Evaluation of Named Entity Recognition using Ensemble in *Transformers* Models for Brazilian Public Texts

Eutino Júnior Vieira Sirqueira[1,2], **Flávio de Barros Vidal**[1]

[1] Department of Computer Science, University of Brasília, Brasília, Brazil
[2]Federal Institute of Piauí, Campus Corrente, Corrente, Brazil

Abstract. *Natural Language Processing (NLP) has experienced significant advances, driven mainly by developing deep learning models using Transformers. In the Brazilian context, the analysis of open data, such as official documents published in the Official Federal Gazette (DOU), is crucial for transparency and access to information. In this work, we propose an evaluation of ensemble models, using Transformers models, applied for the Named Entity Recognition (NER) task in Brazilian Public Texts. The proposed evaluation tested a set of models based on the Bidirectional Encoder Representations from Transformers (BERT) model variations and combinations of ensemble strategies, reaching improvements of up to 11% in the proposed corpus when compared with classic NER approaches using only BERT-based models.*

1. Introduction

The area of Natural Language Processing (NLP) has experienced significant advances, driven mainly by developing deep learning models using *Transformers* [Tay et al. 2022]. One of the NLP tasks, among the many existing in the literature and with great practical applicability, is NER, which aims to identify and classify relevant entities in texts, such as names of people, organizations, locations, dates and, others [Li et al. 2022]. The use of NER has great potential to assist in the analysis and management of documents, especially in contexts such as legal and governmental, where accurate extraction of information is crucial [Rodríguez and Bezerra 2020]. In the Brazilian context, according to [Possamai and de Souza 2020], the analysis of open data, such as official documents published in the DOU, is crucial for transparency and access to information. According to [Possamai and de Souza 2020], it is a challenging process with many possibilities. Because of this, identifying named entities can become even more relevant when applying to public Agreements. Public Agreements are transfer contracts and partnership terms on financial resources between the Brazilian state and other government entities (or NGOs), aiming for a common objective. NER tasks enable data extraction that can collaborate with initiatives seeking to improve public management transparency and efficiency, such as the Deep Vacuity platform [de Carvalho et al. 2022].

In this study, the objective is to evaluate how combining predictions from models based on *Transformers* using an ensemble approach in NER tasks can improve the effectiveness of recognizing entities named in documents from the DOU with an emphasis on public agreements. A set of *ensemble* combination approaches is compared to individual prediction *transformers* models and evaluated for performance in terms of precision, recall, and F1 score, metrics demonstrated in [Dalianis and Dalianis 2018]. The results of this study provide valuable insights into the potential of *ensembles* arrangements in *Transformers* models, which can have significant implications for compliance, data discovery, and evidence-based decision-making.

Regarding related works focusing on Brazilian official, legal, and governmental documents, this work distinguishes itself by addressing the recognition of a greater variety of entities named in public Agreement documents published in the DOU, using ensemble techniques applied to *Transformers* models. Not finding works exclusively in this specific context that provided adequate data, a corpus for Public Agreements was annotated with 192,900 publications from the DOU, covering 27 classes of entities[1]. A strategy was also developed to automate the corpus annotation process. With this, seven models based on *Transformers*, according to Table 2, were trained using the Transfer learning and fine-tuning approach and tested. Finally, ensemble voting techniques were applied and evaluated for precision, recall, and F1 score metrics. The results showed significant improvements in evaluation metrics, even with more entities and without the use of fine adjustments in the models. The automated annotation of a new corpus in an exclusive context also contributed to advances in the study of techniques that seek transparency and efficient management of public documents, paving the way for developing more effective tools.

This work is structured as follows: In Section 2, related works on NER in Brazilian legal and government documents are presented, in addition to NER with *Transformers* models. Section 3 shows ensembles with *Transformers* models. Section 4 details the proposed methodology, including collecting and annotating the corpus, training models, implementing ensemble strategies, and comparing results. The results are discussed in Section 5, and conclusions and future directions are presented in Section 6.

2. Related Works

2.1. NER in Brazilian Public Documents

Several studies have explored the application of NER to improve the analysis and management of Brazilian public and governmental data using varied approaches. In the work of [de Araujo et al. 2018], they focused on creating specific corpora to improve entity extraction. In [Alles et al. 2018], tools such as OpenNLP, CoreNLP, NLTK, and Syntaxnet were compared and applied to the DOU, 11 categories of entities were identified, and the challenges in defining relevant entities were highlighted.

In [de Araujo et al. 2018], is presented the Dataset for NER in Brazilian Legal Text (LeNER-Br), a legal corpus with six categories of entities, which used a Long Short-Term Memory (LSTM) network and Conditional Random Fields(CRF). Both studies showed significant improvements in extraction quality using specific corpora, although they faced challenges in generalization.

