PortOldBERT: Portuguese Historical Language Models

Tomás Freitas Osório and Henrique Lopes Cardoso
LIACC, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
tomas.s.osorio@gmail.com, hlc@fe.up.pt

Abstract

Historical language models play a crucial role in the study of languages, and can benefit tasks such as named-entity recognition (NER), part-of-speech (PoS) tagging, and post-OCR correction, among others. Despite their relevance, most efforts have been concentrated on English. To the best of our knowledge, no such model exists for historical Portuguese. In this work, we introduce PortOldBERT, the first historical Portuguese encoder language model. We demonstrate its usefulness by comparing PortOldBERT's performance with Albertina, the encoder on which it is based, across multiple tasks—pseudo-perplexity, NER, PoS tagging, word error rate (WER) prediction, and OCR error detection—and for different historical periods. PortOldBERT consistently outperforms Albertina in historical data, demonstrating its ability to effectively integrate historical linguistic contexts while retaining the ability to process contemporary text.

1 Introduction

With the rise of computational and digital humanities resources, the publication of large language models (LLMs) has surged, with new models being introduced on a weekly basis. These models are designed to serve various purposes and increasingly include multilingual capabilities, reflecting the expanding landscape of computational linguistics research. Nevertheless, despite the increasing digitisation efforts by libraries and archives, there is still a notable scarcity of language models focused on historical languages (Gabay et al., 2022).

Pre-trained language models (PLMs) specialised in historical language are invaluable due to the dynamic nature of linguistic structures. Languages evolve over time, exhibiting variations across regions and social groups. Changes in pronunciation, the introduction of new words, shifts in the meanings of existing words, and morphological transformations are all driven by many factors, such as contact between different communities during language learning, social differentiation, or other natural processes of language usage (Zampieri and Becker, 2013; Alatrash et al., 2020; Schieffelin and Ochs, 1986; Fishman, 1964; Kerswill, 2006; Blank, 1999; Hamilton et al., 2016). Another issue with historical texts is that they often have non-consolidated spelling (Manjavacas Arevalo and Fonteyn, 2022) and contain a high degree of orthographic variation—common in Western European languages before the standardisation of spelling norms in the 18th and 19th centuries. Therefore, having PLMs specialised in historical language can benefit many tasks, such as automatic post-OCR correction (Rijhwani et al., 2021) and linguistic annotation (Camps et al., 2021), among others.

Portuguese, a Western Romance language that originated in the Iberian Peninsula, is spoken by approximately 250 million native speakers and 24 million second-language speakers (Eberhard et al., 2023). It holds official status in several countries, including Angola, Brazil, Cabo Verde, Mozambique, Guinea-Bissau, Portugal, and São Tomé and Príncipe (Comunidade dos Países de Língua Portuguesa , CPLP), and is a co-official language in East Timor, Equatorial Guinea, and Macau.

Despite Portuguese being spoken by millions and not being a low-resource language, to the best of our knowledge, there is no available PLM specialised in historical Portuguese. To address this gap, we introduce PortOldBERT, a specialisation of Albertina (Rodrigues et al., 2023) on historical Portuguese. Our training data consists of documents from eighteen corpora spanning the 12th to the 20th centuries. We divided the training data by historical period and developed specialised versions of PortOldBERT for each period, as well as a version encompassing all periods.

To evaluate the resulting models, we compared their historical adaptation capabilities with the original Albertina models using pseudo-perplexity and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2691–2705

the performance on NER, PoS tagging, WER prediction, and OCR error detection.

Beyond introducing PortOldBERT, a key contribution of this work is the evaluation tasks based on OCR–transcription pairs, specifically WER prediction and OCR error detection. These tasks exploit a relatively abundant source of annotated data that has remained under-explored for benchmarking the quality of historical PLMs.

PortOldBERT[1] is publicly available to the research community via HuggingFace under the Creative Commons Attribution Non-Commercial 4.0 license, together with the corpora used for its training[2].

2 Related Work

Large language models (LLMs) contain hundreds of millions to billions of parameters trained on massive corpora (Shanahan, 2024). These models have demonstrated remarkable capabilities across a broad spectrum of natural language processing tasks and fields (Fan et al., 2024). LLMs can be categorised into general-purpose and domain-specific (Zhang et al., 2023). General-purpose LLMs, trained on a wide range of topics and domains, excel in various language-related tasks and have made significant contributions to the AI community (Zhao et al., 2023). However, while they perform well across multiple tasks, general-purpose LLMs may encounter challenges in specific domain-focused tasks. Consequently, there is an ongoing effort to develop domain-specific models, utilising corpora from particular domains. These models are designed to grasp domain-specific knowledge, terminology, and stylistic nuances, enhancing performance in specialised downstream tasks within those domains (Wang et al., 2023; Gururangan et al., 2020).

Looking at historical text, the development of PLMs has primarily focused on a select few languages, with English receiving the most attention. Researchers have pursued two main strategies for creating historical PLMs. The first strategy involves continuing the training of an encoder-based model pre-trained on contemporary data with historical data. For example, Hosseini et al. (2021) adapted a contemporary BERT model to nineteenth-century English, while Singh et al. (2021) applied this approach to ancient and Byzantine Greek.

The second strategy consists of training a model from scratch using historical corpora. Notably, D'AlemBERT (Gabay et al., 2022), a RoBERTa-based (Liu et al., 2019) model, was developed for historical French. Following this approach, MacBERTh (Manjavacas Arevalo and Fonteyn, 2021), a BERT-based model, was trained for English. Riemenschneider and Frank (2023) introduced four models, GRϵBERTA and PHILBERTA (RoBERTa-based), and GRϵTA and PHILTA (T5-based (Raffel et al., 2020)), focusing on monolingual Ancient Greek, with the former two being trained on Greek, Latin, and English corpora. Riemenschneider and Frank found that monolingual models typically outperform multilingual ones in morphological and syntactic tasks. Bamman and Burns (2020) also trained a BERT-based model from scratch for Latin on 642.7 million tokens originally written over 22 centuries, from 200 BCE to today. Similarly, GysBERT (Manjavacas Arevalo and Fonteyn, 2022) is a BERT-base model for historical Dutch between 1500 and 1950 trained from scratch. As we previously saw, some models are multilingual, such as hmBERT (Schweter et al., 2022), supporting historical German, English, French, Finnish, and Swedish.

