], Hidelberg O. Albuquerque[d], Douglas Vitório[e], Nádia F. F. da Silva[f]

[a] Institute of Mathematics and Computer Sciences, University of São Paulo, São Paulo, Brazil, rafael.p.gouveia2@usp.br, https://orcid.org/0000-0002-4684-8037

[b]Institute of Mathematics and Computer Sciences, University of São Paulo, São Paulo, Brazil, andre@icmc.usp.br, https://orcid.org/0000-0002-4765-6459

[c]MiningBR Research Group, Federal Rural University of Pernambuco, Recife, Brazil, ellen.ramos@ufrpe.br, https://orcid.org/0000-0002-7706-4809

[d]MiningBR Research Group, Federal Rural University of Pernambuco, Recife, Brazil, hidelberg.albuquerque@ufrpe.br, https://orcid.org/0000-0003-2277-8860

[e]Centro de Informática, Federal University of Pernambuco, Recife, Brazil, damsv@cin.ufpe.br, https://orcid.org/0000-0003-2285-574X

[f]Institute of Informatics, Federal University of Goiás, Goiás, Brazil, nadia.felix@ufg.br, https://orcid.org/0000-0002-3875-2211

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. This work investigates the application of Active Learning methodologies for data annotation in Named Entity Recognition (NER) tasks mainly when used in documents from the legal domain in Portuguese. Its aim is to determine an algorithm able to improve the efficiency of the annotation process and reduce the human cost involved, without compromising the quality of the classifiers trained in these corpora. Three sample selection methods were explored: (i) Multi-Criteria Active Learning, using informativeness, representativeness, and diversity as selection criteria, (ii) Dynamic Selection Guided by Entity Volume, and (iii) Random Sentence Selection (used as a baseline for evaluating the other two). The study was conducted using the BERT model for classification, employing different amounts of labeled data for each approach (annotation budgets). The results show that, although Multi-Criteria Active Learning performed better in some scenarios, Dynamic Selection Guided by Entity Volume consistently showed good performance, especially for low annotation budgets, in addition to being computationally more efficient. Thus, the analysis of the results suggests that the volume of named entities is a good predictor for selecting informative samples. This study contributes to the Active Learning field by applying these techniques to modern language models and providing efficient solutions for reducing costs in data annotation for Named Entity Recognition.

Keywords. Active Learning, Named Entity Recognition (NER), Selection Criteria, Information Retrieval, Natural Language Processing, Machine Learning.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.945

1. Introduction

The exponential growth of data generated daily poses a significant challenge for the efficient retrieval of information. The ever-growing magnitude of information collections has made subsequent querying increasingly unfeasible for humans. This difficulty is particularly pronounced when the target of the query involves finding information not based on a formal label (such as a book title) but rather on some description of the desired information. One domain where this challenge is exacerbated is the legislative domain. The Brazilian Chamber of Deputies alone has analyzed over 144,000 projects, each generating documents detailing the process. Moreover, during the drafting of each project, it is necessary to consult related documents such as laws and bills, which are numerous and dispersed among a vast number of unstructured documents (Brandt, 2020).

To address this problem, investments have been made in developing tools that assist in the process of information retrieval from document collections using Natural Language Processing (NLP) techniques—a subfield of artificial intelligence aimed at solving problems related to the automatic understanding and generation of data in human natural languages (Albuquerque et al., 2022). One of the challenges in NLP is automating the process of Named Entity Recognition (NER), which involves detecting textual elements representing names of objects, concepts, etc., and potentially classifying them by entity type. The ability to perform this task automatically and reliably is highly desirable, as much of the subject matter in sentences can be identified solely by the named entities mentioned in the text. For this reason, NER is a widely used technique in various activities, including information retrieval. (Jehangir et al., 2023) However, there are no morphological elements that allow the creation of deterministic algorithms to reliably detect these entities (like the A* algorithm for determining optimal routes). This necessitates the application of sophisticated techniques to achieve the goal.

For this reason, the current state-of-the-art involves the use of Machine Learning techniques. Mitchell, 1998 defines Machine Learning by describing the concept of a ”computer learning” as follows: ”A computer program is said to learn from experience E with respect to some task T and performance measure P, if its performance on T, as measured by P, improves with E.” Machine Learning algorithms are thus designed to improve over time without requiring humans to continuously provide explicit instructions for improvement, such as conditionals (which would require humans to have such information and enough time to account for all possible cases, which is often unreasonable). Machine Learning algorithms require experience (training) to improve their performance to an acceptable level, often surpassing deterministic algorithms or even human experts. The most commonly used machine learning paradigm for NER algorithms is supervised learning, where the algorithms are provided with a list of possible ”inputs” along with the expected output for each input. After adapting the resolution algorithm (often achieved by updating the mathematical model used to represent the task at hand), the program should demonstrate improved performance (typically defined as ”making fewer errors” when exposed to corresponding inputs). However, obtaining these input-output pairs (referred to as ”labeled examples” or ”annotated examples”) is often a costly process, requiring various domain experts to ensure that the annotation is completed within an acceptable time frame and error rate. This process can be financially expensive or even unfeasible due to time and labor constraints. Such costs frequently become prohibitive, especially when considering that the state-of-the-art for various tasks, such as image processing and NLP, involves programs based on more sophisticated models. These models, while achieving higher performance, often require enormous labeled datasets. For example, during the training of the BERT (bi-directional encoding representations from transformers) model, which was once state-of-the-art for NER tasks, Devlin et al., 2018 used the CONLL-2003 dataset, which contains about 300,000 tokens.

Among the initiatives to address problems related to large-scale legislative domain data using NLP is the Ulysses project—a set of initiatives aimed at leveraging artificial intelligence to support legislative activities and improve the relationship between the Brazilian Chamber of Deputies and citizens. Various systems have been explored within the project, including systems that utilize NER models in their architecture to enhance information retrieval systems for use by the Legislative Consultancy of the Chamber of Deputies, as explored by Albuquerque et al., 2024. Therefore, achieving optimal performance for these systems requires highly performant NER models, which, as mentioned earlier, necessitate the availability of a large annotated domain-specific dataset. Although there are large relevant datasets, such as the recently introduced Ulysses Tesemõ by Siqueira et al., 2024 (an extensive corpus in the legal and governmental domain with over 3 million documents sourced from various origins), annotating corpora of this magnitude is unfeasible. Even if only a sample of it is annotated, the decision on which samples to annotate is often arbitrary and potentially inefficient for building a corpus that enhances the performance of classifiers trained with these samples (Albuquerque et al., 2022).

