# **Análise Técnica do Modelo BERTimbau: Metodologia, Configuração e Desempenho**

## **I. Visão Geral do BERTimbau**

### **A. Introdução**

BERTimbau refere-se a uma família de modelos de linguagem baseados na arquitetura BERT (Bidirectional Encoder Representations from Transformers), especificamente pré-treinados para o português brasileiro.1 Desenvolvidos por Fábio Souza, Rodrigo Nogueira e Roberto Lotufo, estes modelos representaram um avanço significativo para o Processamento de Linguagem Natural (PLN) em português.1 A publicação principal que descreve o trabalho é "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", apresentada na conferência BRACIS em 2020 1, com trabalhos relacionados e atualizações subsequentes detalhando análises adicionais.2

Os modelos foram disponibilizados em duas variantes principais, seguindo a nomenclatura padrão do BERT: BERTimbau-Base e BERTimbau-Large, ambas na versão *cased* (sensível a maiúsculas e minúsculas).1 A sua disponibilização em plataformas de código aberto como Hugging Face e GitHub foi um fator crucial para sua ampla adoção pela comunidade de pesquisa e desenvolvimento.1

### **B. Motivação e Contexto**

A criação do BERTimbau foi motivada pela necessidade de modelos de linguagem de alta performance especificamente adaptados ao português brasileiro. Antes de seu desenvolvimento, a comunidade dependia largamente de modelos multilingues, como o Multilingual BERT (mBERT), que, embora capazes de processar português, frequentemente apresentavam desempenho subótimo em comparação com modelos treinados monolingualmente.1 Modelos multilingues enfrentam o que é conhecido como a "maldição da multilinguisidade", onde a capacidade fixa do modelo é compartilhada entre múltiplos idiomas, potencialmente diluindo a eficácia para qualquer idioma específico.2 BERTimbau visou preencher essa lacuna, fornecendo um modelo robusto treinado extensivamente em um corpus representativo do português brasileiro.4

Desde o seu lançamento, BERTimbau tornou-se um modelo fundamental para o PLN em português, servindo como base para inúmeras pesquisas subsequentes e aplicações práticas. Isso inclui adaptações para domínios específicos, como o jurídico (Legal BERTimbau) 7, e seu uso em uma variedade de tarefas downstream, como Reconhecimento de Entidades Nomeadas (NER), Classificação de Texto, Resposta a Perguntas (QA), entre outras.6

### **C. Implicações da Criação e Disponibilização**

A introdução do BERTimbau representou um marco democratizador para o PLN de ponta em português brasileiro. Ao superar o desempenho de modelos multilingues como o mBERT em tarefas específicas do idioma 1, BERTimbau forneceu à comunidade uma ferramenta mais eficaz. Antes dele, pesquisadores e desenvolvedores que trabalhavam com português tinham que aceitar o desempenho inferior do mBERT ou investir recursos computacionais significativos para treinar seus próprios modelos do zero. A disponibilização pública e de código aberto 1 removeu essa barreira, permitindo que um número maior de pessoas aproveitasse modelos de linguagem pré-treinados de alta qualidade, acelerando assim a pesquisa e o desenvolvimento de aplicações em português.

A estratégia de lançamento aberto foi fundamental para o impacto e a adoção generalizada do BERTimbau. A presença dos modelos em repositórios como GitHub 1 e Hugging Face 1 facilitou o acesso e a integração em projetos existentes. A vasta gama de trabalhos subsequentes que utilizam, adaptam ou comparam com BERTimbau 6 é uma consequência direta dessa acessibilidade. Isso não apenas fomentou uma comunidade mais ativa em torno de recursos de PLN para o português, mas também estabeleceu o BERTimbau como um *baseline* padrão de fato para a avaliação de novos modelos e técnicas no idioma.