Comparative research between both strategies was also done. For instance, Manjavacas and Fonteyn (2022) demonstrate that models trained from scratch generally outperform those using the first strategy across various NLP tasks for historical English. On GHisBERT (Beck and Köllner, 2023) work, this was also shown in the eighth-century till the seventh-century German. Palmero Aprosio et al. (2022) also explore both methods, where the BERToldo was trained from scratch, and ContBERToldo is the outcome of continuing the pre-training of a contemporary PLM on historical corpora. However, in this work, ContBERToldo shows superior performance in their evaluations. Similarly, Konle and Jannidis (2020) explored the effects of continuing the pre-training, on historical text, of a BERT-based model that has been pre-trained on contemporary German versus training from scratch an ELECTRA (Clark et al., 2020) model and have found that applying the first strategy improves performance significantly compared to training from scratch.

Palmero Aprosio et al. (2022) also extended their study of BERToldo and ContBERToldo by analysing the impact of restricting the training cor-

[1] https://huggingface.co/LIACC/PortOldBERT

[2] https://huggingface.co/LIACC/HistoricalPortugueseCorpora

pus to specific historical periods. Their findings show that the most effective strategy is pre-training a PLM on the full available historical data. This suggests that, rather than limiting training to individual periods, leveraging the entire dataset yields stronger downstream performance.

To the best of our knowledge, no PLM has been released for historical Portuguese at the time of writing.

3 Language models for historical Portuguese

In this work, we present PortOldBERT, an encoder model derived from Albertina (Rodrigues et al., 2023), a DeBERTa-based model (He et al., 2021) pre-trained on contemporary Portuguese and released under the MIT license. Rather than training from scratch, we opted for continuing pre-training, as the former is often less effective when constrained by limited corpora (see Section 2).

PortOldBERT features several specialised versions, each tailored to capture the unique linguistic characteristics of different historical periods. In addition to these period-specific models, we trained a version that integrates data from all historical periods. The following sections explain the methodology taken.

3.1 Corpora

Osório and Lopes Cardoso (2024) identified twenty-two historical Portuguese corpora, twenty of which are freely available. From these, we chose eighteen corpora, which are detailed in Table 10 in Appendix A. The remaining two, although freely available, were excluded, one due to overlap with existing corpora and the other because the exact document dates were unavailable.

Pichel Campos et al. (2018) state that the Portuguese language can be divided into seven historical periods: the Medieval Period (12th-15th centuries), the Renaissance Period (16th-17th centuries), the 18th century, the first and second half of the 19th century, and the first and second half of the 20th century. However, due to the scarcity of documents from specific periods, we consolidate these into the Medieval Period (before the 16th century), the Renaissance Period combined with the 18th century (16th to 18th centuries), and the first and second halves of both the 19th and 20th centuries. Table 1 presents the number of documents and tokens used to train the different versions of PortOldBERT for historical Portuguese. Additionally, it provides the subword fertility rate (SFR) for each historical period. The SFR, defined as the average number of subwords generated per word (Rust et al., 2021), which measures the "aggressiveness" of the tokenizer's segmentation. Notably, the tokenizer produced zero out-of-vocabulary (OOV) tokens across all historical periods.

Table 1: Number of tokens and documents used to pre-train PortOldBERT per split and subword fertility rate (SFR) per period.

Period	Train		Test		SFR
	Docs	Tokens	Docs	Tokens	
<16th	246	5.0M	17	101k	2.15
16th–18th	2063	9.2M	144	650k	2.08
19th-1	1409	7.6M	112	603k	2.01
19th-2	566	14.4M	50	1.8M	2.01
20th-1	5729	11.3M	441	718k	2.00
20th-2	379	807k	22	55k	1.96
Total	10392	48.5M	786	3.9M	2.03

To evaluate PortOldBERT on NER, we use the BDCamões (Grilo et al., 2020), ELTeC (Santos, 2021), and PPM (Vieira et al., 2021) corpora. The distribution of entities across the training and test sets is provided in Tables 11, 12 and 13 in Appendix A. The ELTeC and BDCamões corpora were automatically tagged using PALAVRAS (Bick, 2006, 2014) and LX-Suite (Branco and Silva, 2006), respectively, both tools primarily designed for contemporary Portuguese. The PPM corpus includes a single manually annotated document, which we use for testing, while the remaining documents were automatically annotated.

BDCamões contains 195 automatically annotated literary documents across various genres, including novels, chronicles, poems, and short stories from the 16th to the 21st centuries. ELTeC comprises 100 original novels written between 1840 and 1920, each featuring a comparable internal structure in terms of nature, scope, and quality. PPM consists of responses to a 60-question survey sent to parish priests overseeing dioceses across the Portuguese Kingdom collected between 1758 and 1761.

Despite BDCamões containing automatically annotated documents spanning from the 16th to the 20th century, the limited number of texts available from the pre-19th century period led us to group BDCamões into three periods: pre-19th century, 19th century, and 20th century. Table 2 presents the number of tokens and documents used for training

Table 2: Number of tokens and documents per split for the BDCamões NER task.

Period	Train		Test	
	Docs	Tokens	Docs	Tokens
<19th	10	102k	2	16k
19th	66	2.4M	6	237k
20th	95	1.3M	16	102k
Total	171	3.8M	24	355k

Table 3: Number of tokens and documents per split for the ELTeC and PPM NER task.

Corpus	Train		Test	
	Docs	Tokens	Docs	Tokens
ELTeC	91	7.4M	9	568k
PPM	366	720.2k	1	27.4k

and testing in each period.