Solving this problem is the goal of the field of Active Learning (AL). The methodology employed by active learning involves generating a selection criterion based on a pre-trained model capable of automatically selecting the best inputs—that is, those whose respective outputs, if labeled and provided to a model, would result in the greatest improvement in its performance, often using, in the NER domain, its uncertainty regarding the labels of the entities it classified. Ideally, this would produce the best possible model by providing the smallest number of labeled examples. This would reduce annotation effort (and, as a side effect, also reduce algorithm training times), making the use of Machine Learning algorithms via supervised learning more accessible (de Sá Vitório, 2020). This work aims to determine an effective active learning algorithm to be employed in the expansion of corpora in the legislative domain in Portuguese intended for training modern named entity recognition models, efficiently reducing the amount of data to be manually labeled for the creation of such models that have a vast array of applications in the streamlining of legislative-related government tasks. To achieve this goal, algorithms will be proposed and implemented that consider different sample selection criteria described in prior literature. These algorithms will be applied in the context of corpus expansion for NER in the legislative domain, with the aim of determining:

- The best algorithm for executing active learning aimed at corpus expansion for training named entity recognition models in the legislative domain;
- The effectiveness of using the proposed attributes for sample selection compared to the annotation of random samples (effectively the non-use of active learning);
- The best hyperparameters for the selected algorithm, given the proposed task.

2. Background and Related Work

In this chapter, a literature review is presented, highlighting previous approaches to reducing the annotation effort required for generating a dataset capable of creating an efficient NER classifier using active learning-based techniques. These serve as the foundation for the methodologies proposed and evaluated.

2.1. Less is More: Active Learning with Support Vector Machines

One of the pioneers in using active learning for text classification had it's main innovation being the use of a simpler metric for selecting annotation candidates: while previous studies employed complex statistical methods to determine the impact of including an example in the training corpus (often requiring retraining the model multiple times considering all annotation possibilities for each candidate), the author proposed that the informativeness of an example should be considered as simply inversely proportional to its margin relative to the hyperplane separating the classes in the current model, given that the model used was an SVM (Support Vector Machine) (Schohn & Cohn, 2000).

The proposed algorithm was tested on document classification tasks using the "20 Newsgroup" corpus, which contains various posts from different online discussion groups. The goal was to perform binary classification on data subsets drawn from any two of the twenty discussion groups. Additionally, the algorithm was tested on topic classification of documents using a dataset of news articles from the Reuters-21578 Distribution 1.0 corpus.

The results were promising, as the learning curves were consistently better for active learning than for the case in which training examples were randomly selected. In some cases, the model's performance with a small number of selected examples even surpassed the performance achieved when using the entire dataset. This suggests that, for certain models, a smaller set of meaningful examples may offer better generalization capability than simply using large random samples.

2.2. Multi-Criteria-based Active Learning for Named Entity Recognition

In the first work in the area of active learning for named entity recognition (NER), the authors aimed to propose an active learning approach based on three criteria and to apply it effectively to facilitate the training of NER models, with the goal of minimizing human annotation effort by selecting the most valuable examples for labeling (Shen et al., 2004).

To maximize the contribution of selected examples, the authors proposed, in addition to informativeness, to also consider representativeness and diversity. To incorporate these three selection criteria at the entity level, two algorithms were proposed. The first aims to ensure representativeness and diversity of the samples with respect to the entire corpus through clustering techniques, based on a similarity metric between tokens. The second proposes a scoring system that weighs informativeness and representativeness using global metrics and then iteratively checks whether the diversity of the selected sample matches the established standard.

The algorithms were tested and compared against random selection and selection based solely on the most informative entities. These tests involved training several SVM models for NER using two corpora for comparison: GENIA (biomedical, containing protein-related names) and MUC-6 (news domain, containing names of people, places, and organizations). After initial training using only part of each corpus to create initial models to determine informativeness, the algorithms and number of selected entities were varied, and additional training and performance testing were conducted using the new examples.

The results were promising, showing that it is possible, using active learning, to select samples comprising only 20% of the corpus size and still achieve performance comparable to a classifier trained on the full dataset (F1-Score of 63.3, which is low by current standards). Moreover, it was concluded that the algorithm which checks for local diversity consistently outperformed the others, and that the active learning process more effectively reduced the required sample size for MUC-6. This suggests that the technique's performance may depend on the complexity of the target task.

However, this work did not explore the effect of varying the hyperparameters of the proposed algorithms. It is also noteworthy that the methods select at the entity level; therefore, in a real-world scenario, entities would be sent for annotation in isolation. Nonetheless, it is important to highlight that the lack of context can hinder the performance of human annotators.

2.3. Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection

In 2008, a paper presented an active learning-based technique that shares the same use case—reducing the number of manually labeled examples required through intelligent sample selection. To this end, it also utilizes feedback from a previously trained classifier for the target task (Tsuruoka et al., 2008).

However, the suggested method diverges from classical active learning by not using uncertainty-based selection criteria. Instead, it selects examples based on the number of named entities that the model predicts exist in the example. This technique is based on the premise that named entities are sparse in the domain where it is applied. Therefore, for better performance, examples containing the largest number of entities should be prioritized.

Thus, the method consists of iteratively selecting sentences with the highest number of predicted named entities and sending all these selected sentences for token-level annotation by a human annotator. The process is repeated until the estimated fraction of annotated named entities reaches a desired value.

Experiments were conducted using a CRF (Conditional Random Fields)-based probabilistic graphical model classifier and the NLPBA (biomedical, a variant of GENIA containing protein names) and CoNLL-2003 (general domain) corpora. It was found that a reduction of approximately 52.4% in annotated sentences was possible to annotate 99% of the named entities. However, the impact of the remaining 1% of unannotated entities, or the effect of these samples on classifier performance, was not explored.

2.4. Active Learning with Pre-trained Language Models for Named Entity Recognition in Requirements Engineering

In 2024, it was published an article about an experiment evaluating active learning techniques for NER using a BERT classifier (a modern model based on transformers, the technology behind state-of-the-art models for various NLP tasks). These techniques were based on informativeness criteria, all derived from the probability distribution provided by the model when predicting the label of a candidate token for manual annotation (Riesener et al., 2024).

The criteria used for sample selection were "least confidence" (token with the lowest probability for the class it was assigned), "margin sampling" (token with the largest difference between the two most probable classes), and "entropy-based sampling." Since the selection for annotation occurred at the sentence level, while the scores were at the token level, a "weakest link" policy was used, where a sentence's informativeness was defined by the informativeness of its most informative token.

Experiments were conducted on NER tasks using datasets with aerospace domain documents from various sources, where the performance of the different techniques across various sentence budgets was compared to that of random sampling.

The results indicated that "margin sampling" was the most effective technique and confirmed that results similar to those obtained with the full corpus could be achieved using only a fraction of the data (36% in the experiments), even for tasks involving deep learning models.

3. Research Methods

This study aims to evaluate the use of specific criteria (informativeness, representativeness, diversity, and entity volume) for sentence selection to support manual annotation through two active learning algorithms, with the goal of improving named entity recognition (NER) models. The basic workflow of the study is illustrated in Figure 1.

