Assessing European and Brazilian Portuguese LLMs for NER in specialised domains

Rafael Oleques Nunes[1 [0009−0007−8842−421X]], Joaquim Santos[2 [0000−0002−0581−4092]], Andre Spritzer[1 [0009−0002−4232−1585]], Dennis Giovani Balreira[1 [0000−0002−0801−9393]], Carla Maria Dal Sasso Freitas[1 [0000−0003−1986−8435]], Fernanda Olival[3 [0000−0003−4762−3451]], Helena Freire Cameron[3 [0000−0001−7719−6994]], and Renata Vieira[4 [0000−0003−2449−5477]]

[1] Federal University of Rio Grande do Sul, Brazil
{ronunes,dgbalreira,spritzer,carla}@inf.ufrgs.br
[2] University of Vale do Rio dos Sinos, Brazil
nejoaquim@edu.unisinos.br
[3] CIDEHUS - Portalegre Polytechnic University, Portugal
helenac@ipportalegre.pt
[4] CIDEHUS - University of Évora, Portugal
{mfo,renatav}@uevora.pt

Abstract. This paper discusses the impact of Portuguese variants in Large Language Models for the task of named entity recognition (NER) in specialised domains. The tests were made on a Brazilian Portuguese legal and a European Portuguese historical corpora. The models taken into account are BERTimbau (PT-BR), Albertina (PT-PT and PT-BR), and XML-R (multilingual). The impact was more evident in the Portuguese historical corpus, which resulted in higher F1 measures compared to previous works that did not consider the same language variant. Additionally, the study underscores the impact of model architecture on performance, highlighting the critical role of both linguistic alignment and model size in enhancing NER in specialised domains.

Keywords: Named entity recognition · Large Language Models · Portuguese language variants.

1 Introduction

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP). The objective of NER is to identify and classify specific terms in a sentence, such as the names of people, organisations, locations, dates, and other entities. This task is fundamental for various NLP applications, including information extraction, question answering, and automatic text summarisation. Although many studies and models have been developed with good results, most have focused on general entities, relying on datasets such as CoNLL 2002 for English [23] and HAREM for Portuguese [17].

However, specialised domains often cannot utilise these general classifiers effectively because of domain-specific language or the need to identify different

types of entities not commonly present in general datasets. For instance, while entities such as gene names and diseases are crucial in the biomedical domain, in legal texts, entities such as law citations and court names are more relevant. Previous studies have addressed these challenges by delving into specific domains and proposing tailored datasets and models [20, 18, 13].

As specialised models on Portuguese variants became recently available, [21, 15], this study evaluates the performance of these models on specialised texts written in two variants of the Portuguese NER annotated corpus (European and Brazilian). We aim to compare how these models perform in these variants, considering two specialised domains, historical and legislative, that present unique challenges and require domain-specific knowledge.

By advancing NER capabilities in domain-specific Portuguese texts in two variations of the language, we aim to facilitate more accurate and contextually relevant information extraction, supporting various applications, from historical research to legal document processing. The legislative and historical fields have the advantage of being comprehensive despite some specialization. This study contributes to the academic understanding of domain-specific NER and provides practical insights and tools for improving NLP applications in specialised fields.

Our main contributions are: (i) a comprehensive evaluation of different models on NER tasks within Portuguese texts, (ii) an analysis of the differences in model performance when pre-trained with European and Brazilian variants of the Portuguese language, and (iii) insights into applying these models across different specialised domain corpora: legal and historical.

2 Related Work

Named Entity Recognition (NER) involves identifying and classifying entities such as locations, organisations, and persons within a text. General entities have been extensively explored in various languages and corpora. Popular datasets for evaluating NER in different languages include CoNLL 2002 for English [23] and HAREM for Portuguese [17]. These datasets have been instrumental in advancing NER research and establishing model performance benchmarks.

However, the classifiers trained on these datasets are not always sufficient for all domains. For example, Silva et al. (2023) [20] proposed a dataset focused on *cachaça*, a distilled spirit made from sugarcane juice, where important entity categories include classification, price, storage time, and sensory characteristics, such as colour. These particular entities cannot be adequately represented using only the traditional NER categories.

The legal domain has also been found to benefit from the use of specialised entities, such as legal citations from legislation or court cases [3]; products of law and legal basis [2]; courts, origins of legal procedure, and trial dates [7]; and decisions and sentences [5]. Recent studies have explored these areas using various techniques with transformer models, including domain adaptation [4, 9], fine-tuning [25], in-context learning [14], and self-learning [13].

Historical texts represent another important domain with unique characteristics. Studies focusing on NER tasks in historical texts [10, 24, 18] have highlighted the challenges posed by different word spellings and sentence structures compared with contemporary texts in the same language.

All the works discussed so far have focused on NER tasks within the context of Portuguese language texts. In this study, we extend this focus by evaluating the impact of Portuguese-specific models on specialised texts, specifically within the historical and legal domains, exploring both European and Brazilian language variants.