## **II. Corpus de Pré-treinamento e Metodologia**

### **A. Corpus de Pré-treinamento: BrWaC**

O principal conjunto de dados utilizado para o pré-treinamento do BERTimbau foi o BrWaC (Brazilian Web as Corpus).1 Este é um corpus extenso e diversificado, composto por textos extraídos da web brasileira. Contém aproximadamente 2.68 bilhões de tokens distribuídos em cerca de 3.53 milhões de documentos, oferecendo uma ampla cobertura de domínios e estilos linguísticos do português brasileiro.5 Fontes posteriores mencionam que etapas de limpeza foram aplicadas ao BrWaC antes do treinamento, incluindo a remoção de tags HTML e a correção de problemas de codificação de caracteres (mojibakes) usando bibliotecas como BeautifulSoup e ftfy.10 Adicionalmente, artigos da Wikipédia em português foram utilizados, particularmente durante a fase de geração do vocabulário.8

### **B. Objetivos de Pré-treinamento**

O pré-treinamento do BERTimbau seguiu os objetivos padrão estabelecidos pelo modelo BERT original: Masked Language Model (MLM) e Next Sentence Prediction (NSP).8

Uma distinção importante na implementação do MLM foi o uso de *Whole Word Masking* (WWM).1 Ao contrário do mascaramento padrão do BERT, que seleciona aleatoriamente tokens de subpalavras (WordPieces) para mascarar, o WWM garante que, se um token pertencente a uma palavra for selecionado para mascaramento, todos os outros tokens que compõem a mesma palavra também sejam mascarados. O objetivo então é prever a palavra inteira original. Essa abordagem força o modelo a aprender representações que podem ser mais sensíveis aos limites das palavras, o que pode ser benéfico para tarefas downstream que dependem fortemente da semântica de palavras inteiras, como o Reconhecimento de Entidades Nomeadas (NER).

### **C. Tokenização**

O método de tokenização empregado pelo BERTimbau é o WordPiece 8, o mesmo utilizado pelo BERT original. O processo de geração do vocabulário, no entanto, foi adaptado para o português. Inicialmente, um vocabulário foi gerado usando a ferramenta SentencePiece com o algoritmo Byte Pair Encoding (BPE) sobre o corpus BrWaC e textos da Wikipédia em português. Subsequentemente, este vocabulário foi convertido para o formato WordPiece exigido pela arquitetura BERT.8

O tamanho final do vocabulário é de aproximadamente 30.000 unidades de subpalavras.8 Os modelos disponibilizados são *cased*, o que significa que diferenciam letras maiúsculas de minúsculas (bert-base-portuguese-cased, bert-large-portuguese-cased).1 Isso é refletido nos exemplos de uso, que especificam o parâmetro do\_lower\_case=False ao carregar o tokenizador.1

### **D. Implicações da Metodologia de Pré-treinamento**

A escolha do corpus BrWaC, complementado pela Wikipédia em português, juntamente com a criação de um vocabulário WordPiece customizado, constitui uma diferença fundamental em relação a modelos multilingues como o mBERT. Enquanto o mBERT utiliza um único vocabulário compartilhado entre 104 idiomas, treinado principalmente na Wikipédia 2, o vocabulário do BERTimbau foi otimizado especificamente para o português brasileiro, derivado de um corpus nativo muito maior e mais diversificado (BrWaC).1 Essa otimização resulta em uma tokenização mais eficiente para o português – ou seja, menos subtokens são necessários, em média, para representar uma palavra – comparado à abordagem genérica do mBERT.8 Uma tokenização mais eficiente frequentemente se correlaciona com melhor desempenho em tarefas downstream, pois o modelo opera com unidades semanticamente mais significativas.8 A análise da distribuição de subtokens em exemplos de teste confirma essa diferença.23

A adoção da estratégia de *Whole Word Masking* (WWM) 1, em vez do mascaramento padrão de subtokens, pode oferecer vantagens particulares para certas tarefas em português. Ao forçar o modelo a prever palavras inteiras com base no contexto, o WWM pode aprimorar a compreensão da semântica no nível da palavra. Tarefas como NER, onde o BERTimbau demonstrou forte desempenho 1, frequentemente envolvem a identificação de entidades que correspondem a palavras ou frases completas. Acredita-se que o WWM melhora implicitamente o desempenho nessas tarefas ao reforçar a consciência dos limites das palavras durante o pré-treinamento.

O tamanho do vocabulário, cerca de 30.000 subtokens 8, representa um equilíbrio entre a capacidade de capturar a riqueza lexical do português e a manutenção da eficiência computacional. Um vocabulário significativamente maior poderia capturar mais nuances, mas aumentaria o tamanho do modelo e o custo computacional. Um vocabulário muito menor poderia levar a um número excessivo de palavras fora do vocabulário (OOV) ou a uma segmentação excessiva em subtokens. O tamanho escolhido é comparável ao do vocabulário original do BERT em inglês, sugerindo que foi considerado um ponto de equilíbrio razoável para o português, com base nos dados de treinamento disponíveis. Trabalhos posteriores, como o BERTugues 9, exploraram otimizações adicionais, como a remoção de caracteres raramente usados em português, indicando que o design de vocabulários para idiomas específicos continua sendo uma área ativa de pesquisa.