For the ELTeC and PPM corpora, we opt not to divide them into separate periods, as all documents within each corpus originate from a similar historical period. Table 3 shows the distribution of documents and tokens between the training and testing sets.

To evaluate PortOldBERT on PoS tagging, we use four corpora: BDCamões, ELTeC, Tycho Brahe (Galves, 2018), and Colonia (Zampieri and Becker, 2013), all of which were automatically annotated. The Tycho Brahe corpus comprises 47 annotated texts ranging from the 14th to the 19th century, covering various literary forms such as letters, journals, acts, books, and dissertations. Colonia contains texts spanning the 16th to the early 20th century.

The Tycho Brahe and Colonia corpora are split into two periods: pre-19th century and 19th century. Tycho Brahe encompasses 34 PoS tags, while Colonia has 33. Table 4 shows the token and document distribution for Tycho Brahe and Colonia in the PoS tagging task. The ELTeC and BDCamões corpora follow the same training and testing splits used in the NER task, with ELTeC containing 74 PoS tags and BDCamões 78.

For OCR error detection and WER prediction, we use the PORTO (Freitas Osório and Lopes Cardoso, 2025) dataset, which consists of documents written in Portuguese spanning from the 17th to the 20th centuries. It comprises 3,782 image–transcription pairs, accompanied by OCR outputs from four different systems. From these, we selected 3,677 pairs that were not flagged as having complex layouts, in order to ensure higher transcription quality, and written in Portuguese. Table 5 presents the token and document distribution for both tasks, considering only the transcription tokens.

Table 4: Number of tokens and documents per split for the Tycho and Colonia POS tagging task.

Corpus	Period	Train		Test	
		Docs	Tokens	Docs	Tokens
Tycho	<19th	28	1.3M	7	308k
	19th	7	415k	5	198k
	Total	35	1.7M	12	506k
Colonia	<19th	38	1.5M	5	169k
	19th	45	3.3M	5	451k
	Total	83	4.8M	10	620k

Table 5: Number of tokens and documents on PORTO dataset per split.

Corpus	Period	Train		Test	
		Docs	Tok.	Docs	Tok.
ESTER-Pt	19th-20th	1876	380k	149	29k
CHLMP	18th	623	182k	47	14k
Tycho	17th&20th	473	67k	34	5k
AdA	18th	448	165k	27	10k
Total		3420	794k	257	57k

To prevent leakages and ensure data integrity, we always separate full documents when generating the test sets. This approach maintains distinct isolation between training and testing data, ensuring accurate and more reliable model evaluations by avoiding potential biases.

3.2 Evaluation

We employ two approaches to evaluate the language models: we calculate each model's pseudo-perplexity on text samples from different time frames and also rely on fine-tuning models for NER, PoS tagging, OCR error detection, and WER prediction.

Perplexity is a widely used metric for evaluating language models. Perplexity measures how well a language model fits a text sample; a low perplexity indicates effective prediction capabilities, while a high perplexity suggests poor performance.

Perplexity is defined as the exponential average negative log-likelihood of a sequence. If we have a tokenised sequence $X = (x_0, x_1, ..., x_t)$, then the perplexity of X is:

$$PPL(X) = exp\left\{ -\frac{1}{t}\sum_{i}^{t} logP_{LM}(x_i|x_{<i}) \right\} \quad (1)$$

where $logP_{LM}(x_i|x_{<i})$ is the log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$. Therefore, perplexity can be seen as a measure of a model's ability to predict uniformly across a set of tokens in a corpus. However, a masked language model, such as BERT, makes predictions based on surrounding tokens, not just the preceding ones. As a result, PPL is unsuitable for evaluation, so, instead, pseudo-perplexity (Salazar et al., 2020) (PPPL) can be employed, which is computed as follows:

$$PPPL(X) = exp\left\{ -\frac{1}{t}\sum_{i}^{t} logP_{MLM}(x_i|x_n) \right\} \quad (2)$$

where the token $x_{n=i}$ is replaced with [MASK] and predicted using all past and future tokens.

3.3 Model Training

Albertina (Rodrigues et al., 2023) is available in various sizes, ranging from a smaller version with 100 million parameters to a high-capacity version with 1.5 billion parameters. Each model has two versions: one specifically tailored to European Portuguese and the other to Brazilian Portuguese. This linguistic distinction is particularly important when dealing with historical Portuguese text, where orthographic agreements and linguistic evolution play a significant role. PortOldBERT uses the European Portuguese 100 million parameters version.

We train several versions of PortOldBERT using a 128-token sequence truncation and dynamic padding, consistent with the Albertina pretraining setup. The training was conducted using with a learning rate of 1e-5, employing a weight decay of 0.05 and 10% warm-up steps. Each model variant underwent 10 epochs of training with a batch size of 140. These experiments were conducted on two Quadro RTX 8000 GPUs, each with 48 GB of VRAM.

For fine-tuning Albertina and PortOldBERT models for NER, PoS tagging, OCR error detection and WER prediction, we employ Low-Ranked Adaptation (Hu et al., 2021) (LoRA) with a rank of 8, an alpha value of 32, a dropout rate of 0.05, a learning rate of 2×10^{-5}, and a weight decay of 0.05. LoRA experiments were conducted on a single Quadro RTX 8000 GPU. LoRA is a method that reduces the number of training parameters and, consequently, the storage demands of attention-based models. In this process, the weights of the pre-trained model are kept frozen, and additional trainable weight matrices are introduced to the attention layers to tailor the model for specific downstream tasks (Houlsby et al., 2019; Mahabadi et al., 2021; He et al., 2022). This approach was chosen to maintain high performance across multiple tasks while minimising the computational overhead typically associated with fine-tuning transformer architectures.

During fine-tuning, all NER and PoS tagging models use the entire training dataset available for each corpus and are evaluated independently for each period, when applicable. We adopt this approach for simplicity and due to the limited number of examples in certain corpora. Additionally, we do not apply any balancing strategies based on historical periods or entity/PoS distribution.