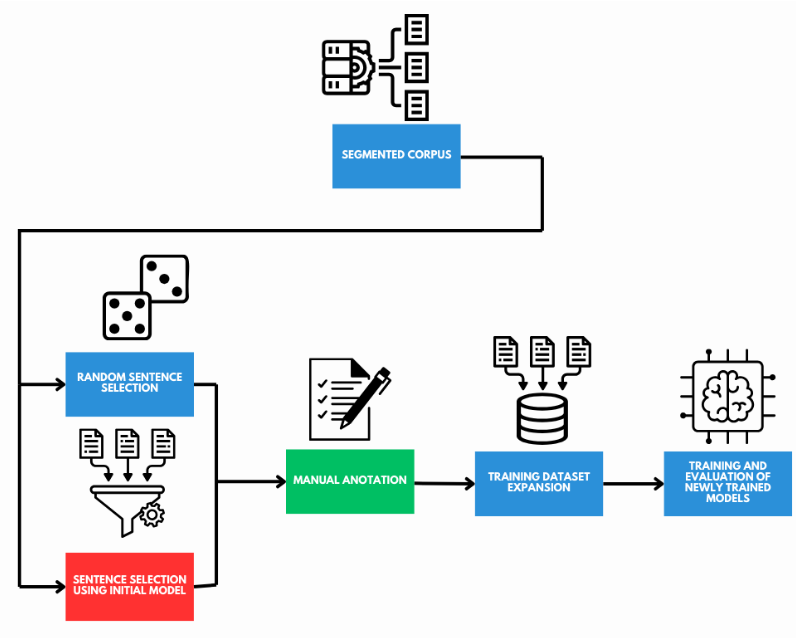
SEGMENTED CORPUS
RANDOM SENTENCE SELECTION
SENTENCE SELECTION USING INITIAL MODEL
MANUAL ANOTATION
TRAINING DATASET EXPANSION
TRAINING AND EVALUATION OF NEWLY TRAINED MODELS

Fig. 1 – Research Flow Schematic

However, the selected dataset, UlyssesNER-Br, is pre-labeled. Thus, the study evaluates model improvement as it is progressively exposed to different samples of this corpus during training, using varying methodologies and sample sizes to define the training sets. The initial model, used to guide the selection criteria, is pre-trained on a fixed sample, achieving low performance compared to the results obtained when trained on the entire corpus.

To achieve this, two algorithms (Multi-Criteria Active Learning and Dynamic Selection Guided by Entity Volume) will be implemented, as detailed in the following subsections. Using the initial model, all proposed algorithms will be executed multiple times, varying hyperparameters and sample sizes to generate diverse training datasets. After retraining the model by concatenating each selected sample with the initial one, the performances will be compared among the different approaches, against a baseline trained on randomly selected samples, and the potential performance gains will be assessed.

3.1. Corpus

The corpus used in these experiments, both for training the initial model used in the selection criteria and for providing sentences to be selected via active learning, is UlyssesNER-Br.

UlyssesNER-Br is a dataset developed for named entity recognition (NER) within the legal domain in Portuguese, as part of the Ulysses project, and is publicly available in an online repository. Entities in this dataset are annotated according to 18 entity types, organized into 7 categories, all of which are relevant for information retrieval efforts in legislative texts, as outlined in Table 1 (Albuquerque et al., 2022).

Tab. 1 – UlyssesNER-Br named entity types

Category	Type	Description	Example from corpus
DATA	DATA	Date	01 de janeiro de 2020
EVENTO	EVENTO	Event	Eleições de 2018
FUNDAMENTO	FUNDlei	Law	Lei no 8.666, de 21 de junho de 1993
	FUNDapelido	Law nickname	Estatuto da Pessoa com Deficiência
	FUNDprojetodelei	Bill	PEC 187/2016
	FUNDsolicitacaotrabalho	Legislative consultation	Solicitação de Trabalho nº 3543/2019
LOCAL	LOCALconcreto	Concrete place name	Niterói-RJ
	LOCALvirtual	Virtual place name	Jornal de Notícias
ORGANIZAÇÃO	ORGpartido	Political party	PSB
	ORGgovernamental	Governmental organization	Câmara dos Deputados
	ORGnãogovernamental	Nongovernmental organization	Conselho Reg. de Medicina (CRM)
PESSOA	PESSOAindividual	Person name	Jorge Sampaio
	PESSOAgrupoind	Group of people	Família Setúbal
	PESSOAcargo	Occupation	Deputado
	PESSOAgrupocargo	Group of people named by their occupation	Parlamentares
PRODUTO DE LEI	PRODUTOsistema	System	Sistema Único de Saúde (SUS)
	PRODUTOprograma	Program	Programa Minha Casa, Minha Vida
	PRODUTOoutros	Other law products	Fundo partidário

Initially, the dataset consisted of legislative bills retrieved from the Brazilian Chamber of Deputies' website and internal legislative consultation requests (these being the objects referenced by the entities "FUNDlei" and "FUNDprojetodelei," respectively). The dataset was subdivided into two corpora: the PL-corpus and the ST-corpus. However, the ST-corpus was deemed confidential and was not made publicly available; thus, it was excluded from this study.

Later, the corpus was expanded the dataset to include the "C-corpus", which comprises informal comments made by citizens on legislative bills, extracted from a web platform provided by the Chamber of Deputies. This expansion aimed to test the hypothesis that exposing the model to formal sentences from the other two corpora could enhance the predictive capacity of a model trained on a same-domain but informal corpus. This hypothesis was confirmed by the experiments published in the same article (Costa et al., 2022).

Finally, the issue of data leakage present in the PL-corpus was addressed and a higher-quality version of the corpus was provided (Nunes et al., 2024). This updated version was used in this study, along with an updated version of the C-corpus, containing more annotated documents (da Costa, 2023).

3.2. Classifier

The named entity recognition model used in this work was the HuggingFace implementation of the BERT architecture, specifically for token classification, called "AutoModelForTokenClassification" (Hugging Face, 2019). More specifically, the model employed was BERTikal, a checkpoint (pre-trained BERT model) created by training the "BERTimbau base cased model" on a corpus of Brazilian legal documents (Polo et al., 2021). BERTimbau, in turn, is a checkpoint developed by training a BERT model on the BrWaC corpus, short for "Brazilian Web as Corpus," a large Portuguese-language dataset obtained via web scraping (Souza et al., 2020).

This model was repeatedly fine-tuned (training a classification layer using the BERT model embeddings generated for each token sequence as input) with the UlyssesNER-Br corpus so that it could recognize the named entities of interest. It was then employed in the active learning processes evaluated in this work, including both the selection of sentences for annotation and the evaluation of the quality of classifiers trained using such samples.

3.3. Multi-criteria Active Learning

The first three criteria to be explored are informativeness, representativeness, and diversity (the latter two obtained using similarity measures between tokens). These criteria are incorporated into a single algorithm, with the weight assigned to each controlled by hyperparameters.

This algorithm was adapted from previous ones to select not only entities for annotation but also textual elements containing them, with a granularity arbitrarily defined by the user of the active learning method (in this study, sentences are used) (Shen et al., 2004). Unlike the original work, where the metrics of each token in the entity are aggregated, this study considers the highest score among all tokens in a sentence as representative of the sentence's overall value. This approach aims to facilitate scalability for various granularities and avoid introducing arbitrariness into the metrics, similar to more recent methodologies (Riesener et al., 2024).

