JabuticaBERT: Modern Portuguese Encoders from Scratch with RTD and Long-Context Training

Thiago Porto[1], Gabriel Gomes[1], Alexandre Bender[1], Ulisses Corrêa[1], Larissa Freitas[1], William Cruz[2], Marcellus Amadeus[2]
[1]Federal University of Pelotas, Brazil [2]Amadeus AI, Brazil
{trporto,gagomes,ulisses}@inf.ufpel.edu.br
{alexandre.thurow,larissaaf,williamalberto.cruz,7marcellus}@gmail.com

Abstract

Encoder-based language models remain essential for natural language understanding tasks such as classification, semantic similarity, and retrieval-augmented generation. However, the lack of high-quality monolingual encoders for Brazilian Portuguese poses a significant challenge to performance. In this work, we systematically explore the training of Portuguese-specific encoder models from scratch using two modern architectures: DeBERTa, trained with Replaced Token Detection (RTD), and ModernBERT, trained with Masked Language Modeling (MLM). All models are pre-trained on the large-scale Jabuticaba corpus. Our DeBERTa-Large model achieves results comparable to the state-of-the-art, with F1 scores of 0.920 on ASSIN2 RTE and 0.915 on LeNER. Crucially, it matches the performance of the 900M-parameter Albertina model while utilizing significantly fewer parameters. We also release custom tokenizers that reduce token fertility rates compared to multilingual baselines. These findings provide evidence that careful architectural choices and monolingual tokenization can yield competitive performance without massive model scaling.

1 Introduction

Encoder-based language models play a central role in natural language understanding tasks such as text classification, semantic similarity, information retrieval, and retrieval-augmented generation (RAG) (Devlin et al., 2018; Lewis et al., 2021). Although recent advances in natural language processing (NLP) have been dominated by large autoregressive decoders, encoders remain the backbone of systems that require robust semantic representations, efficiency at inference time, and scalable indexing over large document collections.

In the context of Brazilian Portuguese, the development of encoder models has lagged behind the rapid evolution seen in English. Most widely used Portuguese models, such as BERTimbau (Souza et al., 2020), still rely on early Transformer architectures (Devlin et al., 2018) that enforce a maximum sequence length of 512 tokens. While sufficient for sentence-level tasks, this constraint creates a significant bottleneck for modern workloads like long-document understanding and retrieval-augmented generation, where critical information often spans thousands of tokens (Beltagy et al., 2020; Wang et al., 2024). Conversely, the English ecosystem has introduced techniques explicitly designed to solve these scaling challenges, including Rotary Positional Embeddings (RoPE), FlashAttention, and alternating local–global attention patterns (Dao et al., 2022; Su et al., 2024b; Warner et al., 2024). The absence of these advancements in current Portuguese benchmarks highlights a growing technological gap and a clear opportunity for modernization.

A distinct but equally critical limitation lies in tokenization efficiency. Many existing Portuguese encoders rely on multilingual or weakly adapted vocabularies, which tend to fragment Portuguese text into significantly more subword units than comparable English models (Pires et al., 2019; Zago and Agnoletti dos Santos Pedotti, 2024). This inefficiency has compounding effects: it increases memory usage, slows down both training and inference, and effectively reduces the semantic capacity of the context window. Consequently, even a model with a nominally large context size may fail to exploit its full potential if the tokenizer artificially inflates sequence lengths.