For OCR error detection and WER prediction, we trained another PortOldBERT using the same setup, but with the PORTO overlapping documents excluded from the training set to prevent data leakage. In both tasks, punctuation and letter casing errors were ignored, as they have minimal impact on semantic correctness.

WER prediction is framed as a regression task in which the model estimates the WER between an OCR output and its corresponding transcription. This task is particularly valuable because it enables automatic assessment of transcription quality without requiring ground-truth references. Such estimations facilitate quality control and lower validation costs.

OCR error detection, on the other hand, is a word-level classification task, framed similarly to a NER task, aimed at identifying words affected by insertion, substitution, or deletion errors. This capability allows for the automatic localisation of transcription mistakes, significantly reducing manual correction efforts and improving the overall efficiency of OCR post-processing workflows.

During the training of the PortOldBERT models and its fine-tuning for downstream tasks, the validation set is created by randomly splitting the training data with a fixed seed, where 7% of the data is allocated for validation, while the remaining is used for training.

4 Results & Discussion

We present the results of PortOldBERT and compare them with Albertina across different time frames and tasks to evaluate their versatility and robustness. By analysing how the model adapts to different temporal domains, we aim to illustrate the impact of temporal domain-specific pre-training.

4.1 Pseudo-Perplexity

Figure 1 presents the pseudo-perplexity results for each model in each given period, with detailed results presented in Table 14 in Appendix A. The 100-million-parameter Albertina models were pre-trained on contemporary corpora, with the pt-PT version accessing only European Portuguese and the pt-BR version focusing on Brazilian Portuguese. PortOldBERT refers to the Albertina European Portuguese model that was further pre-trained with historical data from a specific period or, if noted as "All", on all available historical corpora.

As shown in Figure 1, PortOldBERT-All consistently achieves the lowest pseudo-perplexity across all temporal periods, indicating a better generalisation over time.

In the <16th period, only the PortOldBERT models trained specifically for this period achieve a lower pseudo-perplexity score: PortOldBERT <16th and PortOldBERT-All. This outcome could be attributed to the significant linguistic differences between this period and later periods, particularly because no standard orthographic conventions existed at the time.

The Albertina versions exhibit higher pseudo-perplexity scores the farther the test sets are from contemporary times, except on the <16th, which has a much lower pseudo-perplexity when compared with the following period. The performance of Albertina models significantly deteriorates with documents older than the second half of the 19th century.

Considering all these results, it appears that a single historical model suffices, PortOldBERT-All. This is likely due to insufficient data available for each period, preventing any model from properly acquiring temporally relevant linguistic knowledge from a given period alone. Interestingly, despite all model parameters being trained, the PortOldBERT models still retain some of the contemporary knowledge from the Albertina models, as shown by their low pseudo-perplexity scores in documents from the second half of the 20th century. This suggests that continuing pre-training effectively integrates historical context while preserving the model's initial knowledge.

On the remaining downstream tasks, we chose to use only PortOldBERT-All, as it performed notably better than the other models in terms of pseudo-perplexity score. For comparison, we also have fine-tuned the Albertina European-Portuguese and Brazilian-Portuguese versions.

4.2 Named Entity Recognition

Table 6 summarises the NER performance on the ELTeC, BDCamões and PPM corpus, with detailed results presented in Table 15, 16 and 17 in Appendix A.

When comparing PortOldBERT-All's NER performance with both Albertina variants, the gains are not as pronounced as one might expect based on the pseudo-perplexity results. One contributing factor could be the way named entities were annotated; in particular, the ELTeC and BDCamões corpora were automatically tagged. As a result, these automatically generated annotations may introduce inconsistencies or errors when applied to historical texts, potentially limiting model performance.

In contrast, the PPM corpus, where test annotations were manually curated, shows a significant performance improvement for PortOldBERT-All compared to the Albertina variants, where PortOldBERT-All achieves an F1 score of 47.32%, surpassing Albertina pt-PT by 5.06% and Albertina pt-BR by 7.02%.

Nevertheless, PortOldBERT-All consistently achieves the highest recall and F1 score across the NER experiments, which may indicate its superior ability to generalise better, capturing more entities than the Albertina versions. The only exception occurs in the <19th period of BDCamões, where it does not achieve the top F1 score. However, given the limited number of entities and documents in this test set, the reliability of this result is reduced, though PortOldBERT–All still maintains the highest recall in that period.

4.3 Part-of-Speech Tagging

PoS tagging results are reported in Table 7. Similarly to the named entity annotations, the datasets used for PoS tagging rely on automatically generated annotations.

On the experiments, PortOldBERT-All consistently achieves the highest recall and accuracy in