3 Corpora

3.1 European Portuguese Annotated Historical Texts

The first *sub-corpus* considered in this study is a subset of the historical PT-Eu *corpus*, the *Parish Memories* collection. It is an 18^{th}-century *corpus* composed of the answers to a survey sent in 1758 to the Portuguese priests to obtain feedback about the state of the territory after the 1755 Lisbon big earthquake and to gather information to form a Geographical Dictionary of Portugal. The sub-corpus under analysis contains 71 Memories of the municipalities of the main cities in the largest region in Portugal, Alentejo, i.e., Portalegre, Évora and Beja. We also added the Memories from the municipality of Vila Viçosa, a historically relevant municipality, dominated by the House of Bragança. The original manuscripts have been transcribed and normalised to contemporary standard European Portuguese spelling. The *sub-corpus* was annotated with named entities, customised to the historical reality.

In a first approach (Anonymous, 2021) three basic categories were considered (*person, local, organisation*), as, for a historian, they aim to answer the main questions: *Who, Where, and When*. After this initial approach, considering the need to describe past realities better, a *corpus*-based study was conducted to define extensions of these categories (Anonymous, 2022) according to their relevance to the historian's inquiry. For that, the main category of *person* was broken down into several subcategories.

The category *person* (PER) refers to references made by name, first name, and family name (PER_NAM); occupation (PER_OCC); or social category (PER_CAT). All these attributes reflect the hierarchical structure of the 18^{th}-century Portuguese society as, frequently, titles and occupation positions were almost part of a person's name and identity. An example of mentions to persons by occupation is *Reitor da Universidade de Évora* (Rector of the University of Évora).

Still related to the PER category, and because they constitute specific details of *person*, we established other subcategories to label saints (PER_SAINT); divinities (PER_DIV); groups of persons (PER_PGRP); and authors (PER_AUT). The subcategory for groups of persons is used to annotate organic groups, families, and members of an organisation, among others. Monges Cartuxos (Carthusians monks) and Sarracenos (the Saracens) are examples of this category.

Concerning the *local* category, we generalised it to place (PLC). This category includes geopolitical entities (PLC_GPE), aquifers(PLC_AQU), mountains (PLC_MOUNT), facilities (PLC_FAC), and one extra subcategory for other locations (PLC_LOC) as for instance, Bispado de Portalegre (Bishopric of Portalegre, PLC_GPE) and Rio Guadiana (Guadiana river,PLC_AQU).

Geopolitical entities were included to avoid ambiguities among locations and organisations, as this categories aggregates them indistinctly. Other references to geographical points, such as rivers and mountains, are essential for geo-references.

The remaining categories are for organisation, time, and authored work. The ORG category labels several organisations as, for example, Companhia de Jesus (Society of Jesus), and Universidade de Coimbra (University of Coimbra).

For TIM_CRON, we only annotated specific references to dates, for instance, *o ano de 1755* (the year of 1755). We also established a category, AUTWORK, that allows us to treat the text sources mentioned in the *corpus* to recognise the text sources mentioned by priests.

An extended set of NE categories to account for past ages also implies more complexity in annotation and their computational processes. All the documents of the *sub-corpus* were manually annotated based on the consensual judgment of four annotators, and it was made using the INCEPTION platform[5].

As can seen in Table 1, we have 5,031 annotated NEs. The major classes are geo-political entities, person names, and saints. Persons are referenced only by category, and mountains are the least represented. For training, development, and testing, the distribution is 70, 10, and 20%.

Table 1. Frequency of named entities in the Parish Memories for each type.

CATEG	Train	Dev	Test	Overall NE
AUTWORK	106	12	19	137
ORG	287	52	54	393
PER_AUT	101	13	15	129
PER_CAT	37	4	8	49
PER_DIV	119	25	40	184
PER_NAM	520	62	136	718
PER_OCC	88	11	25	124
PER_PGRP	153	25	21	199
PER_SAINT	435	76	133	644
PLC_AQU	147	13	68	228
PLC_FAC	202	18	69	289
PLC_GPE	785	84	232	1101
PLC_LOC	336	24	87	447
PLC_MOUNT	50	10	13	73
TIM_CRON	217	33	66	316
Total	3,583	462	986	5031

[5] https://inception-project.github.io

3.2 Brazilian Portuguese Annotated Legislative Texts

Table 2. Frequency of named entities in UlyssesNER-Br for each type.