In [Albanaz 2020] developed a model to identify companies winning bids, achieving 90% accuracy on a dataset of 19,321 publications. According to [Silva et al. 2022], they have applied Deep Learning methods to segment and classify legal documents from the Official Gazette of the Federal District (DODF), obtaining the best average F1 score of 0.885. Both works showed the effectiveness of Deep Learning models in specific NER tasks. However, they still claim challenges in adapting to different contexts and the need for large volumes of annotated data.

Recent studies, such as [Wang et al. 2020] and [Belém et al. 2022], used advanced deep learning and *Transformers* models. [Wang et al. 2020] proposed the Sequence Tagging Model (STM), which combines Iterated Dilated Convolutional Neural Networks (IDCNN) and Bi-LSTM, achieving an F1 score of 0.9323 on LeNER-Br, suggesting future improvements with optimization of the Bidirectional Encoder Representations from *Transformers* (BERT) model. For [Belém et al. 2022] SpERT method for NER and relationship extraction in documents from the Official Gazette of Minas Gerais (DOMG), implementing preprocessing and post-processing strategies that resulted in significant improvements in metrics precision, recall, and F1 score. As described in [da Silva 2022], NER is applied in bidding documents using deep learning in experiments with the Portuguese language corpora Harem, Paramopama, and LeNER-Br. Furthermore,

[1] All dataset is available for download at *information removed in the review of manuscript.*.

they developed a manually annotated corpus from 67 tender notices, considering only eight possible classes of relevant named entities. The results indicate that BERT-based models, particularly the BiLSTM-CRF and W2V-BERTLarge, outperformed state-of-the-art approaches in selective evaluation scenarios.

The work presented by [Wang et al. 2020] demonstrates the potential of advanced deep learning techniques and *Transformers*, although there may be ongoing needs for model refinement. Finally, [Guimarães et al. 2024] introduce the DODFMiner tool, which combines pre-processing, rule-based classification, and NER with machine learning to extract named entities from DODF. Despite being promising in extracting information, this tool faces challenges in adapting to different standards of official journals.

2.2. NER Tasks, *Transformers* and Ensemble Models

Ensemble Named Entity Recognition (E-NER) methods combine multiple models to improve performance. Some relevant works use *Transformers* models for this purpose, as in [Singh et al. 2023, Zheng and Sun 2023], the ensemble technique was applied to improve the NER performance of 6 categories in Hindi documents. Combining multiple models using *soft voting* improved accuracy, precision, and recall metrics, increasing from 0.68 to 0.71. As presented in [Zheng and Sun 2023], they used ensemble techniques with four models based on BERT and RoBERTa for NLP in Old Chinese, addressing word segmentation, grammatical tagging, and NER of 3 types of entities (person, location and time). In [Sun et al. 2021], five models were combined for the recognition of 18 classes of entities in Chinese medical dialogues. Model merging via *voting* reduced single model bias, improving overall performance. According to [Rouhizadeh and Teodoro 2022], they used an ensemble of three transformer models for NER in the multilingual context of task 11 of SemEval-2022. The combination by *hard voting* resulted in the 20th position, with an F1 score of 0.652.

Although these works demonstrate the use of transformer ensembles for NER in different languages and domains, no similar approach exists for Brazilian public agreements in Portuguese. Our work seeks to fill this gap, applying ensemble techniques with transformer models to recognize various entities in documents from the DOU.

3. Ensembling *Transformers* Models

As described in [Sagi and Rokach 2018], Ensemble is a general term for methods that combine several basic models to make a decision, typically in supervised machine learning tasks. In another view, according to [Sagi and Rokach 2018], ensemble methods are machine learning techniques for combining predictions from several individual models to improve the accuracy and stability of predictions. These particular models, also known as base learners, can be of different types, such as decision trees, neural networks, and linear regression. An ensemble of these techniques can generate a final model with superior predictive performance [Khan et al. 2024].