## **III. Configuração de Treinamento e Arquitetura**

### **A. Arquiteturas do Modelo (Base vs. Large)**

BERTimbau foi lançado em duas variantes arquiteturais, seguindo o padrão BERT 1:

* **BERTimbau-Base:**
  + Camadas (Transformer Blocks): L=12
  + Tamanho da Camada Oculta: H=768
  + Cabeças de Auto-Atenção: A=12
  + Número Total de Parâmetros: Aproximadamente 110 milhões.4
* **BERTimbau-Large:**
  + Camadas (Transformer Blocks): L=24
  + Tamanho da Camada Oculta: H=1024
  + Cabeças de Auto-Atenção: A=16
  + Número Total de Parâmetros: Aproximadamente 340-355 milhões 4 (O padrão BERT Large é 340M).

Essas arquiteturas são implementações diretas das configurações Base e Large propostas no artigo original do BERT.4

### **B. Hiperparâmetros e Procedimento de Treinamento**

Ambas as variantes do BERTimbau foram pré-treinadas por um total de 1.000.000 de passos.1 Os detalhes específicos diferem ligeiramente:

* **BERTimbau Base:**
  + Taxa de Aprendizagem (Learning Rate): Pico de 1×10−4, com aquecimento (*warmup*) durante os primeiros 10.000 passos, seguido por um decaimento linear.23
  + Tamanho do Lote (Batch Size): 128 sequências.23 Presume-se que este seja o tamanho do lote global, embora não explicitamente rotulado como tal.
  + Comprimento da Sequência: 512 tokens durante todo o treinamento.23
* **BERTimbau Large:**
  + Taxa de Aprendizagem (Learning Rate): A taxa de aprendizagem de pico para o pré-treinamento não é explicitamente declarada na fonte principal 23, mas provavelmente seguiu um cronograma semelhante ao Base. A taxa de aprendizagem usada para *fine-tuning* em tarefas downstream foi mencionada como 1×10−5.23
  + Tamanho do Lote (Batch Size): 256 sequências para os primeiros 900.000 passos, reduzido para 128 sequências nos últimos 100.000 passos.23
  + Comprimento da Sequência: 128 tokens para os primeiros 900.000 passos, aumentado para 512 tokens nos últimos 100.000 passos.23

### **C. Hardware e Tempo de Treinamento**

O treinamento dos modelos BERTimbau foi realizado utilizando instâncias TPU (Tensor Processing Unit) v3-8 do Google Cloud.23 O projeto contou com créditos de nuvem do Google, conforme agradecimentos nos materiais de divulgação.1

A duração total do processo de pré-treinamento foi:

* **BERTimbau Base:** Aproximadamente 4 dias.23
* **BERTimbau Large:** Aproximadamente 7 dias.23

### **D. Janela de Contexto e Tokens de Entrada Máximos**

O comprimento máximo da sequência, ou janela de contexto, que os modelos BERTimbau podem processar é de 512 tokens (max\_position\_embeddings).23 Isso significa que o número máximo de tokens de entrada para qualquer exemplo é efetivamente limitado a 512, incluindo tokens especiais como e. Embora algumas fontes baseadas em inspeção de configuração online indiquem que essa informação não estava disponível 1, o documento PDF original do exame de qualificação confirma o valor de 512 para a configuração final dos modelos.23

### **E. Implicações da Configuração de Treinamento**

A estratégia de treinamento em duas fases adotada para o BERTimbau Large – começando com um comprimento de sequência menor (128) e um tamanho de lote maior (256), e depois mudando para um comprimento de sequência maior (512) e um tamanho de lote menor (128) nos últimos 10% do treinamento 23 – reflete as práticas do treinamento original do BERT. Esta é uma abordagem pragmática para gerenciar os recursos computacionais intensivos necessários. Treinar um modelo grande como o BERT Large com sequências longas (512) desde o início é extremamente caro em termos de memória e tempo. A fase inicial com sequências mais curtas permite passos de treinamento mais rápidos e lotes maiores, acelerando a convergência em padrões contextuais mais locais. A fase final, com o comprimento total da sequência, permite que o modelo aprenda as dependências de longo alcance que são cruciais para muitas tarefas de PLN complexas. Essa abordagem representa uma otimização prática que equilibra a velocidade do treinamento com a capacidade final do modelo, mas também sublinha o orçamento computacional substancial necessário (7 dias em uma TPU v3-8).