CATEG	Train	Dev	Test	Overall NE
DATA	433	72	98	609
EVENTO	9	5	9	23
FUNDapelido	123	24	34	181
FUNDlei	359	81	85	522
FUNDprojetodelei	8	2	5	15
LOCALconcreto	333	139	88	560
LOCALvirtual	36	6	13	55
ORGgovernamental	324	60	68	452
ORGnaogovernamental	88	10	22	120
ORGpartido	23	11	4	38
PESSOAcargo	224	38	40	302
PESSOAgrupocargo	121	17	20	158
PESSOAindividual	283	59	59	401
PRODUTOoutros	173	38	42	253
PRODUTOprograma	42	6	11	59
PRODUTOsistema	15	2	1	18
Total	2594	570	599	3763

The legislative *corpus* used in this work is UlyssesNER-Br [2]. It comprised 150 bills from the Brazilian Chamber of Deputies (BCoD). Annotations were performed in three phases by two undergraduate students and one graduate student as a curator.

UlyssesNER-Br [2] included both coarse and fine-grained levels. The coarse-grained level comprises seven entity categories, whereas the fine-grained level includes eighteen entity types. The entities follow the traditional HAREM [16] entities (*person, location, organisation, event,* and *date*) with the addition of domain-specific entities such as law foundations and law products.

The entity PESSOA (person) was specialised in three types: PESSOAindividual (individual), PESSOAcargo (occupation), and PESSOAgrupocargo (group of occupations). In this division, it is possible fine-grained levels of person citations, such as Deputado HILDO ROCHA (Deputy HILDO ROCHA), and 16 de março de 2011 [March 16, 2011].

FUNDAMENTO (law foundation) refers to various legal entities such as laws, bills, and legislative consultations requested by congressmen. This category includes fine-grained entities, such as FUNDlei (legal norm), FUNDapelido (legal norm nickname), and FUNDprojetodelei (bill). Examples of fine-grained entities include art. 34 do Estatuto do Idoso (art. 34 of the Elderly Statute), and Código Brasileiro de Trânsito (Brazilian Traffic Code).

The final category, PRODUTODELEI (law product), pertains to anything created due to legislation. This class also includes three fine-grained types: PRO-

DUTOsistema (system product), PRODUTOprograma (program product), and PRODUTOoutros (other products). Examples of each fine-grained class are Sistema Único de Saúde (Unified Health System), and salário mínimo (Minimum wage).

4 Framework and Models

We adopted the Flair[1] framework, a NER library for multiple languages developed in PyTorch[6]. This framework provides pre-trained language models, named entity recognition models, and neural networks for language model training and sequence tagging. With Flair, we can construct pipelines for training token classifiers and feed them with various types of language models, such as Word Embeddings, Transformer-based models, and Flair Embeddings itself.

Herein, we analyse four versions of language models.

BERTimbau [21] is a pre-trained transformer-based language model trained specifically for Brazilian Portuguese. It was trained on the *brWaC corpus*[8], which amounts to a total of 2.6 billion tokens, resulting in 17.5 GB of preprocessed data. BERTimbau was trained using token masking in input sentences. In other words, it is a Masked Language Model (MLM). We chose this model because the current state-of-the-art[22] in NER for Portuguese uses this model. We used the Large version of BERTimbau [7].

Albertina PT-* [15] is a large language model designed explicitly for Portuguese. It functions as an encoder within the BERT family and is built upon the DeBERTa model using the Transformer neural architecture. Albertina PT- has two variants: Albertina PT-PT and Albertina PT-BR. Both variants are distributed free of charge, under a permissible license.

Albertina PT-PT is the European Portuguese version. This model is available in three sizes, specifically with 1.5 billion parameters, 900 and 100 million parameters. The pre-training *corpora* of the Albertina PT-PT 1.5B comprises general and legislative domains.

Albertina PT-BR focuses on Brazilian Portuguese. Its largest version, Albertina 1.5B PT-BR [19], has a more permissive licensed model without using the BrWac dataset, consisting of a 36 billion token dataset compiled from a multilingual *corpus*. Since this *corpus* includes both European and Brazilian Portuguese, additional filtering was applied to retain only documents with metadata indicating Brazil's internet country code top-level domain.

XLM-RoBERTa [6] is a multilingual model designed to understand 100 languages without requiring language-specific tensors, as it can identify the language directly from the input identifiers. It incorporates techniques from RoBERTa [12] into the XLM framework [11], focusing solely on masked language modeling for single-language sentences, and does not employ translation language modeling.

Table 3 compares the four described models according to some of their features.

[6] https://pytorch.org/

[7] https://huggingface.co/neuralmind/bert-large-portuguese-cased

Table 3. Comparison of the four Portuguese Large Language Models.

Feature	Albertina PT-PT	Albertina PT-BR	XLM-R	BERTimbau
Params	1.5B	1.5B	550M	355M
Corpus	CulturaX,DCEP,Europarl,ParlamPT	CulturaX	Multi-data	brWaC
Arch.	DeBERTa,24L,16H	DeBERTa,24L,16H	Trans,24L,16H	BERT,24L,16H
Lang.	PT (PT)	PT (BR)	100 languages	PT (BR)
Domain	General,Legislative	General	General	General
Year	2023	2023	2019	2021

5 Experimental Evaluation

We conducted our experiments on a GPU A100 with 80GB of RAM and an RTX 4090 with 64GB of RAM. The experiments used Python 3.7.6 and the Flair library to use pre-trained models.