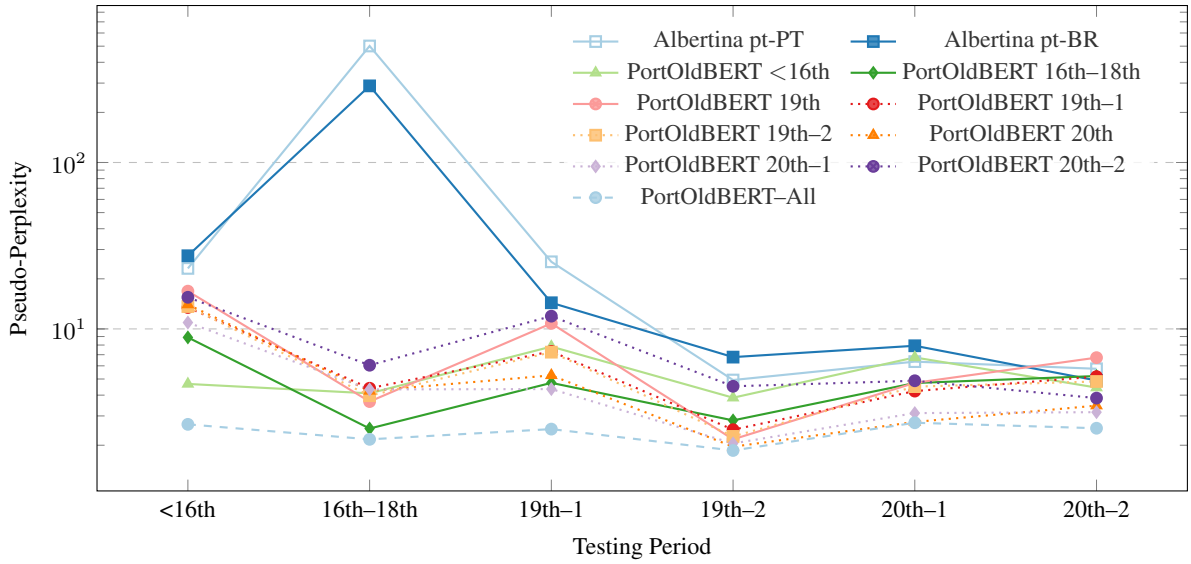

Figure 1: Pseudo-perplexity across temporal testing periods (lower is better). Training temporal periods are indicated in PortOldBERT's version name.

Table 6: Results from models fine-tuned on ELTeC, BDCamões and PPM NER datasets: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Period	Corpus	Albertina pt-PT			Albertina pt-BR			PortOldBERT-All		
		P	R	F1	P	R	F1	P	R	F1
19th-20th	ELTEC	74.56	76.79	75.66	73.87	75.14	74.50	**74.94**	**77.40**	**76.15**
<19th	BDCamões	38.75	47.14	42.54	**39.72**	48.10	**43.51**	38.65	**48.86**	43.16
19th		52.25	56.72	54.40	48.24	54.12	51.02	**53.21**	**57.86**	**55.44**
20th		50.76	54.04	52.36	45.96	51.38	48.52	**51.54**	**55.45**	**53.42**
18th	PPM	42.12	42.39	42,26	44.64	43,97	40.30	**48.04**	**46.62**	**47.32**

nearly all experiments, except for the <19th period of BDCamões. A similar pattern observed in the NER experiments, which may be explained by the limited number of documents available for evaluation in this period—a consequence of our document-level data splitting strategy designed to minimise potential biases. Overall, these results indicate that PortOldBERT-All may have generalised more effectively than the Albertina variants. A possible explanation for this is that PortOldBERT-All was pre-trained on a broader and more representative dataset, allowing it to capture a wider range of linguistic patterns.

4.4 WER Prediction

For the WER prediction and OCR error detection tasks, as previously described, we trained a variant of PortOldBERT that excludes the PORTO overlapping documents from the training set. This variant, referred to as PortOldBERT-All-EP, follows the same experimental configuration as PortOldBERT-All, differing only in the exclusion of these overlapping documents.

Table 8 reports the performance of the models fine-tuned for WER prediction; detailed results for each dataset subset are provided in Table 18 in Appendix A. Across all evaluation metrics, PortOldBERT-All-EP consistently outperforms both Albertina variants, demonstrating clear gains in predictive quality.

Relative to Albertina pt-BR, PortOldBERT-All-EP improves R^2 by 24.6% and correlation by 10.4%, while also achieving lower MAE and RMSE. The improvements are even more pronounced when compared with Albertina pt-PT, further reinforcing PortOldBERT-All-EP robustness and effectiveness on this task.

4.5 OCR error Detection

Table 9 presents the results for OCR error detection. PortOldBERT-All-EP surpasses both Albertina variants, achieving the best performance across all metrics—precision, recall, and F1 score. Results broken down by dataset subset are reported

Table 7: Results from models fine-tuned on ELTeC, Colonia, Tycho and BDCamões (BDC) PoS datasets: A (Accuracy), P (Precision) and R (Recall).

Period	Corpus	Albertina pt-PT			Albertina pt-BR			PortOldBERT-All		
		A	P	R	A	P	R	A	P	R
19th-20th	ELTeC	93.76	**44.26**	42.05	93.98	43.29	40.55	**94.50**	43.52	**44.42**
<19th	Colonia	90.92	68.11	67.14	90.68	67.23	66.88	**91.48**	**68.24**	**68.48**
20th		95.91	73.56	68.48	95.80	**75.10**	68.38	**96.23**	71.52	**69.01**
<19th	Tycho	96.25	87.30	84.51	95.91	87.34	83.07	**97.17**	**88.62**	**86.64**
19th		93.20	**86.07**	83.28	93.07	86.02	81.40	**93.52**	85.91	**83.37**
<19th	BDC	91.62	68.69	67.59	**91.70**	**70.41**	68.24	91.38	68.50	**69.05**
19th		95.48	55.30	52.28	95.35	**55.61**	**52.10**	**95.51**	54.24	51.23
20th		94.97	60.01	56.13	94.88	58.73	54.33	**95.05**	**61.12**	**56.46**

Table 8: Results of models fine-tuned for WER prediction: R^2 (coefficient of determination), MAE (mean absolute error), RMSE (root mean squared error), and Corr (correlation).

Model	R^2	MAE	RMSE	Corr
Albertina pt-PT	0.597	0.104	0.163	0.792
Albertina pt-BR	0.639	0.105	0.155	0.815
PortOldBERT-All–EP	**0.796**	**0.075**	**0.116**	**0.900**

Table 9: Results from models fine-tuned on OCR error detection: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Model	P	R	F1
Albertina pt-PT	0.903	0.831	0.865
Albertina pt-BR	0.901	0.824	0.861
PortOldBERT-All-EP	**0.920**	**0.859**	**0.888**

in Table 19 in Appendix A. This improvement suggests that PortOldBERT-All-EP captures orthographic and lexical variation more effectively, which is crucial for detecting OCR-related noise in historical texts.