To address current limitations, we systematically evaluate two complementary architectures: DeBERTa (He et al., 2021), utilizing the data-efficient Replaced Token Detection (RTD) objective (Clark et al., 2020; He et al., 2023), and ModernBERT (Warner et al., 2024), which integrates FlashAttention (Dao et al., 2022) and RoPE Embeddings (Su et al., 2024b) to enable scalable document-level

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 933–942

processing. We support these designs by training custom tokenizers and pre-training all models (Base and Large) from scratch on the Jabuticaba corpus (Amadeus et al., 2024). This setup minimizes sequence inflation (Pires et al., 2019), optimizes attention costs (Brenndoerfer, 2025), and ensures a controlled comparison free from English-centric initialization biases (Martin et al., 2020). Importantly, our comparison reflects two *bundled* modern design choices commonly used in practice: DeBERTa with RTD and ModernBERT with MLM. As a result, objective and architecture are not fully disentangled in this study; we therefore focus on establishing strong, reproducible Portuguese baselines and analyzing practical trade-offs (accuracy, parameter efficiency, and long-context readiness), leaving cross-over ablations (e.g., ModernBERT+RTD, DeBERTa+MLM) and long-context downstream evaluations as future work.

This work makes the following contributions:

- We train a suite of Brazilian Portuguese encoder models *from scratch* on the Jabuticaba corpus, covering two modern families: DeBERTa-style encoders with RTD and ModernBERT-style encoders with long-context training up to 8,192 tokens.
- We introduce Portuguese-specific tokenizers trained on large-scale Portuguese data and quantify token-efficiency improvements across distinct text domains.
- We provide a reproducible training and evaluation pipeline, and release the resulting models and tokenizers to support further research and deployment in Portuguese NLP.[1]

2 Background

The Transformer architecture serves as the foundation for state-of-the-art NLP, yet its efficacy is heavily dependent on specific design choices regarding training objectives, attention mechanisms, and input representation. To contextualize the architectural decisions made in this work, this section reviews the technical evolution of encoder models. We first contrast the standard MLM objective with the more sample-efficient RTD used in DeBERTa. Subsequently, we discuss recent advancements in attention efficiency—namely FlashAttention and RoPE Embeddings—that enable long-context processing in ModernBERT. Finally, we examine the critical role of custom tokenization strategies in optimizing computational efficiency and semantic representation for morphologically rich languages like Portuguese.

2.1 Evolution of Masked Language Modeling

The standard paradigm for pre-training encoder models was established by BERT (Devlin et al., 2018), utilizing the MLM objective. In this approach, approximately 15% of the input tokens are randomly replaced with a special `[MASK]` token, and the model is trained to reconstruct the original token based on the surrounding context. While effective, MLM suffers from significant sample inefficiency: the model only calculates loss and updates weights based on the small fraction of masked tokens, effectively ignoring the vast majority of the input sequence during backpropagation.

To address this limitation, recent architectures have adopted RTD, a method originally proposed by ELECTRA (Clark et al., 2020). Instead of masking tokens, RTD employs a generator network to replace a subset of tokens with plausible alternatives. The main model (the discriminator) is then tasked with determining whether each token in the sequence is original or replaced. Crucially, this binary classification task provides a *denser learning signal* because the loss is computed over every token in the sequence, rather than just the masked subset.

Building on these efficiency gains, the DeBERTa architecture (He et al., 2021) combines the RTD objective with a disentangled attention mechanism. Unlike BERT, which sums content and position embeddings into a single vector, DeBERTa represents each token using two distinct vectors: one for content and one for relative position. Attention scores are computed using disentangled matrices that explicitly model content-to-content and content-to-position dependencies. This separation allows the model to capture more granular syntactic relationships, which, combined with the denser signal from RTD, typically yields superior performance on natural language understanding tasks compared to standard MLM baselines.

2.2 Efficient Long-Context Architectures

A fundamental limitation of traditional Transformer encoders is the quadratic time and mem-

[1] Released models: JabuticaBERT-XSmall, JabuticaBERT-Large, modernJabuticaBERT-Base-1k, modernJabuticaBERT-Base-8k, modernJabuticaBERT-Large-1k, and modernJabuticaBERT-Large-8k.

ory complexity of self-attention with respect to sequence length N, which has historically restricted most BERT-based models to a maximum context of 512 tokens. This constraint severely limits applicability to long-document understanding tasks, such as legal analysis and scientific literature processing.