Informativeness is a criterion commonly used in classical active learning methods (Schohn & Cohn, 2000). An example is considered highly informative for training if the respective model exhibits high uncertainty in its classification. This information can be extracted directly from models that provide the probability distribution of possible classes for a given token. As such, informativeness is calculated as follows in the Equation 1:

$$\text{Informativeness}(u) = \text{Uncertainty}(u) = 1 - \mu\left(u, \left(\arg\max_{E} \mu(u, E)\right)\right) \tag{1}$$

Where μ is the token and $\mu(u, E)$ is the probability that the token μ belongs to class E. Since the token is classified into the class with the highest probability, the formula calculates the probability of the token belonging to any class other than the chosen one.

A min-max normalization is also applied to the scores to produce values more representative of the token's informativeness relative to others, as shown in the Equation 2:

$$\text{Informativeness}_{\text{normalized}}(u) = \frac{\text{Informativeness}(u) - \text{min}}{\text{max} - \text{min}} \tag{2}$$

Where min and max are the smallest and largest informativeness values obtained in the corpus, respectively.

For the model used (BERT), the predicted probability distribution for each token's classes can be easily extracted. Hugging Face's implementation (Hugging Face, 2019), used in this study, provides logits (values from the final classification layer, one for each possible class, with the largest value indicating the predicted class). By applying the Softmax function, a corresponding probability distribution can be obtained.

To avoid annotating outliers—entities that, despite high informativeness, are very rare and thus provide little value for frequent model usage—representativeness is considered. An example is representative if it is, on average, highly similar to other tokens.

The similarity is calculated using the token representation provided by BERT (the final layer before the classification head), which incorporates semantic information. For a pair of tokens, similarity is calculated using cosine similarity as in Equation 3:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \tag{3}$$

Where A and B are the two vectors (in this case, token embeddings) that will have their similarity evaluated, the numerator represents the dot product between them and the denominator represents the product of their respective euclidean lengths. Cosine similarity produces a value between -1 and 1, indicating the similarity between two vectors (Manning et al., 2008). The representativeness score is then defined as the average similarity of a token to all others in the corpus as in Equation 4:

$$\text{Representativeness}(u) = \frac{\sum_{v \in \text{dataset}} \text{sim}(u, v)}{|\text{dataset}|} \tag{4}$$

A min-max normalization is also applied to representativeness scores, similarly to informativeness.

To avoid annotating redundant entities—those that, despite being highly informative, are very similar to others already selected and thus do not add new information to the model—diversity is incorporated. An example is considered highly diverse when it has low similarity to other selected examples.

Diversity is ensured by discarding selected tokens that have a similarity above a certain threshold with any other selected tokens.

The algorithm itself involves calculating informativeness and representativeness scores for all tokens in the corpus, combining them using a weighting parameter λ (the weight assigned to representativeness) as shown in Equation 5:

$$\text{score}(u) = (1 - \lambda)\, \text{Informativeness}_{\text{normalized}}(u) + \lambda\, \text{Representativeness}_{\text{normalized}}(u) \tag{5}$$

Sentences (or other textual elements containing the token, depending on the chosen granularity) are then iteratively selected for annotation in descending order of their scores. Sentences are not repeated, and tokens are discarded if their similarity to any already selected token exceeds a threshold γ. The algorithm is illustrated in Figure 2.

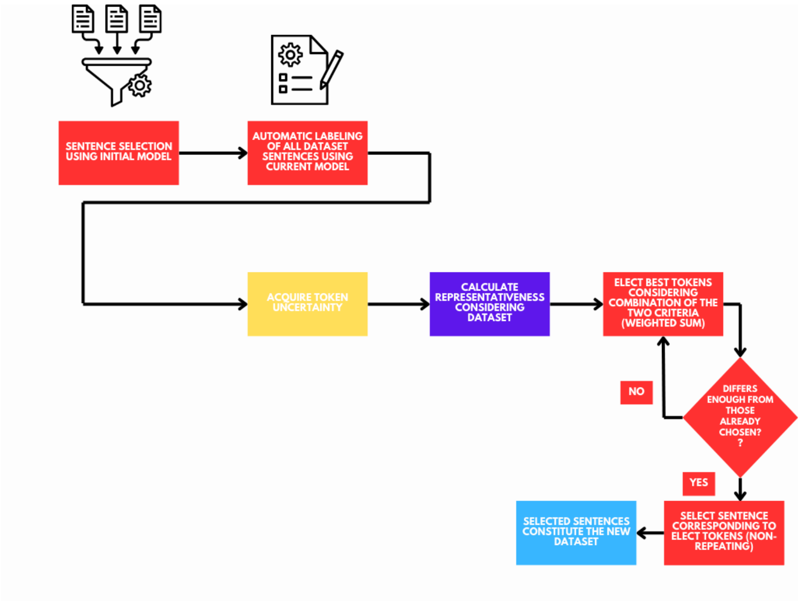
SENTENCE SELECTION USING INITIAL MODEL
AUTOMATIC LABELING OF ALL DATASET SENTENCES USING CURRENT MODEL
ACQUIRE TOKEN UNCERTAINTY
CALCULATE REPRESENTATIVENESS CONSIDERING DATASET
ELECT BEST TOKENS CONSIDERING COMBINATION OF THE TWO CRITERIA (WEIGHTED SUM)
DIFFERS ENOUGH FROM THOSE ALREADY CHOSEN? ?
NO
YES
SELECT SENTENCE CORRESPONDING TO ELECT TOKENS (NON-REPEATING)
SELECTED SENTENCES CONSTITUTE THE NEW DATASET

Fig. 2 – Multi-criteria Active Learning Pipeline Schematic

It's noteworthy that, by the definition given in Equation 4, the calculation of the informativeness value of a single token includes n cosine similarity calculations, in which n is the number of tokens in the dataset. This makes the informativeness calculation complexity proportional to the size of the dataset as a whole, which is unique among all the criteria used in this study. Consequently, in order to calculate the representativeness of all n tokens in the dataset, in order to be able to start the process of selecting sentences, kn^2 cosine similarity operations are needed (where k is a constant to account for basic optimizations), leading to the representativeness calculation part of the algorithm being considered of quadratic time complexity with regards to the total number of tokens in the dataset of sentences to be forwarded to annotation. This could lead to poor scalability, because, as the pool of sentences one can choose for labeling grows (the use case of interest for active learning), the time taken to calculate representativity may become prohibitively large.

3.4. Dynamic Selection Guided by Entity Volume

The fourth criterion to be explored is solely the predicted quantity of named entities present in the analyzed sentence. This criterion is addressed using an algorithm adapted from Tsuruoka et al., 2008. This information can also be obtained directly from models that provide the probability distribution of possible classes for each processed token (such distribution is derived similarly to that used for informativeness calculation, as described earlier).