5 Conclusions

Our study highlights the advantages of continuing to train the models on a broad temporal domain. The PortOldBERT-All model version achieves superior performance concerning pseudo-perplexity evaluation. This broad temporal pre-training enables the model to effectively integrate historical linguistic contexts while retaining some of the contemporary knowledge. However, in downstream tasks such as NER and PoS tagging, the performance gains are less pronounced. This could be due to the limited evaluation data, as we opted to use full documents rather than partial excerpts to reduce bias. Another contributing factor might be the reliance on automatically generated annotations, affecting result reliability. In contrast, for the PPM corpus, where the test set was manually annotated, the advantage of PortOldBERT-All is more evident, outperforming fine-tuned Albertina variants.

When looking into WER prediction and OCR error detection, PortOldBERT-All-EP outperforms the Albertina models, particularly in the WER prediction task. These results suggest that continuing pre-training a PLM on historical corpora provides a stronger representational capacity for error-sensitive tasks, making it better suited to handle the noisy and heterogeneous OCR outputs of historical documents.

For future work, a clear path would be to train an encoder from scratch rather than continuing to train a model pre-trained on contemporary text, as suggested by previous studies (Manjavacas and Fonteyn, 2022; Beck and Köllner, 2023). Another direction would be to train a decoder-based language model specifically for historical Portuguese, adopting a similar approach to this work but using perplexity instead of pseudo-perplexity as the evaluation metric.

Limitations

Despite the promising results, our study has certain limitations. The quality and quantity of the textual data used for pre-training and evaluation could introduce biases or inconsistencies. For example, historical texts may contain transcription errors, which could impact the model's ability to generalise effectively. Furthermore, the model's performance was assessed using a limited set of tasks that rely mainly on automatically annotated data. This may not fully capture its overall capabilities, as automatic annotations can introduce inaccuracies in training and evaluation.

Acknowledgments

This work was financially supported by UID/00027/2025 of the Artificial Intelligence and Computer Science Laboratory (LIACC), with DOI https://doi.org/10.54499/UID/00027/2025, funded by Fundação para a Ciência e a Tecnologia, I.P. / MECI through national funds.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Arquivo dos Açores. 2023. https://hdl.handle.net/21.11129/0000-000D-F8C0-2. Accessed: 16-5-2023.

Arquivo Pessoa. 2023. http://arquivopessoa.net/. Accessed: 15-05-2023.

As Memórias Paroquiais de 1758. 2023. http://www.cidehusdigital.uevora.pt/portugal1758. Accessed: 15-05-2023.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *ArXiv*, abs/2009.10053.

Christin Beck and Marisa Köllner. 2023. GHisBERT – training BERT from scratch for lexical semantic investigations across historical German language stages. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 33–45, Singapore. Association for Computational Linguistics.

Eckhard Bick. 2006. Functional aspects in portuguese ner. In *Computational Processing of the Portuguese Language*, pages 80–89, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eckhard Bick. 2014. *PALAVRAS - A Constraint Grammar-Based Parsing System for Portuguese*, pages 279–302. Bloomsbury Academic.

Andreas Blank. 1999. *Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change*, pages 61–90. De Gruyter Mouton, Berlin, Boston.

António Branco and João Ricardo Silva. 2006. A suite of shallow processing tools for Portuguese: LX-suite. In *Demonstrations*, pages 179–182.

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. 2021. Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. *Journal of Data Mining & Digital Humanities*, 2021.

Chancelaria de D. Afonso III: documentos em português. 2023. https://hdl.handle.net/21.11129/0000-000D-FE7C-B. Accessed: 16-5-2023.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Comunidade dos Países de Língua Portuguesa (CPLP). 2023. https://www.cplp.org/id-2597.aspx. Accessed: 2022-10-26.

Corpus Eletrônico de Documentos Históricos do Sertão. 2023. http://www5.uefs.br/cedohs/view/home.html. Accessed: 16-5-2023.

Corpus Histórico da Linguagem da Medicina em Português (Século XVIII): Terminologia Diacrônica e Humanidades Digitais. 2023. https://sites.google.com/view/projeto38597. Accessed: 16-5-2023.

CTACorpus. 2023. http://teitok.clul.ul.pt/cta/. Accessed: 13-12-2022.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas.

Margarida Falcão, Mariana Dias, and Carla Teixeira Lopes. 2022. Manual transcriptions of typewritten digital representations of portuguese cultural heritage documents from the 20th century. https://rdm.inesctec.pt/dataset/cs-2022-005.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A bibliometric review of large language models research from 2017 to 2023. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Joshua A. Fishman. 1964. Language maintenance and language shift as a field of inquiry. a definition of the field and suggestions for its further development. *Linguistics*, 2(9):32–70.

Tomás Freitas Osório and Henrique Lopes Cardoso. 2025. Portuguese post-ocr resources for text optimisation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6361–6366, New York, NY, USA. Association for Computing Machinery.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Charlotte Galves. 2018. The tycho brahe corpus of historical portuguese: Methodology and results. *Linguistic Variation*, 18:49–73.

GMHP. 2023. http://www.usp.br/gmhp/CorpI.html. Accessed: 14-06-2022.

Mariana Gomes, Ana Guilherme, Leonor Tavares, and Rita Marquilhas. 2012. Project FLY: a multidisciplinary project within linguistics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2833–2837, Istanbul, Turkey. European Language Resources Association (ELRA).

Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The bdcamões collection of portuguese literary documents: a research resource for language technology and digital humanities. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 849–854.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. A large portuguese corpus on-line: Cleaning and preprocessing. *Computational Processing of the Portuguese Language. PROPOR*, pages 113–120.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english. *Journal of Open Humanities Data*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Paul Kerswill. 2006. *Migration and Language*, volume Volume 3. De Gruyter Mouton, Berlin • New York.

Leonard Konle and Fotis Jannidis. 2020. Domain and task adaptive pretraining for language models. In *Workshop on Computational Humanities Research*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Tomás Freitas Osório and Henrique Lopes Cardoso. 2024. Historical Portuguese corpora: a survey. *Language Resources and Evaluation*.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. BERToldo, the historical BERT for Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 68–72, Marseille, France. European Language Resources Association.