Recent encoder architectures overcome this bottleneck without relying on sparse or approximate attention by incorporating FlashAttention (Dao et al., 2022). FlashAttention is an IO-aware algorithm that reduces memory accesses between GPU high-bandwidth memory and on-chip SRAM through tiling, enabling exact attention computation with substantially lower memory overhead. In practice, this reduces the effective memory footprint from quadratic to linear with respect to sequence length, allowing efficient training on long inputs (e.g., 8,192 tokens).

In parallel, positional encoding mechanisms have evolved to better support long contexts. Traditional absolute positional embeddings often fail to generalize beyond the sequence lengths observed during training. RoPE address this limitation by encoding position as a rotation applied to query and key vectors (Su et al., 2024a). This formulation naturally captures relative positional information and enables more stable extrapolation to longer sequences.

Training exclusively on maximum-length sequences is computationally inefficient and can destabilize optimization. Consequently, modern pipelines adopt context expansion or sequence length curricula (Li et al., 2022), in which models are initially trained on shorter sequences and progressively exposed to longer inputs. This strategy facilitates stable learning of long-range dependencies while significantly reducing overall computational cost.

2.3 Tokenization in Low-to-Medium Resource Languages

The efficiency of a Transformer model is intrinsically linked to its tokenizer, which dictates how raw text is mapped to numerical input. In multilingual models such as mBERT or Llama, the vocabulary is shared across dozens of languages. Because vocabulary size is fixed (typically between 30k and 250k tokens), high-resource languages like English dominate the allocation. Consequently, medium-resource languages like Portuguese often suffer from suboptimal segmentation, where common words are fractured into an excessive number of subword units. This phenomenon, known as *high token fertility*, artificially inflates sequence lengths.

High fertility has two detrimental effects on model performance. First, it dilutes the effective capacity of the context window; a fixed limit of 8,192 tokens captures significantly less semantic content in Portuguese than in English if the average tokens-per-word ratio is high. Second, it increases computational overhead, as the model must process more distinct units to encode the same amount of information.

To mitigate these issues, recent research emphasizes the importance of language-specific tokenization. By training Byte Pair Encoding (BPE) or WordPiece algorithms exclusively on monolingual corpora, it is possible to generate vocabularies where the majority of domain-specific and common words are represented as single tokens. This optimization reduces the tokens-per-word ratio, thereby increasing the information density of the input embeddings and directly improving inference speed and memory efficiency without altering the underlying model architecture.

3 Related Works

The development of pre-trained language models has substantially advanced natural language understanding across a wide range of tasks. For Brazilian Portuguese, early approaches primarily relied on multilingual models such as mBERT, which benefit from cross-lingual transfer but often underperform monolingual counterparts due to vocabulary fragmentation and limited exposure to language-specific phenomena due to the fact that Portuguese is a low-resource language (Kalyan et al., 2021). These limitations motivated the creation of monolingual encoder models tailored to Portuguese.

BERTimbau (Souza et al., 2020) represents one of the first widely adopted monolingual BERT-based models for Brazilian Portuguese. It follows the original BERT architecture and is pre-trained using the MLM objective on large Portuguese corpora. Both base and large variants have been released and extensively used as reference models in downstream tasks such as textual entailment, semantic similarity, and named entity recognition. The success of BERTimbau demonstrates the importance of monolingual pre-training and language-specific tokenization for Portuguese natural language understanding.

More recently, Albertina (Rodrigues et al., 2023) has been introduced as a large-scale Portuguese encoder, significantly increasing the parameter count to approximately 900 million. Pre-trained using DeBERTa (He et al., 2021) as the base model, Albertina continues to rely on the MLM objective and follows a transformer-based encoder design, with performance gains largely attributed to model scaling. While this approach yields strong results, it also entails substantial computational and memory costs, raising questions about efficiency and accessibility for practical applications.