Hyperparameters were set to the default values recommended by the library: a learning rate of 5e-5, a mini-batch size of 4, and training for 10 epochs. Truncation was applied to the maximum length, and padding was set to true.

We employed standard metrics for model evaluation, including accuracy, precision, recall, and micro F1-score using the CoNLL-2002 script [23]. These metrics provided a comprehensive assessment of the model's performance across different classes, ensuring a thorough evaluation of its effectiveness.

6 Results and Discussion

6.1 Overall Results

First, we detail the results for each corpus individually. Then, we combine the results for comparative analysis, highlighting the influence of linguistic context and textual domain on the performance of specialised NER models.

Model	Precision	Recall	F1	$\Delta \uparrow$
Albertina PT-PT	**72.76**	**76.10**	**74.39**	+3.02
Albertina PT-BR	69.71	73.11	71.37	+0.61
XLM-R-Large	68.31	73.38	70.76	+0.23
BERTimbau-Large	67.36	74.00	70.53	*bl*

Table 4. Parish Memories models results.

Parish Memories corpus. Table 4 presents the overall results, highlighting that Albertina, trained in European Portuguese, achieved superior F1-Score results compared to other models. This success is likely not solely due to Albertina having more parameters than previous models but primarily due to its pre-training on European Portuguese texts. Given that, although the corpus consists of 18^{th}-century Portuguese texts, they were used in their normalized

version in European Portuguese, and this linguistic alignment may likely have contributed significantly to Albertina's performance.

Another notable observation concerns the differences in F1-Scores, as indicated in the column $\Delta \uparrow$. The largest discrepancy among the models' results is observed with European Portuguese Albertina, which markedly outperforms its Brazilian Portuguese counterpart.

The recall metric also yields insightful conclusions. BERTimbau-Large exhibits the highest recall among Brazilian Portuguese and multilingual models, surpassing Albertina and XLM-R-Large, suggesting that its smaller architecture can achieve comparable entity identification. However, precision trends align with model size, indicating that larger models tend to identify true positive instances more precisely. Leveraging a model trained on the same Portuguese variant as the *corpus* consistently yields the best results across all metrics.

Model	**Precision**	**Recall**	**F1**	$\Delta \uparrow$
Albertina PT-PT	86.83	**91.32**	**89.02**	+0.08
Albertina PT-BR	**87.93**	89.98	88.94	+0.18
BERTimbau-Large	84.14	90.32	87.12	+1.18
XLM-R-Large	82.39	89.82	85.94	*bl*

Table 5. UlyssesNER-Br types level models results.

Brazilian Legislative Texts. Table 5 presents the results for the fine-grained level in the UlyssesNER-Br *corpus*. It highlights that larger models can achieve better results, but the contribution of language specificity to enhanced performance remains inconclusive. The challenge of language specificity is evident in Table 5, where the increment in F1-Score for the two best models (the variations of Albertina) is only 0.08, possibly due to the legislative data during the pre-training of European Portuguese Albertina (see Table 3), or possibly to random classifier initialisation and result fluctuations.

Comparing the models, a significant finding is BERTimbau's high recall compared to larger models, underscoring its effectiveness in accurately identifying positive instances within specific legislative contexts, even if it is the smallest model. However, its lower precision suggests potential misses in positive instances.

Larger models, such as XLM-R and Albertina, excel in capturing fine-grained categories. While BERTimbau demonstrates superior recall in fine-grained analysis, the trade-off between precision and recall results in larger models achieving better F1-Score results with comparable recall. The Albertina models achieve the best balance between precision and recall. This balance indicates their effectiveness in both capturing a large number of relevant instances and maintaining accuracy.

Comparative analysis. The linguistic variation between Brazilian Portuguese and European Portuguese differed significantly across the two *corpora*, highlighting how the *Parish Memories corpus* could leverage models tailored to

its specific variant more effectively than Brazilian legislative texts. One plausible explanation lies in the domain and structure of the texts.

The *Parish Memories corpus* comprises letters written by priests with varying levels of academic background and formality. These texts differ from contemporary European Portuguese but maintain significant linguistic proximity, especially in their textual references. While the entities mentioned refer to things, people, and events of the 18^{th}-century, many of these references are still used and mentioned today, such as place names like Coimbra and Évora and local parish names and devotions that continue in the country. Therefore, it is reasonable to argue that a model trained on European Portuguese is better equipped to extract entities from this *corpus*.

Conversely, even though the Brazilian legislative *corpus* contains contemporary texts, many of these details were not necessarily learned from general texts in either Portuguese variant by the structure and specific jargon of legislative texts. Thus, it becomes that, for this *corpus*, the model size was more decisive than the linguistic variation.