In this way, Ensembles is based, according to [Sagi and Rokach 2018], on the idea that by aggregating the predictions of several models, where the individual errors of each one tend to be compensated, resulting in greater robustness and precision. Several methods can be used to build an Ensemble, each with its features and applicability. Among the most popular techniques according to recent works [Sagi and Rokach 2018, Khan et al. 2024], as follows: **Bagging (Bootstrap Aggregating)** - Bagging involves training several independent models on different subsets of the original dataset (obtained by sampling with replacement) and then combining their predictions; **Boosting** - Boosting works sequentially, where each model is trained to correct the errors made by the previous model in the sequence; **Stacking (Stacked Generalization)**: Stacking combines the predictions of several individual machine learning models using another model (meta-model) to learn how to integrate the predictions of the base models; **Voting** - Combines the predictions of several simple machine learning models (such as classifiers or regressors) and returns the prediction that the individual models most frequently chose. Each method has specific variations (or submethods) that differ in combining the base models. As presented in [Khan et al. 2024], multiple machine learning models are trained independently, and then their predictions are combined through a voting process to determine the final prediction. There are two main types of voting: **Hard Voting** - The majority of votes from individual models determines the final prediction. The class (in the case of classification) or the average value (in the case of regression) that receives the most votes is chosen as the final prediction; **Soft Voting** - Each model assigns a weight to its predictions based on its confidence (e.g., class probabilities). The weighted predictions are then combined to obtain the final prediction.

As demonstrated by [Singh et al. 2023, Zheng and Sun 2023], voting and its variations are simple and effective for improving the accuracy and robustness of NER tasks. Our work implemented variations of the Hard Voting and Soft Voting methods in transformer models trained for NER. Furthermore, a variation is developed where the Model with the highest F1 score metric determines the final classification of named entities.

4. Proposed Methodology

The methodology adopted in this work was organized into four major stages, presented below in Figure 1.

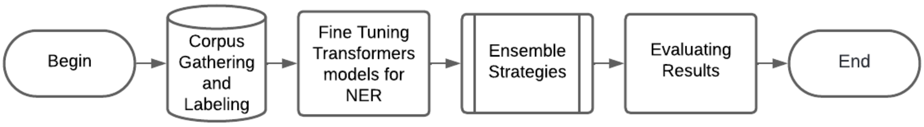

Begin
Corpus Gathering and Labeling
Fine Tuning Transformers models for NER
Ensemble Strategies
Evaluating Results
End

Figure 1. Fluxogram of the Proposed Methodology.

4.1. Corpus Gathering and Labeling

The corpus was created in several methodological steps. Firstly, entity mapping was done to identify the entities in the text data. This mapping was done using data files provided and available on the *Portal da Transparência* of the Brazilian government, which is the official government website that allows you to monitor the use of public resources, containing data of public agreements with 27 types of potential entities. Initially, a sample of 50 public agreements was analyzed to define which entities would be relevant in the search for publications in the DOU based on criteria of uniqueness and frequency. Uniqueness is given by the efficiency in searching for publications, and frequency concerns the recurrence of entities in publications.

Next, the publication's search is focused on collecting and storing DOU information related to the public agreements data. This involved web data extraction techniques to extract data from DOU pages using mapped entities and specific filters. Validation of the Publications was the third step, which was essential to ensure that the publications obtained corresponded to the public Agreements of interest in the *Portal Transparência* data. Using frequent entities, the minimum presence of three entities was verified in each publication to validate its relevance. Finally, the Publication Annotation step prepared the data for training the machine learning models, using the spaCy PhraseMatcher tool [Honnibal et al. 2020] to identify and label entities in the texts of DOU publications according to specific terminology. This terminology was obtained from the Comma-separated values (CSV) file downloaded from the *Portal Transparência* related to the documents. One hundred ninety-two thousand nine hundred publications were annotated from the

DOU of 71,287 Agreements of *Portal Transparência*[2]. Table 1 shows the number of annotations per entity, which is the labeling defined by the Brazilian Government in Portuguese.

Table 1. Description of the used Entities from "Portal da Transparência".

Entity ID	**# of Annotations**	**Description**
Número Convênio	92.891	Identify the agreement
UF	1.361.105	Federative Unit of the grantor
Código SIAFI Município	23.166	Identify the municipality of the grantor
Nome Município	68.081	Name of the municipality of the grantor
Situação Convênio	4.004	Status of the agreement
Número Original	25.771	Original number of the agreement
Número Processo do Convênio	38.371	Number of the agreement process
Objeto do Convênio	114.439	Object agreed by the entities
Código Órgão Superior	1.114	Code of the superior granting agency
Nome Órgão Superior	1.495.540	Name of the superior granting agency
Código Órgão Concedente	17.774	Code of the granting agency
Nome Órgão Concedente	66.591	Name of the granting agency
Código UG Concedente	301.681	Code of the granting UG
Nome UG Concedente	1.383	Name of the granting UG
Código Convenente	102.837	Code of the recipient
Tipo Convenente	191	Type of the recipient
Nome Convenente	149.594	Name of the recipient
Tipo Ente Convenente	2.274.833	Type of the recipient entity
Tipo Instrumento	3.059	Type of instrument
Valor Convênio	171.283	Amount of the grantor's contribution
Valor Liberado	162	Total amount released by the government
Data Publicação	1	Agreement publication date
Data Início Vigência	2.190.278	Start date of the agreement's validity
Data Fim Vigência	508.623	End date of the agreement's validity
Valor Contrapartida	137.369	Amount of the recipient's contribution
Data Última Liberação	8.763	Date of the last resource release
Valor Última Liberação	1.275.212	Amount of the last resource release