Despite recent advances in encoder architectures for Brazilian Portuguese, limitations remain in how these models are adapted and evaluated in monolingual settings. In particular, although RTD has been shown to provide denser supervision and improved sample efficiency in prior work, its adoption in Portuguese encoders has been limited. Most Portuguese encoder models continue to rely on MLM as their primary pre-training objective. Although modern architectures have been adopted these models are trained using MLM rather than RTD.

In addition, prior work frequently reuses tokenizers and base-model configurations originally developed for multilingual or high-resource language settings. Such tokenization strategies may not optimally capture Portuguese-specific morphological and lexical properties. Moreover, many studies adopt continued pre-training or fine-tuning from existing checkpoints rather than training both the tokenizer and the encoder entirely from scratch. This reliance on inherited tokenization and parameter initialization complicates the analysis of how architectural design, pre-training objectives, and tokenizer choices interact in fully monolingual Portuguese encoder models.

To address these limitations, we adopt a fully monolingual pre-training strategy for Brazilian Portuguese, training both encoder models and their tokenizers from scratch on large-scale Portuguese data. Modern encoder architectures are instantiated from random initialization, preserving their architectural design while avoiding reliance on multilingual or English-centric checkpoints. For DeBERTa, we further investigate RTD as the primary pre-training objective, in contrast to the commonly used MLM. By jointly training the tokenizer and encoder and avoiding continued pre-training from existing models, this approach enables a clearer analysis of how architectural choices, pre-training objectives, and language-specific properties interact in monolingual Portuguese encoders.

4 Methodology

4.1 Pre-training Data

All models in this study were trained from scratch using the Jabuticaba corpus, a large-scale collection of Brazilian Portuguese text designed specifically for language model pre-training (Amadeus et al., 2024). The corpus aggregates data from multiple domains, including web crawls, news articles, legal documents, and academic texts, providing broad linguistic and stylistic coverage.

To ensure controlled comparisons across architectures and training objectives, the same pre-training corpus was used for all models, including DeBERTa and ModernBERT variants, as well as for tokenizer training. No additional domain-specific filtering or task-oriented data selection was applied. This design choice isolates the effects of architectural decisions, pre-training objectives, and tokenization strategies from variations in data distribution.

Prior to training, the corpus underwent standard preprocessing steps commonly adopted in large-scale language modeling pipelines, including document-level deduplication, normalization of Unicode characters, and removal of corrupted or extremely short samples. These steps aim to reduce redundancy and noise while preserving the natural distribution of Brazilian Portuguese text.

4.2 Model Architectures

This work investigates two encoder architectures that reflect recent advances in Transformer design while remaining suitable for large-scale monolingual training: DeBERTa and ModernBERT. Rather than proposing a new architecture, our goal is to systematically adapt and evaluate these modern designs in the context of Brazilian Portuguese.

DeBERTa. Builds upon the original BERT encoder by introducing disentangled attention mechanisms that separately model content and positional information (He et al., 2023). Instead of summing token and positional embeddings into a single representation, DeBERTa represents each token using distinct vectors for content and relative position. Attention scores are computed through disentangled matrices that explicitly capture content-to-content and content-to-position interactions. This formulation has been shown to improve the mod-

eling of syntactic and semantic relationships compared to standard absolute positional embeddings.

In this study, we follow the architectural design introduced in DeBERTaV3, adopting standard base and large configurations as defined in prior work (He et al., 2023). These configurations allow us to analyze the impact of model scale while keeping architectural choices consistent with established best practices.

ModernBERT. Recently proposed encoder architecture designed to overcome the fixed-context and efficiency limitations of early Transformer models (Warner et al., 2024). It incorporates RoPE embeddings and optimized attention implementations such as Flash Attention, enabling native support for long input sequences with improved memory efficiency (Dao et al., 2022; Su et al., 2024b).

Unlike traditional BERT-style encoders, ModernBERT alternates local and global attention patterns and eliminates unnecessary padding operations, allowing the model to scale to document-level inputs while preserving strong performance on sentence-level tasks. These design choices make ModernBERT particularly suitable for scenarios where long-range dependencies are essential, such as document understanding and retrieval-oriented applications.