Jose Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Project Gutenberg. 2023. https://www.gutenberg.org/browse/languages/pt. Accessed: 15-05-2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, page 441–453. Springer Nature Switzerland.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Diana Santos. 2021. Portuguese novel collection (eltecpor). *European Literary Text Collection (ELTeC)*.

B B Schieffelin and E Ochs. 1986. Language socialization. *Annual Review of Anthropology*, 15(1):163–191.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *CLEF 2022: Conference and Labs of the Evaluation Forum*, pages 1109–1129, Bologna, Italy.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Gael Vaamonde, Ana Luísa Costa, Rita Marquilhas, Clara Pinto, and Fernanda Pratas. 2014. Post scriptum: Archivo digital de escritura cotidiana. *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, pages 473–482.

Renata Vieira, Fernanda Olival, Helena Freire Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*.

Zhonghao Wang, Zijia Lu, Bo Jin, and Haiying Deng. 2023. Mediagpt : A large language model for chinese media. *Preprint*, arXiv:2307.10930.

Maria Francisca Xavier. 2016. *O CIPM – Corpus Informatizado do Português Medieval, fonte de um Dicionário exaustivo*, pages 137–156. De Gruyter, Berlin, Boston.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *Non-Standard Data Sources in Corpus-based Research. ZSM-Studien Series - Vol. 5.*

Shitou Zhang, Jingrui Hou, Siyuan Peng, Zuchao Li, Qibiao Hu, and Ping Wang. 2023. Arcgpt: A large language model tailored for real-world archival applications. *Preprint*, arXiv:2307.14852.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Appendix

Table 10: Corpora used to pre-train PortOldBERT.

Dataset	N. Tokens	Time scope	License
Gutenberg (Project Gutenberg)	19.3M	14th-20th	Open
GMHP (GMHP)	7.03M	15th-20th	CC BY-NC-ND
ELTeC (Santos, 2021)	6.55M	1840-1920	CC BY 4.0
Colonia (Zampieri and Becker, 2013)	4.58M	16th-20th	CC BY-NC-ND
BDCamões (Grilo et al., 2020)	3.68M	15th-21th	MS-NC-NoReD-ND
Tycho Brahe (Galves, 2018)	3.37M	14th-19th	CC BY-NC-ND
CIPM (Xavier, 2016)	3.32M	13th-16th	CC BY-NC-ND
Pessoa (Arquivo Pessoa)	1.20M	20th	No License
LT Corpus (Généreux et al., 2012)	1.48M	19th-20th	ELRA end-user NC
PS Corpus (Vaamonde et al., 2014)	763k	16th-19th	tv
PPM (As Memórias Paroquiais de 1758)	272k	18th	CC BY
CEDOHS (Corpus Eletrônico de Documentos Históricos do Sertão)	192k	16th-20th	No License
CHLMP (Corpus Histórico da Linguagem da Medicina em Português , Século XVIII)	192k	18th	No License
AdA (Arquivo dos Açores)	182k	19th-21th	MS-NC-NoReD-ND
Fly Corpus (Gomes et al., 2012)	131k	20th	CC BY-NC-ND 3.0
EPISA (Falcão et al., 2022)	68.0k	20th	Open Def. 2.1
CTA (CTACorpus)	29.4k	13th-16th	CC BY-NC 4.0
CDAIII (Chancelaria de D. Afonso III: documentos em português)	18k	13th	CC BY-NC-SA

Table 11: Number entity-labeled words per split in the BDCamões NER task.

Entity	<19th		19th		20th		Total	
	Train	Test	Train	Test	Train	Test	Train	Test
Person	1.8k	314	40.2k	3.8k	15.3k	1.1k	57.3k	5.2k
Location	1.0k	111	32.9k	2.9k	13.3k	709	46.3k	3.6k
Organisation	218	44	7.8k	921	2.9k	149	10.9k	1.1K
Miscellaneous	294	37	3.6k	336	1.3k	94	5.2k	467
Work	180	18	2.7k	202	1.2k	63	4.1k	283
Event	4	2	611	253	575	22	1.2k	277

Table 12: Number entity-labeled words per split in the ELTeC NER task.

Entity	Train	Test
Person	132.8k	9k
Role	40.7k	2.8k
Location	29.9k	1.8k
Miscellaneous	11.1k	689
Organisation	10.4k	465
Work	4.0k	145
Demo	2.8k	186
Event	810	44

Table 13: Number entity-labeled words per split in the PPM NER task.

Entity	Train	Test
Location	16.4k	349
Person	8.4k	336
Organisation	1.2k	261

Table 14: Pseudo-Perplexity results. Training temporal periods are indicated in PortOldBERT's version name. Columns indicate the testing temporal periods.

Models	<16th	16th-18th	19th-1	19th-2	20th-1	20th-2
Albertina pt-PT	23.10	501.36	25.31	4.93	6.36	5.76
Albertina pt-BR	27.52	288.72	14.36	6.77	7.92	4.90
PortOldBERT <16th	4.67	4.10	7.83	3.86	6.74	4.44
PortOldBERT 16th-18th	8.89	2.52	4.74	2.82	4.73	5.21
PortOldBERT 19th	16.86	3.66	10.79	2.17	4.74	6.72
PortOldBERT 19th-1	13.45	4.40	7.33	2.48	4.23	5.17
PortOldBERT 19th-2	13.63	3.97	7.25	2.26	4.50	4.83
PortOldBERT 20th	14.01	4.24	5.26	1.96	2.76	3.46
PortOldBERT 20th-1	10.92	4.33	4.35	2.05	3.12	3.16
PortOldBERT 20th-2	15.49	6.05	11.93	4.52	4.88	3.84
PortOldBERT-All	**2.67**	**2.17**	**2.50**	**1.86**	**2.73**	**2.53**