4.2. Fine Tuning *Transformers* models for NER

To train the NER models was used the spaCy [Honnibal et al. 2020] structure due to its ability to integrate annotated data and its architecture that supports integration with several pre-trained *Transformers* models available on the Hugging Face platform [Wolf et al. 2019]. The training phase used the corpus of annotated publications and pre-trained *Transformers* models in other domains—this phase employed transfer learning through additional training of pre-trained models on the annotated Agreements corpus.

The training process was organized in the following format: Corpus Division - The annotated corpus was randomly divided into training (70%) and testing (30%) sets. Model Configuration and Training - Seven different Transformer models, detailed in Table 2, were configured and trained with the same hyperparameters (max. number of epochs = 200, learning rate = 5×10^-5, regularization weights = L2, optimizer = Adam, batch size = 32, dropout = 0.1). Training and

[2] Available at `https://portaldatransparencia.gov.br/`.

Evaluation - The models were trained using the Spacy library [Honnibal et al. 2020] with annotated data from the training set and evaluated on the test set using the ”scorer” tool from the same library. This tool provides evaluations in terms of precision, recall, and F1 score. For a more detailed understanding of these metrics, we reference [Dalianis and Dalianis 2018], which explores the evaluation of information retrieval and natural language processing systems, explaining these fundamental evaluation concepts.

Table 2. Models Details.

Model	Description	Reference
M1	Albert-base-v2	[Lan et al. 2019]
M2	bert-base-cased	[Devlin et al. 2018]
M3	Electra-base-discriminator	[Clark et al. 2020]
M4	bart-base,	[Lewis et al. 2019]
M5	legal-bert-ner-base-cased-ptbr	[Domingues 2022]
M6	bert-base-portuguese-cased	[Souza et al. 2020]
M7	xlm-roberta-base	[Conneau et al. 2019]

4.3. Ensemble Strategies

Three different approaches were adopted to create ensembles, but all were based on voting strategies to combine the results of the transformer models. These are presented in Table3 and detailed below:

Table 3. Ensemble Acronym, Ensemble Type and Decision Rule.

Ensemble	Type	Rule
E1	Best Model Voting	Highest F1 Score
E2	Majority Voting	Half plus 1 of the votes
E3	Weighted Voting	Weighted Average F1 Score greater than 0.8

4.3.1. Majority Voting

It is a strategy based on hard voting where an Agreement text published in the DOU is applied as input to all NER models to extract named entities, including their label and position in the text. The label and location of an entity are maintained if only most models agree on the label and the position in the text. Therefore, the strategy can be explained as follows:

- Input: An Agreement text T published in the DOU is analyzed by n NER models.
- Output of each model: Each model M_i (where $i = 1, 2, \ldots, n$) identifies named entities $E_{i,j} = (e_{i,j}, l_{i,j}, p_{i,j})$, where $e_{i,j}$ is the entity, $l_{i,j}$ is the label, and $p_{i,j}$ is the position in the text.
- Vote: - For each entity e identified, count the number of models k that agree on the label l and position p:

  $$k(e, l, p) = \sum_{i=1}^{n} \mathbf{1}_{(e_{i,j}=e)\wedge(l_{i,j}=l)\wedge(p_{i,j}=p)}$$

  where $\mathbf{1}$ is the indicator function worth one if the condition is true and 0 otherwise.
- Final decision: The entity e with label l and position p is maintained only if: $k(e, l, p) > \frac{n}{2}$.

4.3.2. Weighted Voting

In this strategy, based on soft voting, each model is initially evaluated in terms of its performance, measuring the F1 score for each entity class in the test set. These F1 scores range from 0.0 to 1.0 and reflect each model's ability to identify named entities accurately in different linguistic contexts. For each DOU publication referring to a public Agreement, its text is used as input in all ensemble models to extract named entities, including the entity's label and its location in the text. For each model that classifies an entity with the same label and position in the text, the F1 scores of the model for that entity are added. Afterward, the average is calculated by dividing the sum of the F1 scores of the models that classified the total by the total of models. The mean reflects the joint reliability of the classification.