Our results and analysis illustrate how linguistic context and textual domain play a crucial role in shaping the selection and performance of NER models. It underscores the importance of meticulously evaluating the *corpus* to select a suitable model that optimizes the extraction of named entities within specific contexts.

6.2 Results by Categories

CATEG	BERTimbau			XLM-R			Albertina (BR)			Albertina (PT)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AUTWORK	**45.83**	**52.38**	**48.89**	47.83	55.00	51.16	55.00	**52.38**	53.66	**70.00**	**66.67**	**68.29**
ORG	**48.05**	67.27	56.06	53.23	**55.93**	**54.50**	57.14	58.18	57.66	**64.29**	**65.45**	**67.86**
PER_AUT	77.78	**87.50**	**82.35**	78.95	**93.75**	85.71	**75.00**	**93.75**	83.33	**83.33**	**93.75**	**88.24**
PER_CAT	**87.50**	87.50	**87.50**	50.00	**75.00**	60.00	**38.89**	87.50	**53.85**	53.33	**100.00**	69.57
PER_DIV	76.74	82.50	79.52	**69.57**	**80.00**	**74.42**	82.50	82.50	82.50	**87.80**	**90.00**	**88.89**
PER_NAM	**61.04**	**67.63**	**64.16**	66.23	71.83	68.92	61.88	71.22	66.22	**68.59**	**76.98**	**72.54**
PER_OCC	**44.12**	**60.00**	**50.85**	60.71	62.96	61.82	55.17	64.00	56.26	**70.37**	**76.00**	**73.08**
PER_PGRP	**50.00**	**61.90**	**55.32**	55.17	**76.19**	64.00	**69.57**	**76.19**	**72.73**	**69.57**	**76.19**	**72.73**
PER_SAINT	77.37	**79.10**	**78.23**	**75.69**	78.99	**77.30**	**78.79**	**77.61**	78.20	77.21	78.36	77.78
PLC_AQU	**66.20**	**67.14**	**66.67**	72.73	**76.71**	74.67	**81.25**	74.29	**77.61**	80.00	74.29	77.04
PLC_FAC	65.33	**67.12**	**66.22**	**59.52**	66.67	**62.89**	**68.12**	**64.38**	66.20	67.14	**64.38**	65.73
PLC_GPE	**77.87**	**81.55**	**79.66**	78.84	**77.87**	78.35	**79.22**	78.54	78.88	78.45	78.11	**78.28**
PLC_LOC	65.35	74.16	69.47	60.00	72.53	65.67	**55.24**	**65.17**	**59.79**	**65.38**	**76.40**	**70.47**
PLC_MOUNT	**56.25**	**69.23**	**62.07**	75.00	**92.31**	82.76	75.00	**92.31**	82.76	**80.00**	**92.31**	**85.71**
TIM_CRON	69.33	**77.61**	**73.24**	66.67	**65.71**	**66.19**	**70.00**	73.13	71.53	**65.28**	70.15	67.63

Table 6. Parish Memories results per entity.

We present the results for each *corpus* at the entity level. We discuss how each model learned each entity and what is the influence of specialised entities. Subsequently, we present a comparative analysis to conclude how different levels of entities can help or affect the final model results.

Parish Memories. Table 6 shows the results at the entities level. Albertina in European Portuguese tends to consistently achieve the best results in F1-Score in most categories, pointing out that the specific features of the language in the model can be advantageous to the *corpus*. Models like Albertina (PT) and (BR) show a better balance between precision and recall, leading to higher F1 scores across many categories. This balance is important for practical applications where false positives and negatives must be minimised.

The table also highlights that Albertina (PT) achieved the highest results, with no F1-Scores near 50%, unlike a random classifier. In contrast, BERTImbau had three classes with scores near 50% and one with even lower results. Additionally, it is noteworthy that Albertina (BR) had five classes with results near 50% (the highest number among the models). However, the overall result compensated for the higher number of classes with F1 scores of 70% or more.

We observe specific trends and outliers in the F1-Scores in a bird's eye analysis of the categories in Table 6. In AUTWORK and PLC_MOUNT, the result increased when the model size was increased (see Table 3 for reference on model sizes), and the best result was in the Portuguese European Albertina model.

The Albertina (PT) model consistently performs well in identifying specialised persons. However, an outlier emerges in the PER_CAT (social category), where BERTimbau exhibits a notable increase of 17.93 points compared to the second-best result achieved by Albertina (PT). This unexpected outcome could be attributed to BERTimbau's pre-training on the brWaC corpus [8], which includes a diverse range of content potentially encompassing social category data. This may have enhanced BERTimbau's ability to accurately recognize entities within social contexts. Additionally, while BERTimbau shows a slight improvement in the PER_SAINT category with a marginal increase of 0.45 points, this improvement is not significant enough to decisively conclude it performs better than Albertina in this category.