Model Scale. For both architectures, we train smaller (Base or XSmall) and large variants to assess the effect of model capacity under comparable training conditions, based on original DeBERTa and ModerBERT implementations.

4.3 Pre-training Objectives

To analyze the impact of pre-training objectives on Portuguese encoder models, we consider two widely adopted paradigms: MLM and RTD. These objectives differ fundamentally in how learning signals are distributed across input tokens, with direct implications for sample efficiency and convergence behavior.

Masked Language Modeling (MLM). Originally introduced with BERT (Devlin et al., 2018), trains the model to reconstruct a subset of randomly masked tokens based on their surrounding context. In standard configurations, approximately 15% of input tokens are selected for masking, and loss is computed only on these positions. While effective, this formulation is inherently sample-inefficient, as the majority of tokens in each sequence do not contribute directly to the optimization signal during training.

In this work, MLM is used as the pre-training objective for ModernBERT models, following the reference training procedure described in prior work (Warner et al., 2024). This choice reflects the canonical usage of ModernBERT and enables a faithful evaluation of its architectural advantages in long-context settings.

Replaced Token Detection (RTD). Discriminative pre-training objective introduced by ELECTRA (Clark et al., 2020) and later adopted by DeBERTaV3 (He et al., 2023). Instead of masking tokens, a generator model replaces a subset of input tokens with plausible alternatives. The main encoder, acting as a discriminator, is trained to predict whether each token in the sequence is original or replaced.

Unlike MLM, RTD computes loss at every token position, yielding a denser supervision signal per training example. This property has been shown to improve sample efficiency and accelerate convergence, particularly in scenarios where training data or computational resources are constrained.

4.4 Tokenizer Training

Tokenization directly affects the computational efficiency and effective context utilization of Transformer-based encoders, as the number of generated tokens determines memory usage, attention cost, and semantic density within a fixed context window (Brenndoerfer, 2025).

To mitigate the inefficiencies of multilingual or weakly adapted vocabularies, we train custom tokenizers specifically for Brazilian Portuguese. All tokenizers are trained from scratch on the Jabuticaba corpus, using a 10% random subset of the data employed for model pre-training, ensuring alignment between the tokenizer vocabulary and the linguistic statistics of the training distribution.

By optimizing subword segmentation for Portuguese morphology and lexical frequency, these tokenizers reduce token fragmentation and yield shorter token sequences on average compared to multilingual alternatives.

Improved token efficiency has two direct benefits. First, it increases the effective information density of input representations, allowing more semantic content to be encoded within the same maximum sequence length. Second, it reduces training and inference costs by lowering the number of at-

tention operations required to process equivalent textual inputs.

This efficiency gain is particularly important for long-context architectures such as ModernBERT: even when models support extended context windows, suboptimal tokenization can substantially limit the usable semantic context. By combining Portuguese-specific tokenizers with modern encoder architectures, we maximize the practical benefits of extended context lengths rather than relying solely on architectural capacity.

4.5 Training Procedure

We pre-train all encoders from scratch on the Jabuticaba corpus using a unified pipeline on 8×A100 GPUs. DeBERTa models are trained with RTD, while ModernBERT models are trained with MLM and the canonical efficiency features (bf16, unpadding, and sequence packing) from the reference implementation).

DeBERTa (JabuticaBERT) — RTD, 512 tokens. Both DeBERTa variants use fp16 and AdamW with warmup-linear schedule, lr 1×10^{-4} and wd 0.01. We train (i) **XSmall** for ∼250k steps with global batch size 1500, and (ii) **Large** for ∼700k steps with global batch size 304. We apply standard stabilization (e.g., gradient clipping) and use gradient accumulation only when needed to match effective batch sizes.