A decision threshold is established to keep only entities with high reliability, in this case, 0.8, for the average F1 score. In other words, only entities that coincide in label and position in the text and whose average F1 scores of the models that classified them are greater than 0.8 are kept as the final result of the Ensemble. This way, you can follow the following steps:

- Entry:
  - A text T of an Agreement published in the DOU.
  - n NER models.
- Output of each model: Each model M_i (where $i = 1, 2, \ldots, n$) identifies named entities $E_{i,j} = (e_{i,j}, l_{i,j}, p_{i,j})$, where i represents the index of the model and j represents the index of the named entity within the set of entities that the model M_i identified in the text T. This way:
  - $e_{i,j}$ is the identified entity,
  - $l_{i,j}$ is the entity label,
  - $p_{i,j}$ is the position of the entity in the text.
- Soft Voting: Each entity is classified by all models. For each entity e with label l and position p, we calculate the F1 score of each model M_i for that entity.
- Sum of F1 scores: Let $F1_i(e, l, p)$ be the F1 score of the model M_i for the entity e with label l and position p. We sum the F1 scores of all models that classified the entity and with the same label l and position p:

$$S(e, l, p) = \sum_{i=1}^{n} F1_i(e, l, p)$$

- average F1 score: We calculated the average of the F1 scores of the models that classified the entity:

$$\bar{F}1(e, l, p) = \frac{S(e, l, p)}{n}$$

  where $\bar{F}1(e, l, p)$ is the average F1 score for the entity e with label l and position p.
- Final decision: A threshold τ is established for the average F1 score to maintain an entity with high reliability. The entity e with label l and position p is maintained if $\bar{F}1(e, l, p) > \tau$

4.3.3. Best Model Voting

It is a voting-based ensemble strategy. An Agreement text published in the DOU is used as input in all NER models to extract named entities, including their labels and positions in the text. The classification criterion is the F1 score value per entity for each model. Thus, the final classification of a named entity identified in a part of the text is determined by the model that has the highest value of the F1 score metric for the assigned entity category. This way, the entity label and position are selected based on the model with the best F1 score performance per entity. This way, you can follow the following steps:

- Entry: A text T of an Agreement published in the DOU. n NER models.
- Output of each model: Each model M_i (where $i = 1, 2, \ldots, n$) identifies named entities $E_{i,j} = (e_{i,j}, l_{i,j}, p_{i,j})$. $e_{i,j}$ is the identified entity, $l_{i,j}$ is the entity label, $p_{i,j}$ is the entity's position in the text.
- F1 score per model: Each model M_i has an F1 score $F1_i$ for the entity category $l_{i,j}$.
- Determination of the model with the highest F1 score: For each entity e identified with label l and position p, there is the model M_{i^*} with the highest F1 score for that entity category.

  $$i^*(e, l, p) = \arg \max_{i \in \{1,2,\ldots,n\}} F1_i(l_{i,j})$$

  where $i^*(e, l, p)$ represents the model index M_i that will be chosen to determine the entity (e, l, p) and $\arg\max_{i \in \{1,2,\ldots,n\}}$ is the operation that returns the value of the index i that maximizes the expression $F1_i(l_{i,j})$. The index i varies within the set $\{1, 2, \ldots, n\}$, where n is the total number of NER models. in turn, in $F1_i(l_{i,j})$, $F1_i$ represents the F1 score of the model M_i and $l_{i,j}$ is the label of the entity e identified by the model M_i.
- Final classification: The entity e with label l and position p in the text is determined by the model M_{i^*} that has the highest F1 score:

  $$E_{final} = \{(e, l, p) \mid (e, l, p) = (e_{i^*,j}, l_{i^*,j}, p_{i^*,j})\}$$

  Where E_{final} is the final set of named entities and $(e, l, p) = (e_{i^*,j}, l_{i^*,j}, p_{i^*,j})$ is the condition specifies that the entity e, its label l, and its position p are equal to an entity $e_{i^*,j}$, its label $l_{i^*,j}$, and its position $p_{i^*,j}$ identified by the model M_{i^*}.

4.4. Evaluating Results

A systematic approach evaluated the different models and ensembles in named entity recognition based on precision, recall, and F1 score metrics. Initially, entity categories were selected, and then precision (proportion of true positives among positive predictions), recall (proportion of true positives among truly positive cases), and F1 score (harmonic mean of precision and recall) were measured for each entity category. We reference [Dalianis and Dalianis 2018] for more details on these metrics. The best model or Ensemble was identified for each metric in general, as shown in Figure 3 and by category, as recorded in Figure 2a. The gains provided by entity category and general were calculated, highlighting the largest and smallest gains in precision, recall, and F1 score. We also sought to obtain how many entity categories each model or Ensemble stood out as the best in each metric. Figure 2b was used to quantify and provide a clear view of overall performance. Based on the data collected, a comparison of the performances of models and ensembles was carried out, interpreting the results to understand the advantages of using ensembles in recognizing named entities. It is expected that this seeks to reach a comprehensive and comparative assessment, clearly highlighting the benefits and limitations of each methodology used.