Similarly, place entities consistently performed well with the Albertina (PT) model, often achieving results that were either the best or very close to it. For instance, in PLC_AQU, PLC_FAC, and PLC_GPE, Albertina (PT) demonstrated high precision and recall, with marginal differences from the top performer ranging from 0.57 to 1.38 (the largest difference observed). This suggests Albertina (PT) can identify and classify various place-related entities within the *corpus*.

However, it's notable that BERTimbau occasionally surpassed Albertina (PT) in specific categories, such as TIM_CRON, indicating its capability to excel in contexts where temporal references are crucial. These nuances of performance highlight the strengths of each model in different entity categories. It emphasises the importance of considering specific contexts and the balance in the results distributions in evaluating effectiveness across varied entity types.

CATEG	BERTimbau			XLM-R			Albertina (BR)			Albertina (PT)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DATA	**100.00**	**94.23**	**97.02**	**96.00**	96.97	97.92	95.92	**96.91**	98.00	**100.00**	**98.99**	**98.99**
EVENTO	**77.77**	**100.00**	**87.50**	87.50	77.78	82.35	**100.00**	**66.67**	80.00	85.71	**66.67**	**75.00**
FUNDapelido	**94.12**	**100.00**	**97.00**	**90.91**	**88.24**	**89.55**	93.75	**88.24**	90.91	**93.75**	**88.24**	90.91
FUNDlei	**98.82**	90.32	**94.38**	**74.51**	**89.41**	**81.28**	91.11	96.47	93.71	90.22	**97.65**	93.79
FUNDprojetodelei	**20.00**	**100.00**	33.33	**100.00**	**20.00**	33.33	**100.00**	40.00	**57.14**	50.00	**20.00**	**28.57**
LOCALconcreto	**94.32**	**84.69**	**89.25**	**87.37**	94.32	90.71	87.50	**95.45**	91.30	89.36	**95.45**	**92.31**
LOCALvirtual	38.46	**23.81**	**29.41**	**44.44**	**61.54**	**51.61**	**33.33**	46.15	38.71	**44.44**	**61.54**	**51.61**
ORGgovernamental	**82.35**	**78.87**	**80.58**	**77.92**	**88.24**	82.76	80.82	86.76	83.69	80.00	**88.24**	**83.92**
ORGnaogovernamental	90.91	80.00	85.11	**77.27**	**77.27**	**77.27**	**94.44**	**77.27**	85.00	90.00	**81.82**	**85.71**
ORGpartido	**100.00**	**66.67**	**80.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
PESSOAcargo	**97.50**	**90.70**	93.98	88.64	**97.50**	92.86	92.86	**97.50**	**95.12**	**86.67**	**97.50**	**91.76**
PESSOAgrupocargo	**95.00**	90.48	**92.68**	90.48	**95.00**	**92.68**	94.74	**90.00**	92.31	**85.71**	**90.00**	**87.80**
PESSOAindividual	96.61	**96.61**	96.61	**96.72**	**100.00**	**98.33**	96.61	**96.61**	96.61	**93.55**	98.31	**95.87**
PRODUTOoutros	**76.19**	**64.00**	69.57	**53.33**	76.19	**62.75**	**66.67**	**80.95**	**73.12**	64.71	78.57	70.97
PRODUTOprograma	**54.55**	**75.00**	**63.16**	**100.00**	**54.55**	**70.59**	**100.00**	**54.55**	**70.59**	**100.00**	**54.55**	**70.59**
PRODUTOsistema	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Table 7. UlyssesNER-Br type level results per entity.

Brazilian Legislative texts. Table 5 shows the results at the entity level. Similar to the conclusions drawn in Section 6.1, we observe that top results are well-distributed among the models BERTimbau, XLM-R, Albertina (BR), and Albertina (PT), with 4, 6, 6, and 8 instances of achieving the best scores, respectively. Additionally, while there are several instances of closely competitive results across models, having the worst result in a category does not imply poor performance, as evidenced by five instances where the worst result still exceeds 89%. This demonstrates the models' effective learning of the representations of many target entities.

Regarding the specific classes, Date (DATA) is notably well learned by all the models, with the minimum F1-Score varying between 97.02% and 98.99%. It is an expected result since Date is the largest class, with 433 examples in the training set, and carries specific patterns in its format.

The class Event (EVENTO) was also learned by all models, although results exhibited greater variability compared to the Date class, ranging from 75.00% to 87.50%. Event is a minority class in the training set, comprising only 9 examples. BERTimbau Large stood out by achieving the highest performance for this class. This contrasts with previous findings [13], which demonstrated that BERTimbau Base failed to learn Event, resulting in an F1-Score of 0%. This observation supports the hypothesis discussed in Section 6.1 that model size, rather than idiomatic variation, plays a crucial role in performance on this *corpus*.