Table 15: Detailed results from models fine-tuned on ELTeC NER dataset: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Entity	Albertina pt-PT			Albertina pt-BR			PortOldBERT-All		
	P	R	F1	P	R	F1	P	R	F1
Person	83.87	**86.51**	**85.17**	83.01	85.63	84.30	**83.98**	85.26	84.62
Role	73.81	76.47	75.12	73.81	72.63	73.21	**77.56**	**82.44**	**79.93**
Location	62.14	71.05	66.30	60.21	67.72	63.75	**62.85**	**71.50**	**66.90**
Misc.	35.83	22.21	27.42	**38.41**	24.53	29.94	37.06	**28.88**	**32.46**
Org.	**21.53**	**30.32**	**25.18**	19.38	27.10	22.60	17.19	26.88	20.97
Work	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Demo	50.45	**60.22**	54.90	50.92	59.68	54.95	**62.15**	59.14	**60.61**
Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall	74.56	76.79	75.66	73.87	75.14	74.50	**74.94**	**77.40**	**76.15**

Table 16: Detailed results from models fine-tuned on BDCamões NER dataset: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Period	Entity	Albertina pt-PT			Albertina pt-BR			PortOldBERT-All		
		P	R	F1	P	R	F1	P	R	F1
<19th	Person	36.02	48.41	41.30	**37.76**	51.59	**43.61**	35.82	**51.91**	42.39
	Location	37.84	50.45	43.24	39.29	49.55	43.82	**40.85**	**52.25**	**45.85**
	Org.	**15.00**	**6.82**	**9.37**	5.88	2.27	3.28	6.25	2.27	3.33
	Misc.	**80.43**	**100.00**	**89.16**	74.47	94.59	83.33	79.55	94.59	86.42
	Work	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Overall	38.75	47.14	42.54	**39.72**	48.10	**43.51**	38.65	**48.86**	43.16
19th	Person	61.68	60.26	60.96	59.24	57.31	58.26	**61.93**	**60.60**	**61.26**
	Location	**57.32**	67.76	62.10	54.60	67.69	60.44	56.72	**71.37**	**63.21**
	Org.	**19.92**	**25.62**	**22.41**	12.01	18.13	14.45	18.81	22.37	20.44
	Misc.	52.58	81.84	64.03	59.06	70.83	64.41	**60.30**	**83.63**	**70.07**
	Work	0.36	0.50	0.41	0.54	0.99	0.70	**1.40**	**1.49**	**1.44**
	Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Overall	52.25	56.72	54.40	48.24	54.12	51.02	**53.21**	**57.86**	**55.44**
20th	Person	**56.86**	59.44	**58.12**	49.47	55.26	52.21	55.09	**59.89**	57.39
	Location	49.62	55.57	52.43	47.24	54.30	50.52	**51.04**	**58.82**	**54.65**
	Org.	19.19	25.50	21.90	16.27	22.82	18.99	**22.67**	**26.17**	**24.30**
	Misc.	63.89	73.40	68.32	68.27	**75.53**	71.71	**77.53**	73.40	**75.41**
	Work	0.00	0.00	0.00	0.00	0.00	0.00	**0.04**	**1.59**	**2.27**
	Event	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Overall	50.76	54.04	52.36	45.96	51.38	48.52	**51.54**	**55.45**	**53.42**

Table 17: Detailed results from models fine-tuned on PPM NER dataset: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Entity	Albertina pt-PT			Albertina pt-BR			PortOldBERT-All		
	P	R	F1	P	R	F1	P	R	F1
Location	38.15	67.34	48.70	39.40	67.62	49.79	**43.61**	**68.48**	**53.29**
Person	51.88	49.40	50.61	**58.25**	53.57	55.81	57.22	**60.12**	**58.64**
Org.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall	42.12	42.39	42,26	44.64	43,97	40.30	**48.04**	**46.62**	**47.32**

Table 18: Detailed results of models fine-tuned for WER prediction: R^2 (coefficient of determination), MAE (mean absolute error), RMSE (root mean squared error), and Corr (correlation).

Subset	Albertina pt-PT				Albertina pt-BR				PortOldBERT-All-EP			
	R^2	MAE	RMSE	Corr	R^2	MAE	RMSE	Corr	R^2	MAE	RMSE	Corr
AdA	0.575	0.119	0.160	0.764	0.576	0.124	0.160	0.777	**0.742**	**0.091**	**0.124**	**0.888**
CHLMP	0.214	0.024	0.051	0.545	0.008	0.039	0.057	0.569	**0.359**	**0.034**	**0.046**	**0.793**
ESTER-Pt	0.471	0.145	0.202	0.719	0.541	0.136	0.188	0.757	**0.753**	**0.093**	**0.138**	**0.874**
Tycho	0.848	0.086	0.155	0.848	0.870	0.090	0.142	0.870	**0.913**	**0.074**	**0.118**	**0.913**
Overall	0.597	0.104	0.163	0.792	0.639	0.105	0.155	0.815	**0.796**	**0.075**	**0.116**	**0.900**

Table 19: Detailed results from models fine-tuned on OCR error detection: P (Precision), R (Recall) and F1 (F1 score) Metrics.

Subset	Albertina pt-PT			Albertina pt-BT			PortOldBERT-All-EP		
	P	R	F1	P	R	F1	P	R	F1
AdA	0.896	0.791	0.840	0.900	0.779	0.835	**0.910**	**0.832**	**0.870**
CHLMP	0.916	0.865	0.890	0.917	0.858	0.886	**0.931**	**0.885**	**0.908**
ESTER-Pt	0.888	0.800	0.842	0.883	0.795	0.837	**0.913**	**0.836**	**0.873**
Tycho	0.872	0.784	0.826	0.862	0.774	0.815	**0.884**	**0.811**	**0.846**
Overall	0.903	0.831	0.865	0.901	0.824	0.861	**0.920**	**0.859**	**0.888**