5. Results

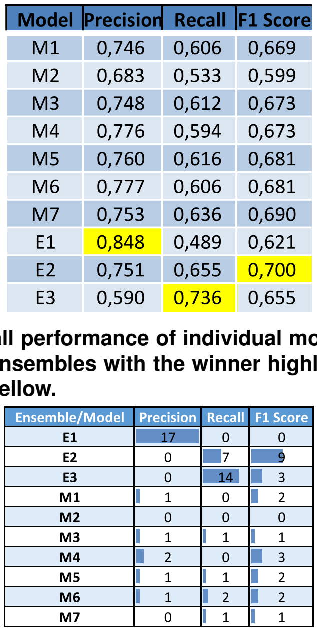
The experiments provided several insights into the performance of ensembles in general and across different categories of entities. It is important to highlight that the potential of ensemble methods for the recognition of named entities may be even greater, given that new experiments with optimization of model hyperparameters and adjustment of ensemble parameters, such as decision rules and thresholds [Kuncheva 2014]. To better understand the results, you should consult the identification of models in Table 2 and ensembles in Table 3. The general results in Figure 3 highlight ensembles' superiority over individual models in all metrics. The Ensemble E1 obtained the best precision (ents_p) with a value of 0.848. This value represents a maximum gain of up to 16.5% and a minimum of 7.1% for individual models. The increase in accuracy suggests that this Ensemble improved the correct identification of positive cases while minimizing false positives. Regarding recall (ents_r), the E3 ensemble achieved the best performance with a rate of 0.736. This performance indicates a significant increase of up to 20.3% compared to individual models, with the smallest gain being 10%. The high recall rate demonstrates the ability to detect most positive cases, reducing the number of false negatives. Finally, in the overall analysis, E2 obtained the best F1 Score (ents_f) of 0.700, with improvements ranging from 1% to 10.1% over the individual models and highlights how these ensembles promote more robust and balanced performances.

In the evaluation by category of entities, Figure 2a shows the best performances obtained by models or ensembles for each metric. It was observed that models E1 and E2 stood out in several categories, with E1 achieving the highest accuracy in several categories, such as "NOME CONVENENTE" (0.940), "NÚMERO CONVÊNIO" (0.999) and "NOME ÓRGÃO CONCEDENTE" (0.980). The E2 model, in turn, stood out in recall for several entities, such as "DATA FINAL VIGÊNCIA" (0.684), "DATA ÚLTIMA LIBERAÇÃO" (0.700) and "VALOR CONVÊNIO" (0.678). The E2 ensemble presented the greatest gain in the F1 score metric, with an increase of 4.3% in "NÚMERO ORIGINAL." The smallest gain for the F1 score was 0.2% in "TIPO INSTRUMENTO," also for the E2 ensemble. For recall, the largest gain was 23.1% in "VALOR CONVÊNIO" with the E3 ensemble, while the smallest gain was 0.1% in "NÚMERO CONVÊNIO", also for E3. Regarding the accuracy metric, the E1 ensemble showed a significant gain of 20.6% in "DATA FINAL VIGÊNCIA", with the smallest gain of 0.3% in "NOME ÓRGÃO SUPERIOR". These gains demonstrate that, although ensemble gains may vary, as shown in Figure 2a, they offer consistent improvements compared to individual models.

Figure 2b provides an overview of the number of entity categories each model, or Ensemble stood out, evaluated by the precision, recall, and F1 score metrics. The E1 model led to the best result in 17 categories in accuracy, indicating a strong ability to predict with high accuracy in various situations. In contrast, the E3 ensemble stood out as the best in 14 categories regarding recall, demonstrating its superior ability to identify positive cases correctly. These data show that ensembles are particularly effective in improving recall, essential for applications where correct identification of all positive cases is crucial. Furthermore, the ensembles offer substantial gains in accuracy and F1 score in several categories, demonstrating their versatility and robustness.

6. Conclusions and Further Work

The results indicate that, even without fine-tuning model hyperparameters and thresholds, the ensembles show general improvements in the three main metrics and specific improvements in most classes of named entities. This experiment also demonstrates that the choice of the optimal Ensemble depends on the particular evaluation priorities: to maximize accuracy and reduce false positives, models like E1 are suitable; to maximize recall and capture as many positive cases as possible, ensembles like E3 are more effective, and to balance precision and recall, ensembles like E2 are recommended.