To calculate this, the expected value for each entity class is considered as fractional occurrences of the entity. The total number of entity occurrences across all tokens in the same sentence is then summed as in Equation 6:

$$\text{Entity Volume}(S) = \sum_{t \in S} (1 - \mu(t, O)) \tag{6}$$

Where S is the sentence, t is the token, and $\mu(t, E)$ is the probability of token t belonging to class E. The null class O corresponds to tokens that are not named entities.

Since all sentences are considered to have the same cost, the selection algorithm involves simply counting the number of entities using the method described above and selecting those sentences with the highest entity count.

The algorithm is illustrated in Figure 3.

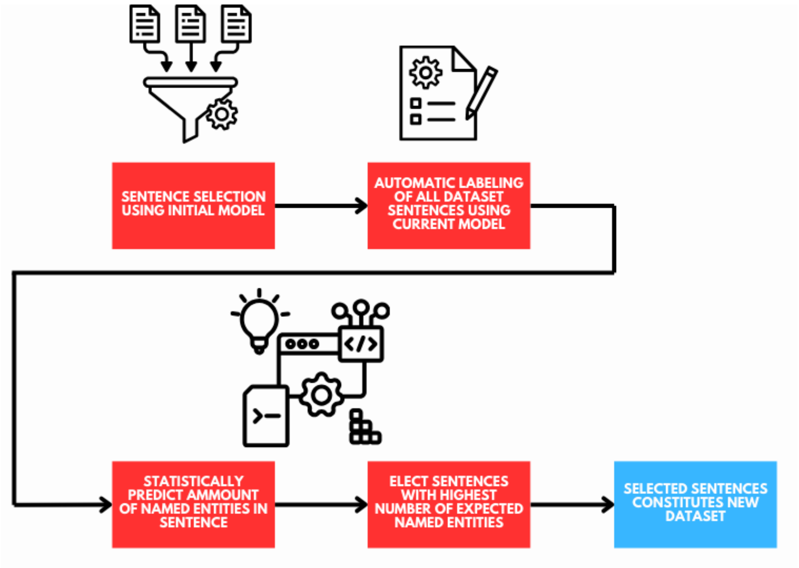
SENTENCE SELECTION USING INITIAL MODEL
AUTOMATIC LABELING OF ALL DATASET SENTENCES USING CURRENT MODEL
STATISTICALLY PREDICT AMMOUNT OF NAMED ENTITIES IN SENTENCE
ELECT SENTENCES WITH HIGHEST NUMBER OF EXPECTED NAMED ENTITIES
SELECTED SENTENCES CONSTITUTES NEW DATASET

Fig. 3 – Dynamic Selection Guided by Entity Volume Pipeline Schematic

3.5. Experiment Structure

Initially, the classifier is trained to classify all 18 different types from UlyssesNER-Br except ”FUNDsolicitacaotrabalho” (since no instance of it occurs in the specific corpora used). This initial process will be done using the C-corpus with an 80/20 train-validation split. Afterward, using the generated model, samples are extracted from the PL-corpus for the three proposed methods: the two described in this chapter and random element selection. Specifically, for the Multi-Criteria Active Learning method, different variations are considered, adjusting the values of its hyperparameters λ and γ. The parameter λ (corresponding to the weight of the representativeness criterion) will be tested for values of 0 (no representativeness), 0.25, 0.5, and 0.75. λ (representing the diversity criterion) will be tested, due to computational and time constraints, only for the cases of 0 (no diversity) and one non-zero value determined through inspection during the experiment, based on the similarity values among tokens present in the dataset.

For each method or variation, four samples will be taken, consisting of 500, 1000, 1500, and 2000 elements for annotation. These elements correspond to the natural division of the dataset, typically segmented by sentences, which in this study are assumed to have equivalent annotation difficulty given the contextual requirements for entity annotation.

Afterward, several classifier training sessions will be conducted using all possible combinations between a generated sample and the C-corpus. Training follows a k-fold cross-validation approach, where the model is trained k times (in this case, 5 times), alternating the portion of the data used for testing, with the remaining portions used for training. This approach yields statistics on the classifier's average performance for the sample generated by each active learning algorithm, mitigating the effects of randomness during training (Scikit-learn, 2024).

The metrics used to compare the performance of the models trained during the experiments are the F1-score and the coefficient of variation. The F1-score is a widely used metric for evaluating classification models, especially in scenarios where class imbalance is present. It is defined as the harmonic mean of precision (the proportion of correctly predicted positive instances out of all predicted positives) and recall (also known as sensitivity, defined as the proportion of true positive instances identified out of all actual positives) (Scikit-learn, 2024).

The coefficient of variation, on the other hand, is the percentage that the variance represents relative to the mean value of the data. This metric is important because it allows us to favor the methods that consistently deliver good results, addressing the fact that focusing solely on average performance can obscure variability in outcomes.

Finally, to ensure greater robustness in the analysis of results, hypothesis tests are conducted using Welch's t-test. This test compares the location parameters (typically the means of normal distributions) of one or two samples. Unlike the more common Student's t-test, Welch's t-test does not assume that the samples have equal variances or sizes, making it more flexible and suitable for real-world data scenarios (Navarro, 2022).

4. Results and Discussion

4.1. Experimental Results

The experiments described earlier tested the three proposed algorithms (Multi-Criteria Active Learning with different weights for representativeness and with or without the diversity criterion, Entity Volume-Oriented Dynamic Selection, and random sentence selection). Samples were generated for training, and following the k-fold cross-validation methodology previously outlined, five fine-tuning processes of the BERT model were conducted, starting from the pre-trained BERTikal model introduced in the previous section. Figure 4, 5 and 6 shows the average F1-score metrics of each of the five models generated, corresponding to each training sample.

It should be noted that each trace represents a configuration of the active learning algorithm. Thus, they illustrate the model's performance evolution as larger training samples are allowed.

Visual analysis of the results reveals that, as expected, all algorithms provided an apparent improvement, albeit marginal, over random selection, except for the multi-criteria algorithm configuration that disregarded diversity and assigned the highest weight to representativeness. This result can be explained by the fact that this configuration favors the creation of a homogeneous sample, which therefore offers little new information to the model (justifying its inferior performance compared to random sampling, especially for larger data sample sizes).