ModernBERT (modernJabuticaBERT) — MLM, 1024 tokens (base training). We train ModernBERT with bf16 and `StableAdamW` under a Warmup–Stable–Decay schedule, enabling **sequence packing**. **Base** is trained for ∼80k steps (lr 8×10^{-4}, wd 1×10^{-5}). **Large** is trained for ∼84k steps (lr 5×10^{-4}, wd 1×10^{-5}), initialized from the Base checkpoint using a Megatron-style / tiled initialization, keeping the same data and training setup. Long-context continuation is described in Sec. 4.6.

4.6 Context Length Expansion Strategy

Training with very long sequences from the start is expensive and can be unstable. Following ModernBERT (Warner et al., 2024), we adopt a *sequence-length curriculum* applied only to ModernBERT: we first train with a shorter maximum length (1024 tokens), then continue training at longer contexts up to 8,192 tokens.

RoPE adaptation for long context. ModernBERT uses local–global attention (local window 128; every third layer global). For context extension, we modify RoPE only for *global* attention layers: we increase the global RoPE θ from 10,000 to 160,000 while keeping the local RoPE θ at 10,000, as in the reference procedure (Warner et al., 2024). This preserves short-range inductive bias in local layers while enabling stable extrapolation in global layers.

Continued training schedule. We perform continued training in two stages: **Stage 1** (∼5k steps) continues with random sampling from the original corpus mixture to adapt optimization to longer sequences; **Stage 2** (∼2k steps) upsamples long-context sources (e.g., books) to reinforce long-range dependencies.

DeBERTa models follow their standard reference configuration with fixed context length (512) and do not use a length curriculum.

4.7 Evaluation Protocol

All models (see table 1) are evaluated on a diverse suite of Brazilian Portuguese downstream benchmarks covering semantic similarity, textual entailment, named entity recognition, and toxicity detection, on 1 A100 GPU.

To ensure comparability across architectures and pre-training objectives, we adopt a unified fine-tuning protocol: identical data splits, optimization settings, and early-stopping criteria are used for every task. We report standard metrics following prior work: Pearson correlation for semantic similarity, F1 for textual entailment and classification, and entity-level F1 for NER. Hyperparameters are tuned over five variables: learning rate sampled categorically from $\{10^{-7}, 10^{-6}, 2 \cdot 10^{-6}, 5 \cdot 10^{-6}, 10^{-5}, 2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}\}$; weight decay sampled uniformly from $[0, 0.1]$; mixed precision with $\texttt{fp16} \in \{\texttt{True}, \texttt{False}\}$; random seed from $\{42, 43, 44, 45\}$; and warmup steps sampled as an integer in $[0,\ 0.1 \times T]$, where T is the total number of training steps (approximated as $\texttt{epochs} \times |\mathcal{D}_{train}|$). For each dataset, we optimize the prescribed validation metric in its defined direction (maximize/minimize) and apply early stopping with patience 4 using epoch-level evaluations.

Model	Params (M)	Vocab (k)
DeBERTinha	40	50
BERTimbau Base	108	30
BERTimbau Large	334	30
JabuticaBERT XSmall	70	128
JabuticaBERT Large	434	128
Albertina 100M	138	50
Albertina 900M	884	128
modernJabuticaBERT Base	149	50
modernJabuticaBERT Large	395	50

Table 1: Model sizes and tokenizer vocabulary used in our comparison. For modernJabuticaBERT there is two context-length variants (1k and 8k) for both Base and Large.

5 Results

5.1 Tokenizer efficiency

This section evaluates the efficiency of the proposed tokenizers trained from scratch for Brazilian Portuguese. Tokenizer efficiency is analyzed through the average number of tokens per word, which serves as an indicator of vocabulary adequacy and subword fragmentation. Comparisons are conducted against the tokenizers of the baseline models and the BERT tokenizer, to assess how well each tokenizer captures Portuguese lexical and morphological patterns.