The location (LOCAL) classes exhibited varying results and trends. First, concrete places (LOCALconcreto) were well-learned entities, with F1 Scores ranging from 89.25% to 92.31%, demonstrating an increase in performance with larger models. Concrete places have specific semantics related to geographical landmarks, which aid the models in identifying context and patterns.

In contrast, virtual places (LOCALvirtual) performed worse, with F1 Scores between 29.41% and 51.61%. A virtual place is a more diverse entity encom-

passing different types of locations, which do not necessarily share the same meaning and can include entities such as newspapers and internet pages [16], making it more challenging for the models to learn. Additionally, annotations in this broader category may lead to semantic conflicts. For example, the term *Jornal Diário Catarinense* (Diário Catarinense newspaper) in the sentence *Destaco que nos anos 90, o Jornal Diário Catarinense realizou uma pesquisa popular que colocou entre os 20 catarinenses do século*[8] could be understood as both a location and a general organisation [14], which is challenging for the model to discern.

organisations (ORG) showed a slight variation in results across different models. The results of the specialised classes of organisations highlight the relationship between the number of training examples in each class and their respective F1 Scores. As seen in Table 2, governmental organisations (ORGgovernamental), non-governmental organisations (ORGnaogovernamental), and political parties (ORGpartido) had 324, 88, and 23 examples in the training set, respectively. Despite the number of examples, the results were inversely related, with the minority class achieving the highest F1-Score and the majority class also achieving a high score, as shown in Table 5. One possible explanation for this behaviour is the specificity of the class. For instance, although the governmental organisation class pertains to a specific domain, it encompasses a wide range of entities, from municipal guards to the Chamber of Deputies. The same applies to non-governmental institutions. Conversely, political parties typically appear in more specific contexts or share similar characteristics in their names, such as following the name of a deputy or including the word "party" in their titles.

The classes related to Law Foundation — Legal Norm Nickname (FUNDapelido), Legal Norm (FUNDlei), and Law Proposals (FUNDprojetodelei) — exhibited varying levels of learning. The discrepancies in results are likely attributed to the number of training examples available for each class. For example, Legal Norm Nickname and Legal Norm achieved high F1 Scores of 97.00% and 94.38%, respectively, with 123 and 359 examples. In contrast, Law Proposals, despite having a specific format, such as the example “PEC 187/2016”, had only 8 examples in the training set. The highest results were concentrated in BERTimbau in Law Foundation classes.

The specialised classes of law products (PRODUTO) exhibited a trend of improved performance with larger models. This was particularly evident in the other products class (PRODUTOoutros), which achieved its best results with Albertina (BR). This class showed strong performance, specifically in the Portuguese models, with the highest result being the Brazilian Portuguese version of Albertina. Regarding the program product class (PRODUTOprograma), the best results were shared among XLM-R and both variations of Albertina, with identical scores across all metrics. This consistency necessitates a more detailed investigation into the class distribution and model learning. Lastly, the system product class (PRODUTOsistema) had only one example in the test set, and the

[8] English translation: *I highlight that in the 90s, Jornal Diário Catarinense carried out a popular survey that placed it among the 20 Santa Catarina citizens of the century.*

same term appeared in the training set, making it difficult to determine whether the 100% F1-Score reflects a well-learned class or if the model specialised in recognising this specific term.

Comparative analysis. Regarding the models, Albertina (PT) demonstrated superior performance in the Parish Memories corpus due to its European Portuguese training, particularly in recognising person and place entities. BERTimbau showcased strengths in identifying entities influenced by its diverse pre-training data, achieving comparable results to larger models, especially in the Brazilian Legislative Texts. XLM-R and Albertina (BR) also showed competitive performance across various entity types, highlighting their versatility.

The focus on European Portuguese-specific entities allowed Albertina (PT) to outperform other models in the Parish Memories *corpus*. Additionally, increasing the model size significantly improved performance in this dataset.

In the Brazilian Legislative Texts *corpus*, the diversity and specificity of the entities, coupled with the number of training examples, greatly influenced the models' performance. Larger models generally performed better, especially in minority and specific entity classes. Moreover, the inclusion of legal data in Albertina's (PT) pre-training likely contributed to its strong performance in this *corpus*.

7 Conclusion

We investigated the influence of Portuguese language variants in pre-trained Language Models on Named Entity Recognition (NER) within specialised domains. Our experiments used two distinct *corpora*: a historical corpus in European Portuguese and a legislative corpus in Brazilian Portuguese. We evaluated models pre-trained specifically for Brazilian Portuguese (BERTimbau and Albertina PT-BR), European Portuguese (Albertina PT-PT), and multilingual contexts (XLM-R). Our analysis delved qualitatively into specific class examples, semantics, and compositions.