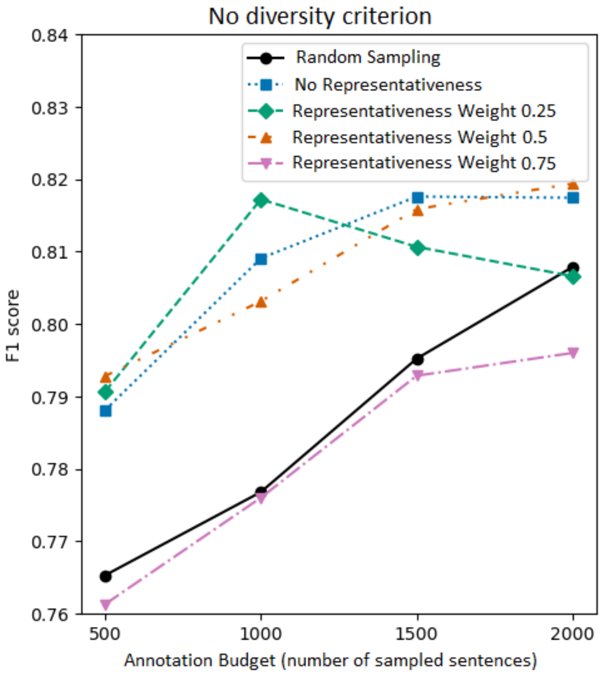

Fig. 4 – Graphs containing the performance of the multi-criteria method without diversity

Regarding the multi-criteria algorithm, although applying the diversity criterion seemingly corrected the issues mentioned above for cases with excessive weight given to representativeness—enabling the model to marginally outperform random sampling—it generally resulted in worse performance on average when diversity was considered. This difference was more pronounced for low-budget scenarios, diminishing as the budget increased. However, this reduction may be attributed to the exhaustion of options, leading the algorithms to select from the same examples (as the PL-Corpus contains only 2,527 data entries).

On the other hand, analyzing the entity volume-oriented algorithm reveals that it performed superiorly, achieving the best model with the smallest number of entries, with an average F1-score of approximately 0.795, about 0.03 higher than the average for random samples with the same budget. The other algorithms that performed better with low budgets only achieved an advantage of approximately 0.015 over random selection. Furthermore, despite the reduction in the advantage relative to random selection with increasing budgets, the difference consistently remained no less than 0.01.

Finally, the entity volume-oriented algorithm is considered superior not only due to its better performance with smaller budgets but also because of its greater simplicity and faster execution. It does not require computationally expensive steps, such as the representativeness calculations necessary for multi-criteria active learning, nor does it depend on optimizing hyperparameters specific to the corpus properties. Instead, it relies only on the assumption of named entities' sparsity being valid.

These results suggest that the predicted volume of named entities is a better predictor of good examples to include in the sample, although informativeness (alone) and representativeness (when properly weighted) can also serve as good predictors, despite the latter being computationally costly. Diversity, however, did not prove to be a good predictor. On the contrary, within the limited scope explored in this study, it was detrimental to the algorithm's performance.

However, it is important to note that this analysis is based on average performance across a series of experiments. Therefore, to confirm these results, hypothesis testing was conducted to exclude the influence of variance on the main outcomes of the analysis.

4.2. Hipotesis Testing

Four hypotheses were tested using Welch's t-test, as described in the Subsection 2.5 The hypotheses were formulated sequentially based on the results of previously tested hypotheses.

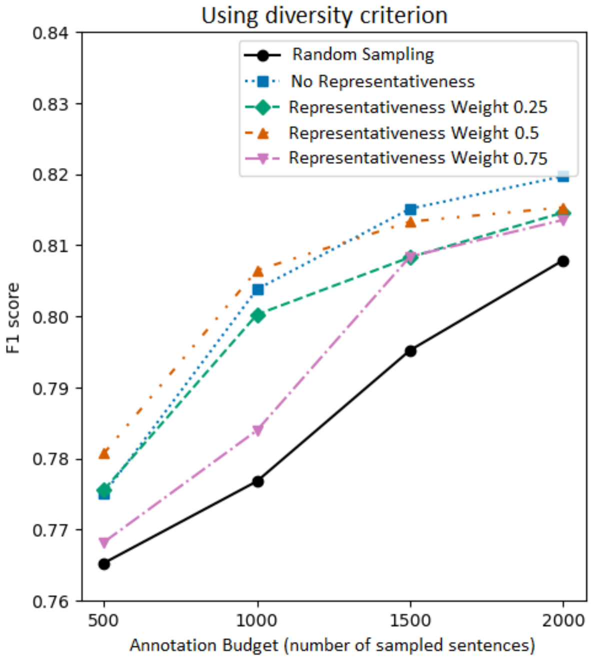

Fig. 5 – Graphs containing the performance of the multi-criteria method including diversity

The first hypothesis tested was whether any of the proposed methods resulted in an improvement compared to random sampling (used as the control in this experiment) in terms of F1-score for low budgets. Low budgets were prioritized in the evaluation since the primary goal of active learning is efficiency with small samples. Thus, the null hypothesis, which we wish to reject, consists of the proposed methods being worse or equivalent to random sampling, as explained below in Equation 7:

$$\begin{cases} H_0 : \mu_{\text{method}} <= \mu_{\text{random}} \\ H_a : \mu_{\text{method}} > \mu_{\text{random}} \end{cases} \tag{7}$$

After calculating the relevant values, Table 2 was generated:

Tab. 2 – Hypothesis Test 1 (Equation 7) - Whether, for a low budget, any method outperformed random sampling

Method	Average F1-Score	Standard Deviation	Var. Coef. %	t-statistic	P-value
Amostragem aleatória	0.765312	0.014761	1.928702	-	-
info.	0.788102	0.020510	2.602401	2.016675	**0.04102**
info. + repres. (0.25)	0.790656	0.017582	2.223775	2.468595	**0.01982**
info. + repres. (0.5)	0.792830	0.015609	1.968725	2.864208	**0.01054**
info. + repres. (0.75)	0.761319	0.010901	1.431814	-0.486585	0.67964
info. + diver.	0.775049	0.013038	1.682240	1.105487	0.15077
info. + repres. (0.25) + diver.	0.775598	0.027827	3.587758	0.730130	0.24622
info. + repres. (0.5) + diver.	0.780737	0.021791	2.791018	1.310466	0.11560
info. + repres. (0.75) + diver.	0.768246	0.010041	1.307029	0.367483	0.36202
volume	0.794829	0.007793	**0.980508**	3.954136	**0.00366**

For a significance level of $\alpha = 0.05$, only Multi-Criteria Active Learning with high representativeness weight or including the diversity factor failed to reject the null hypothesis. For all other methods, superiority over the random method was statistically confirmed.

Thus, as an initial conclusion, the Entity Volume-Oriented Dynamic Selection was chosen as the best method. This choice was not only due to its significant p-value but also because it demonstrated a considerably lower coefficient of variation than the other methods. Furthermore, its low computational complexity solidified its position as the preferred method, confirmed through subsequent tests.