The analysis is performed on datasets drawn from tweets, Wikipedia articles, and literary texts by Machado de Assis and Monteiro Lobato, using identical preprocessing pipelines to ensure comparability across tokenizers. By evaluating diverse domains and writing styles, this setup enables a detailed examination of tokenizer behavior under varying linguistic conditions.

Table 2 reports the average number of tokens per word across different text domains. As expected, the tokenizer inherited from the English model exhibits the highest fragmentation across all domains, reflecting its limited suitability for Portuguese text. BERTimbau shows strong performance on the Wikipedia domain; however, this result should be interpreted with caution, as the BERTimbau tokenizer was trained primarily on Wikipedia data (Souza et al., 2020). Consequently, comparisons on the formal Wikipedia subset may favor BERTimbau and do not fully reflect its generalization to other domains.

Across the remaining domains, JabuticaBERT Large consistently achieves the lowest token-per-word ratios, indicating the most efficient tokenization overall. When excluding the Wikipedia domain, modernJabuticaBERT Base emerges as the second most efficient tokenizer, outperforming both BERTimbau and Albertina 900M on social media and literary texts. These results suggest that training tokenizers from scratch on large and diverse Portuguese corpora leads to improved morphological coverage and reduced fragmentation, particularly in non-formal and domain-diverse settings.

5.2 Benchmark Performance

We report downstream performance on standard Brazilian Portuguese benchmarks covering textual entailment, semantic textual similarity with the ASSIN 2 dataset (Real et al., 2020), named entity recognition with the LeNER dataset (Luz de Araujo et al., 2018), and toxicity detection with the ToldBR dataset (Leite et al., 2020). Models are evaluated under identical hyper-parameters search for fine-tuning to isolate the effects of architecture, tokenizer, and pre-training objective. Performance is measured using task-appropriate metrics, including F1-score and accuracy for classification and sequence labeling tasks and Pearson correlation for semantic similarity. Results are presented in Table 3, comparing our models with the baselines.

Among the evaluated systems, JabuticaBERT Large consistently achieves strong results across tasks, closely matching the performance of Albertina 900M while using approximately half the number of parameters. In particular, JabuticaBERT Large attains competitive scores on ASSIN2 RTE, STS and ToldBR, and achieves the highest F1 score on LeNER among the JabuticaBERT variants, indicating that high accuracy can be achieved without relying on extreme model scaling.

A distinguishing aspect of JabuticaBERT is the use of RTD during pre-training, which contrasts with the exclusive use of MLM in several baselines. The results suggest that RTD-based pre-training yields strong and consistent performance across tasks when applied in a fully monolingual setting. In contrast, the modernJabuticaBERT models, trained using MLM, exhibit more moderate but stable performance. Also, the extended-context configurations of these models were not fully exploited during evaluation, which may reveal advantages in practical information retrieval and search scenarios. Nevertheless, their consistent results indicate reliable learning behavior, and their architectural flexibility allows future exploration of alternative objectives such as RTD.

Model	Formal (Wiki)	Social (Tweets)	Machado de Assis	Monteiro Lobato	Overall
BERT (EN)	2.148	2.025	1.942	2.245	2.158
BERTimbau	1.585	1.847	1.653	1.920	1.774
modernJabuticaBERT	1.812	1.724	1.636	1.881	1.814
JabuticaBERT	**1.341**	**1.465**	**1.425**	**1.603**	**1.495**
Albertina 900M	1.862	1.913	1.944	2.211	2.060

Table 2: Tokenizer efficiency measured as the average number of tokens per word across different text domains (lower is better).