Several conclusions can be drawn based on our findings. Models tailored to the textual and linguistic context, such as Albertina for European Portuguese, demonstrated superior performance in NER tasks, particularly in the Parish Memories corpus. Albertina excelled in precision and recall, extracting entities from 18th-century texts because of its linguistic alignment. Conversely, larger models such as XLM-R and Albertina showed an enhanced precision-recall balance in the Brazilian legislative texts, underscoring the critical role of model size in handling fine-grained categories.

Future research should focus on granular error analysis and interpretation. Understanding model improvements and challenges is crucial, particularly in ambiguity and linguistic complexity. Exploring the performance of techniques such as semi-supervised learning and transfer learning methods will further advance the field of NER in specific textual domains.

Acknowledgements. This work has received funds from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code

001, the Brazilian funding agency CNPq, and the Portuguese Science Foundation FCT, in the context of the projects CEECIND/01997/2017 and UIDB/00057/2020 https://doi.org/10.54499/UIDB/00057/2020.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)
2. Albuquerque, H.O., Costa, R., Silvestre, G., Souza, E., da Silva, N.F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., et al.: Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In: International Conference on Computational Processing of the Portuguese Language. pp. 3–14. Springer (2022)
3. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: a dataset for named entity recognition in brazilian legal text. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. pp. 313–323. Springer (2018)
4. Bonifacio, L.H., Vilela, P.A., Lobato, G.R., Fernandes, E.R.: A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9. pp. 648–662. Springer (2020)
5. Brito, M., Pinheiro, V., Furtado, V., Neto, J.A.M., Bomfim, F.d.C.J., da Costa, A.C.F., Silveira, R.: Cdjur-br-uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 177–186. SBC (2023)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
7. Correia, F.A., Almeida, A.A., Nunes, J.L., Santos, K.G., Hartmann, I.A., Silva, F.A., Lopes, H.: Fine-grained legal entity annotation: A case study on the brazilian supreme court. Information Processing & Management **59**(1), 102794 (2022)
8. Filho, J.A.W., Wilkens, R., Idiart, M., Villavicencio, A.: The brwac corpus: a new open resource for brazilian portuguese. In: Proceedings of the 11th International conference on language resources and evaluation. pp. 4339–4344 (2018), http://www.lrec-conf.org/proceedings/lrec2018/summaries/599.html
9. Garcia, E.A., Silva, N.F., Siqueira, F., Albuquerque, H.O., Gomes, J.R., Souza, E., Lima, E.A.: Robertalexpt: A legal roberta model pretrained with deduplication for portuguese. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 374–383 (2024)
10. Grilo, S., Bolrinha, M., Silva, J., Vaz, R., Branco, A.: The bdcamoes collection of portuguese literary documents: a research resource for digital humanities and language technology. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 849–854 (2020)
11. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)

12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
13. Nunes, R.O., Balreira, D.G., Spritzer, A.S., Freitas, C.M.D.S.: A named entity recognition approach for portuguese legislative texts using self-learning. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 290–300 (2024)
14. Oleques Nunes., R., Spritzer., A., Dal Sasso Freitas., C., Balreira., D.: Out of sesame street: A study of portuguese legal named entity recognition through in-context learning. In: Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS. pp. 477–489. INSTICC, SciTePress (2024). https://doi.org/10.5220/0012624700003690
15. Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H.L., Osório, T.: Advancing neural encoding of portuguese with transformer albertina pt. In: EPIA Conference on Artificial Intelligence. pp. 441–453. Springer (2023)
16. Santos, D., Cardoso, N.: A golden resource for named entity recognition in portuguese. In: International workshop on computational processing of the portuguese language. pp. 69–79. Springer (2006)
17. Santos, D., Seco, N., Cardoso, N., Vilela, R.: Harem: An advanced ner evaluation contest for portuguese. In: quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006) (2006)
18. Santos, J., Cameron, H.F., Olival, F., Farrica, F., Vieira, R.: Named entity recognition specialised for portuguese 18th-century history research. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 117–126 (2024)
19. Santos, R., Rodrigues, J., Gomes, L., Silva, J., Branco, A., Cardoso, H.L., Osório, T.F., Leite, B.: Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family (2024)
20. Silva, P., Franco, A., Santos, T., Brito, M., Pereira, D.: Cachacaner: a dataset for named entity recognition in texts about the cachaça beverage. Language Resources and Evaluation pp. 1–19 (2023)
21. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS (2020)
22. Souza, F., Nogueira, R.F., de Alencar Lotufo, R.: Portuguese named entity recognition using BERT-CRF. CoRR **abs/1909.10649** (2019), http://arxiv.org/abs/1909.10649
23. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002) (2002), https://aclanthology.org/W02-2024
24. Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., Santos, I.: Enriching the 1758 portuguese parish memories (alentejo) with named entities. Journal of Open Humanities Data **7**, 20 (2021)
25. Zanuz, L., Rigo, S.J.: Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 219–229. Springer (2022)