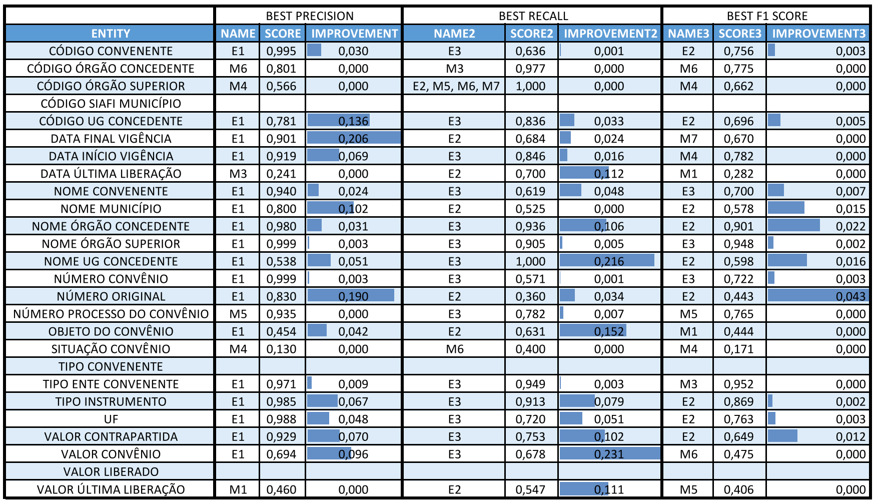
Model	Precision	Recall	F1 Score
M1	0,746	0,606	0,669
M2	0,683	0,533	0,599
M3	0,748	0,612	0,673
M4	0,776	0,594	0,673
M5	0,760	0,616	0,681
M6	0,777	0,606	0,681
M7	0,753	0,636	0,690
E1	0,848	0,489	0,621
E2	0,751	0,655	0,700
E3	0,590	0,736	0,655

(a) Overall performance of individual models and ensembles with the winner highlighted in yellow.

Ensemble/Model	Precision	Recall	F1 Score
E1	17	0	0
E2	0	7	9
E3	0	14	3
M1	1	0	2
M2	0	0	0
M3	1	1	1
M4	2	0	3
M5	1	1	2
M6	1	2	2
M7	0	1	1

(b) Number of named entity categories where each model or Ensemble performed best metrics.

Figure 2. For models and ensembles: (a) Overall best performance in and (b) Number of entity categories where it was the best .



	BEST PRECISION			BEST RECALL			BEST F1 SCORE		
ENTITY	NAME	SCORE	IMPROVEMENT	NAME2	SCORE2	IMPROVEMENT2	NAME3	SCORE3	IMPROVEMENT3
CÓDIGO CONVENENTE	E1	0,995	0,030	E3	0,636	0,001	E2	0,756	0,003
CÓDIGO ÓRGÃO CONCEDENTE	M6	0,801	0,000	M3	0,977	0,000	M6	0,775	0,000
CÓDIGO ÓRGÃO SUPERIOR	M4	0,566	0,000	E2, M5, M6, M7	1,000	0,000	M4	0,662	0,000
CÓDIGO SIAFI MUNICÍPIO									
CÓDIGO UG CONCEDENTE	E1	0,781	0,136	E3	0,836	0,033	E2	0,696	0,005
DATA FINAL VIGÊNCIA	E1	0,901	0,206	E2	0,684	0,024	M7	0,670	0,000
DATA INÍCIO VIGÊNCIA	E1	0,919	0,069	E3	0,846	0,016	M4	0,782	0,000
DATA ÚLTIMA LIBERAÇÃO	M3	0,241	0,000	E2	0,700	0,112	M1	0,282	0,000
NOME CONVENENTE	E1	0,940	0,024	E3	0,619	0,048	E3	0,700	0,007
NOME MUNICÍPIO	E1	0,800	0,102	E2	0,525	0,000	E2	0,578	0,015
NOME ÓRGÃO CONCEDENTE	E1	0,980	0,031	E3	0,936	0,106	E2	0,901	0,022
NOME ÓRGÃO SUPERIOR	E1	0,999	0,003	E3	0,905	0,005	E3	0,948	0,002
NOME UG CONCEDENTE	E1	0,538	0,051	E3	1,000	0,216	E2	0,598	0,016
NÚMERO CONVÊNIO	E1	0,999	0,003	E3	0,571	0,001	E3	0,722	0,003
NÚMERO ORIGINAL	E1	0,830	0,190	E2	0,360	0,034	E2	0,443	0,043
NÚMERO PROCESSO DO CONVÊNIO	M5	0,935	0,000	E3	0,782	0,007	M5	0,765	0,000
OBJETO DO CONVÊNIO	E1	0,454	0,042	E2	0,631	0,152	M1	0,444	0,000
SITUAÇÃO CONVÊNIO	M4	0,130	0,000	M6	0,400	0,000	M4	0,171	0,000
TIPO CONVENENTE									
TIPO ENTE CONVENENTE	E1	0,971	0,009	E3	0,949	0,003	M3	0,952	0,000
TIPO INSTRUMENTO	E1	0,985	0,067	E3	0,913	0,079	E2	0,869	0,002
UF	E1	0,988	0,048	E3	0,720	0,051	E2	0,763	0,003
VALOR CONTRAPARTIDA	E1	0,929	0,070	E3	0,753	0,102	E2	0,649	0,012
VALOR CONVÊNIO	E1	0,694	0,096	E3	0,678	0,231	M6	0,475	0,000
VALOR LIBERADO									
VALOR ÚLTIMA LIBERAÇÃO	M1	0,460	0,000	E2	0,547	0,111	M5	0,406	0,000