Model	ASSIN2 RTE (F1)	ASSIN2 STS (Pearson)	LeNER (F1)	ToldBR (Acc)
DeBERTinha	0.899	0.828	0.901	0.759
BERTimbau Base	0.898	0.842	0.879	0.762
BERTimbau Large	0.905	0.852	0.906	**0.768**
JabuticaBERT XSmall	0.897	0.841	0.892	0.761
JabuticaBERT Large	0.920	0.861	**0.915**	0.757
Albertina 900M	**0.921**	**0.865**	0.914	0.757
Albertina 100M	0.873	0.826	0.913	0.765
modernJabuticaBERT Base 1024	0.902	0.847	0.856	0.739
modernJabuticaBERT Base 8192	0.898	0.854	0.890	0.726
modernJabuticaBERT Large 1024	0.910	0.852	0.889	0.751
modernJabuticaBERT Large 8192	0.915	0.853	0.884	0.736

Table 3: Performance of Portuguese encoder models across ASSIN2, LeNER, and ToldBR benchmarks.

6 Conclusion

We addressed the scarcity of robust Brazilian Portuguese encoders by training DeBERTa-RTD and ModernBERT-MLM models from scratch on the Jabuticaba corpus, together with Portuguese-specific tokenizers. Empirically, DeBERTa-Large achieves strong performance on ASSIN2 and LeNER while remaining substantially more parameter-efficient than large-scale alternatives such as Albertina 900M. In parallel, ModernBERT variants extend monolingual encoders to an 8,192-token context window through efficient attention and rotary positional embeddings, providing a practical foundation for long-document processing. Since objective and architecture are coupled in our current design points, isolating RTD vs. MLM effects via cross-over ablations, as well as evaluating long-context capabilities on long-context downstream tasks, are promising directions for future work.

Limitations

Our study has four main limitations. (i) We evaluate two practical, bundled design points (DeBERTa+RTD and ModernBERT+MLM), so objective and architecture are not fully disentangled; cross-over ablations (e.g., ModernBERT+RTD, DeBERTa+MLM) are left for future work. (ii) Although we pre-train ModernBERT variants up to 8,192 tokens, most Portuguese benchmarks remain short-context, and we do not yet provide a dedicated long-context downstream evaluation. (iii) We do not report long-context efficiency measurements (e.g., tokens/s and peak memory) for 1k vs. 8k settings. (iv) While Jabuticaba provides high-quality pre-training data, we did not perform a formal decontamination analysis against downstream test sets, which may introduce unknown overlap risk.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We thank Amadeus AI and the SoberanIA project for the computational resources, technical support, and infrastructure that enabled the experiments conducted in this work, as well as for making the Jabuticaba corpus available.

References

Marcellus Amadeus, William Alberto Cruz Castañeda, José Roberto Homeli da Silva, and Rodrigo Scotti. 2024. Jabuticaba: The largest commercial corpus for llms in portuguese. *Preprint*, arXiv:2404.13680.

Iz Beltagy, Matthew E. Peters, and Arman Cohan.

2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Michael Brenndoerfer. 2025. Attention complexity: Quadratic scaling, memory limits & efficient alternatives. In *Language AI Handbook*, chapter 78. Michael Brenndoerfer. Chapter 78 of 380 in the *Language AI Handbook*. Published May 26 2025; updated May 29 2025.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing. *Preprint*, arXiv:2108.05542.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Conglong Li, Minjia Zhang, and Yuxiong He. 2022. The stability-efficiency dilemma: investigating sequence length warmup for training gpt models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *Preprint*, arXiv:1906.01502.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

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-*. *Preprint*, arXiv:2305.06721.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024a. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Jianlin Su, Yujie Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2024b. Rotary positional embedding: A method for relative position encoding in transformers. In *Proceedings of the Conference*. ArXiv preprint arXiv:2104.09864.

Jiajia Wang, Jimmy X. Huang, Xinhui Tu, Junmei Wang, Angela J. Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *Preprint*, arXiv:2403.00784.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Ricardo Zago and Luciane Agnoletti dos Santos Pedotti. 2024. Bertugues: A novel bert transformer model pre-trained for brazilian portuguesebertugues: Um modelo transformer bert inovador pré-treinado para o português brasileiro. *Semina: Ciências Exatas e Tecnológicas*, 45:e50630.