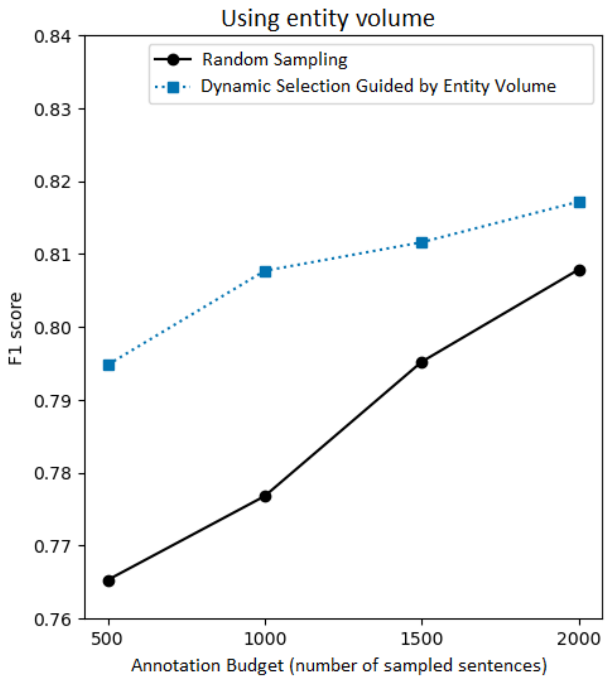

Fig. 6 – Graphs containing the performance of the entity volume method

Next, to confirm the choice of Dynamic Selection Guided by Entity Volume as the best method, the hypothesis that all methods superior to random sampling perform equivalently to this method was tested. If this hypothesis were rejected, it would then be necessary to test whether the other methods are superior to it. Thus, the null hypothesis is equality with respect to the performance of the volume-based method, as shown below in Equation 8:

$$\begin{cases} H_0 : \mu_{\text{volume}} = \mu_{\text{method}} \\ H_a : \mu_{\text{volume}} \neq \mu_{\text{method}} \end{cases} \tag{8}$$

After calculating the relevant values, Table 3 was generated.

Tab. 3 – Hypothesis Test 2 (Equation 8) - Whether, for a low budget, any method was distinct from entity volume sampling

Method	Average F1-Score	Standard Deviation	Var. Coef. %	t-statistic	P-value
Amostragem aleatória	0.765312	0.014761	1.928702	3.954136	**0.00733**
info.	0.788102	0.020510	2.602401	0.685582	0.52271
info. + repres. (0.25)	0.790656	0.017582	2.223775	0.485122	0.64625
info. + repres. (0.5)	0.792830	0.015609	1.968725	0.256235	0.80650
info. + repres. (0.75)	0.761319	0.010901	1.431814	5.591745	**0.00073**
info. + diver.	0.775049	0.013038	1.682240	2.911764	**0.02441**
info. + repres. (0.25) + diver.	0.775598	0.027827	3.587758	1.488121	0.20150
info. + repres. (0.5) + diver.	0.780737	0.021791	2.791018	1.361609	0.23138
info. + repres. (0.75) + diver.	0.768246	0.010041	1.307029	4.67642	**0.00186**
volume	0.794829	0.007793	0.980508	-	-

The analysis showed that distinction in performance could only be confirmed for some methods that were not superior to the control (random sampling). Thus, it was considered reasonable to assume that they are inferior to volume selection.

Consequently, it can be concluded that Dynamic Selection Guided by Entity Volume is not necessarily superior to Multi-Criteria Active Learning, as they demonstrated equivalent performance. However, since both are valid active learning methods for improving models with low budgets (as shown in Test 1), and the volume-based method has lower computational complexity and a lower variation coefficient, it is considered the better method.

The first graph in Figure 4 exhibited an outlier, suggesting the superiority of Multi-Criteria Active Learning for the case without diversity and with a representativeness weight of 0.25. Therefore, the aim is to reject the hypothesis that this method is inferior or equivalent to others at that budget, as shown below in Equation 9:

$$\begin{cases} H_0 : \mu_{\text{info.+repres. }(0.25)} <= \mu_{\text{method}} \\ H_a : \mu_{\text{info.+repres. }(0.25)} > \mu_{\text{method}} \end{cases} \tag{9}$$

After calculating the relevant values, Table 4 was generated.

Tab. 4 – Hypothesis Test 3 (Equation 9) - Whether, for a budget of 1000 sentences, any method outperformed the multicriteria sampling info. + repres. (0.25)

Method	Average F1-Score	Standard Deviation	Var. Coef. %	t-statistic	P-value
Amostragem aleatória	0.776813	0.011072	1.425348	4.428778	**0.00161**
info.	0.809079	0.009620	1.188965	0.926857	0.19408
info. + repres. (0.25)	0.817225	0.017138	2.097133	-	-
info. + repres. (0.5)	0.803147	0.021569	2.685575	1.142693	0.14391
info. + repres. (0.75)	0.776030	0.021806	2.809931	3.321323	**0.00568**
info. + diver.	0.803814	0.011866	1.476264	1.438642	0.09636
info. + repres. (0.25) + diver.	0.800263	0.018085	2.259836	1.522326	0.08326
info. + repres. (0.5) + diver.	0.806487	0.006227	0.772124	1.316829	0.1223
info. + repres. (0.75) + diver.	0.783971	0.013619	1.737138	3.396902	**0.00506**
volume	0.807713	0.011623	1.439041	1.027128	0.16918

At the chosen significance level ($\alpha = 0.05$), it was not possible to reject the hypothesis that the method was inferior or equivalent to the other candidates for the best method.

Thus, it can be concluded that the observed outlier was merely a result of randomness in the experiment's execution and does not represent any significant phenomenon.

Finally, it was necessary to determine whether the result for low budgets (that the best method was Dynamic Selection Guided by Entity Volume) also held true for high budgets. Thus, we sought to reject the null hypothesis that the methods are inferior or equivalent to the volume method, as shown below in Equation 9:

$$\begin{cases} H_0 : \mu_{\text{method}} <= \mu_{\text{volume}} \\ H_a : \mu_{\text{method}} > \mu_{\text{volume}} \end{cases} \tag{10}$$

After calculating the relevant values, Table 5 was generated.

However, for the chosen significance level ($\alpha = 0.05$), the null hypotheses could not be rejected. Thus, for high budgets, other methods remained inferior or equivalent to the method initially deemed the best: Dynamic Selection Guided by Entity Volume.

5. Conclusion

Through the experiments conducted, it was possible to study the performance of algorithms based on relatively older literature in the field of active learning, which were adapted for use alongside modern large language model-based classifiers. Thus, it was concluded that these algorithms are capable of obtaining, as proposed, samples that are significantly more informative for classifiers than random samples, making them a valuable resource for reducing the cost of corpus annotation.

Tab. 5 – Hypothesis Test 4 (Equation 10) - Whether, for high budget, any method outperformed entity volume-based sampling

Method	Average F1-Score	Standard Deviation	Var. Coef. %	t-statistic	P-value
Amostragem aleatória	0.807885	0.016930	2.095648	-0.839945	0.78728
info.	0.817446	0.011734	1.435467	0.026275	0.48989
info. + repres. (0.25)	0.806624	0.020560	2.548955	-0.862897	0.79314
info. + repres. (0.5)	0.819404	0.017945	2.190047	0.194081	0.42547
info. + repres. (0.75)	0.796019	0.013833	1.737745	-2.079001	0.96319
info. + diver.	0.819714	0.016079	1.961481	0.232992	0.41084
info. + repres. (0.25) + diver.	0.814607	0.011369	1.395702	-0.270612	0.6026
info. + repres. (0.5) + diver.	0.815335	0.011909	1.460666	-0.191755	0.57328
info. + repres. (0.75) + diver.	0.813570	0.018985	2.333498	-0.308862	0.61733
volume	0.817192	0.018090	2.213674	-	-

Additionally, this research's main contribution was the determination that, among the two proposed algorithms, the best one for active learning in legal domain corpora in Brazilian Portuguese, given the configurations studied in this work, is the Dynamic Selection Guided by Entity Volume, as it has lower computational and theoretical complexity and significantly better performance for lower annotation budget values.