Figure 3. Best performance of models/ensembles by entity category. Metrics evaluated for precision, recall, and F1 score.

This work details the effectiveness of three different ensemble approaches, which can guide future decisions in model selection and combination to optimize performance in NER tasks. In future work, there is the possibility of adjusting the thresholds of these ensembles, as suggested in the literature, and exploring different combinations of models. Furthermore, strategies can be sought to balance the unbalanced entity classes in the corpus. In summary, ensemble methods offer significant benefits to NER, providing consistent gains in key metrics and improving the robustness and accuracy of classification systems in complex scenarios such as the one presented.

References

Albanaz, J. O. L. (2020). Reconhecimento de entidades nomeadas em resultados de licitações publicados em diários oficiais.

Alles, V. J., Giozza, W. F., and de Oliveira Alburquerque, R. (2018). Natural language processing to classify named entities of the brazilian union official diary.

Belém, F. M., Ganem, M., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforc¸o e delimitação contextual para reconhecimento de entidades e relações em documentos oficiais.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

da Silva, M. G. (2022). Reconhecimento de entidades nomeadas em documentos de editais de compras utilizando aprendizado profundo.

Dalianis, H. and Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical Text Mining: secondary use of electronic patient records*, pages 45–53.

de Araujo, P. H. L., de Campos, T., de Oliveira, R. R. R., Stauffer, M., Couto, S., and de Souza Bermejo, P. H. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text.

de Carvalho, L. R., Mendes, F. L., Chaves, J., Lima, M. C., de Deus, F. E. G., Araújo, A. P., and de Barros Vidal, F. (2022). Deep-vacuity: A proposal of a machine learning platform based on high-performance computing architecture for insights on government of brazil official gazettes. In *WEBIST*, pages 136–143.

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

Domingues, M. (2022). Language model in the legal domain in portuguese. `https://huggingface.co/dominguesm/legal-bert-base-cased-ptbr/`.

Guimarães, G. M. C., Silva, F. M., Queiroz, A. L., Marcacini, R. M., Faleiros, T., Borges, V. R. P., and García, L. (2024). Dodfminer: An automated tool for named entity recognition from official gazettes.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python.

Khan, A. A., Chaudhari, O., and Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation.

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J., Sun, A., Han, J., and Li, C. (2022). A survey on deep learning for named entity recognition.

Possamai, A. J. and de Souza, V. G. (2020). Transparência e dados abertos governamentais: Possibilidades e desafios a partir da lei de acesso À informação.

Rodríguez, M. M. and Bezerra, B. L. D. (2020). Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias).

Rouhizadeh, H. and Teodoro, D. (2022). Ds4dh at semeval-2022 task 11: Multilingual named entity recognition using an ensemble of transformer-based language models.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley-Blackwell*, 8(4).

Silva, F. M., Guimarães, G., Rezende, S. O., Queiroz, A. L., Borges, V. R. P., Faleiros, T., and García, L. (2022). Named entity recognition approaches applied to legal document segmentation.

Singh, A., Singh, S. S., and Tiwary, U. S. (2023). Enhancing hindi named entity recognition through ensemble learning.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Sun, J., Tang, R., Xiang, L., Zhai, F., and Zhou, Y. (2021). Multi-strategy fusion for medical named entity recognition.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient transformers: A survey.

Wang, Z., Wu, Y., Lei, P., and Cheng, P. (2020). Named entity recognition method of brazilian legal text based on pre-training model.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zheng, J. and Sun, J. (2023). Ensembles of bert models for ancient chinese processing.