Finally, from these results, it is concluded that the predicted volume of named entities is a better predictor of good examples for sample selection, although informativeness and representativeness can also be good predictors, despite the latter being computationally costly.

6. Future Work

To further this research, future work could involve testing the proposed methods with models other than BERT to assess the extent to which the results depend on the underlying model. Additionally, it would be valuable to conduct direct comparisons between these methods and other state-of-the-art approaches. Finally, the methodology identified as the most effective could be applied to develop additional annotated corpora of legal domain documents.

Acknowledgement

- **Funding or Grant**: We express our gratitude to research funding agencies, CAPES, FAPESP, and CNPq for supporting this research. We also thank the Center for Applied Research in Artificial Intelligence Recreating Environments (CPA IARA) and the Brazilian National Institute of Artificial Intelligence (IAIA) for supporting this research.
- **Data/Software Access Statement**: The corpus used can be found on the oficial repository: https://github.com/ulysses-camara/ulysses-ner-br
- **Contributor Statement***: Conceptualization: Rafael P. Gouveia, André C.P.L.F. de Carvalho, Ellen Souza; Formal analysis: Rafael P. Gouveia; Funding acquisition: André C.P.L.F. de Carvalho, Ellen Souza; Methodology: Rafael P. Gouveia, Ellen Souza, Hidelberg O. Albuquerque, Douglas Vitório, Nádia F. F. da Silva; Software: Rafael P. Gouveia; Resources: Ellen Souza, Hidelberg O. Albuquerque, Douglas Vitório; Writing - Original Draft: Rafael P. Gouveia; Writing – Editing: Ellen Souza, Nádia F. F. da Silva;
- **Conflict Of Interest (COI)***: There is no conflict of interest.

References

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., & Oliveira., A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. *PROPOR 2022*.

Albuquerque, H. O., Souza, E., Silva, T., Gouveia, R. P., Junior, F., Vitório, D., da Silva, N. F. F., de Carvalho, A. C., Oliveira, A. L., & de Andrade, F. E. (2024, March). Ulyssesnerq: Expanding queries from brazilian portuguese legislative documents through named entity recognition. https://aclanthology.org/2024.propor-1.35/

Brandt, M. B. (2020). *Modelagem da informação legislativa: Arquitetura da informação para o processo legislativo brasileiro* [Doctoral dissertation, UNIVERSIDADE ESTADUAL PAULISTA].

Costa, R., Albuquerque, H. O., Silvestre, G., Silva, N. F. F., Souza, E., Vitório, D., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., de Souza Dias, M., Pereira, F. S. F., Silva1, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., & Oliveira, A. L. I. (2022). Expanding ulyssesner-br named entity recognition corpus with informal user-generated text. *EPIA 2022*.

da Costa, R. P. (2023, Abril). Reconhecimento de entidades nomeadas em textos informais no domínio legislativo. https://repositorio.bc.ufg.br/tede/items/c2783ab8-8aba-487d-bdc2-8fbf0a375a10

de Sá Vitório, D. Á. M. (2020). Avaliando estratégias de seleção de active learning para mineração de opinião com fluxos contínuos de dados. https://repositorio.ufpe.br/bitstream/123456789/38119/1/DISSERTA%c3%87%c3%83O%20Douglas%20%c3%81lisson%20Marques%20de%20S%c3%a1%20Vit%c3%b3rio.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Hugging Face, I. (2019). https://huggingface.co

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition – datasets, tools, and methodologies. *Natural Language Processing Journal*, *3*, 100017. DOI: https://doi.org/10.1016/j.nlp.2023.100017.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mitchell, T. (1998). *Machine learning*. McGraw-Hill Science/Engineering/Math.

Navarro, D. (2022, Dezembro). The independent samples t-test (welch test). https://stats.libretexts.org/Workbench/Learning_Statistics_with_SPSS_-_A_Tutorial_for_Psychology_Students_and_Other_Beginners/10%3A_Comparing_Two_Means/10.04%3A_The_Independent_Samples_t-test_(Welch_Test)

Nunes, R. O., Spritzer, A. S., Freitas, C. M. D. S., & Balreira, D. G. (2024, Setembro). *Reconhecimento de entidades nomeadas e vazamento de dados em textos legislativos: Uma reavaliação da literatura* [Não publicado]. https://www.researchgate.net/publication/384467812_Reconhecimento_de_Entidades_Nomeadas_e_Vazamento_de_Dados_em_Textos_Legislativos_Uma_Reavaliacao_da_Literatura

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., de Lima, L. M. P., do Amaral Maia, A. C., & Vicente, R. (2021, Outubro). Legalnlp – natural language processing methods for the brazilian legal language. https://arxiv.org/pdf/2110.15709

Riesener, M., Kuhn, M., Schümmelfeder, S., & Xiao, D. (2024, Outubro). Active learning with pre-trained language models for named entity recognition in requirements engineering. https://www.sciencedirect.com/science/article/pii/S2212827124007078

Schohn, G., & Cohn, D. A. (2000, June). Less is more: Active learning with support vector machines. https://dl.acm.org/doi/10.5555/645529.657802

Scikit-learn. (2024). Cross-validation: Evaluating estimator performance. *scikit-learn*. https://scikit-learn.org/1.5/modules/cross_validation.html

Shen, D., Zhang, J., Su, J., Zhou, G., & Tan, C.-L. (2004, July). Multi-criteria-based active learning for named entity recognition. https://aclanthology.org/P04-1075/

Siqueira, F. A., Vitório, D., Souza, E. P., & Santos, J. A. P. (2024, July). Ulysses tesemõ: A new large corpus for brazilian legal and governmental domain. https://www.researchgate.net/publication/384467812_Reconhecimento_de_Entidades_Nomeadas_e_Vazamento_de_Dados_em_Textos_Legislativos_Uma_Reavaliacao_da_Literatura

Souza, F., Nogueira, R., & de Alencar Lotufo, R. (2020, Outubro). Bertimbau: Pretrained bert models for brazilian portuguese. https://www.researchgate.net/publication/345395208_BERTimbau_Pretrained_BERT_Models_for_Brazilian_Portuguese

Tsuruoka, Y., Tsujii, J., & Ananiadou, S. (2008, November). Accelerating the annotation of sparse named entities by dynamic sentence selection. https://aclanthology.org/W08-0605.pdf