O uso explícito de TPUs v3-8 e a menção aos créditos do Google Cloud 1 destacam a infraestrutura computacional significativa necessária para pré-treinar modelos de linguagem em larga escala como o BERTimbau. TPUs são hardware especializado otimizado para machine learning. Uma instância v3-8 é uma configuração poderosa, e os tempos de treinamento relatados 23 indicam um custo computacional considerável. Isso sugere que a replicação de tais esforços de pré-treinamento não é trivial e depende fortemente do acesso a recursos computacionais avançados, reforçando o valor da decisão dos autores de liberar publicamente os modelos pré-treinados.

Embora a arquitetura central siga o padrão BERT 4 e os objetivos de pré-treinamento sejam os mesmos 8, é a combinação específica de dados em português (corpus BrWaC 1), um vocabulário adaptado 8, a técnica de WWM 1 e o regime de treinamento descrito 23 que constitui a "receita" única do BERTimbau. É essa combinação de fatores que define o modelo e explica seu desempenho superior em comparação com modelos multilingues genéricos para tarefas em português.1

## **IV. Análise de Desempenho em Benchmarks**

### **A. Visão Geral das Tarefas Avaliadas**

A publicação original e a documentação associada ao BERTimbau focaram sua avaliação em três tarefas de PLN em português, utilizando conjuntos de dados específicos da língua 1:

1. **Similaridade Textual Semântica (Sentence Textual Similarity - STS):** Avaliada no dataset ASSIN 2. A métrica utilizada foi a correlação de Pearson entre as pontuações de similaridade previstas pelo modelo e as anotações humanas.
2. **Reconhecimento de Implicação Textual (Recognizing Textual Entailment - RTE):** Também avaliada no dataset ASSIN 2. A métrica foi o F1-score.
3. **Reconhecimento de Entidades Nomeadas (Named Entity Recognition - NER):** Avaliada no dataset First HAREM / MiniHAREM, com variantes de 5 e 10 classes de entidades. A métrica foi o F1-score.

### **B. Desempenho nos Benchmarks Principais (vs. mBERT e SOTA Anterior)**

Os resultados comparativos, conforme apresentados na documentação do projeto 1, demonstram consistentemente a superioridade do BERTimbau sobre o mBERT e o estado da arte (SOTA) anterior para o português na época da publicação:

| **Tarefa** | **Dataset de Teste** | **BERTimbau-Large** | **BERTimbau-Base** | **mBERT** | **SOTA Anterior** |
| --- | --- | --- | --- | --- | --- |
| STS | ASSIN 2 | 0.852 | 0.836 | 0.809 | 0.83 1 |
| RTE | ASSIN 2 | 90.0 | 89.2 | 86.8 | 88.3 1 |
| NER (5 classes) | MiniHAREM | 83.7 | 83.1 | 79.2 | 82.3 2 |
| NER (10 classes) | MiniHAREM | 78.5 | 77.6 | 73.1 | 74.6 2 |

*Referências na tabela original: 1 Multilingual Transformer Ensembles for Portuguese Natural Language Task. 2 Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition.*

A análise desses resultados é clara: ambas as versões do BERTimbau (Base e Large) superaram o mBERT e os SOTAs anteriores em todas as tarefas e conjuntos de dados reportados.1 O BERTimbau-Large consistentemente apresentou um desempenho ligeiramente superior ao Base. Esses resultados quantitativos validam empiricamente a hipótese central por trás do BERTimbau: o pré-treinamento específico para um idioma resulta em benefícios significativos de desempenho em comparação com modelos multilingues de capacidade fixa.1 A diferença de desempenho destaca as limitações inerentes à abordagem multilingue do mBERT para o português.2

### **C. Desempenho em Resposta a Perguntas (SQuAD-BR)**

Em relação à consulta sobre "squadFIT 2.0", é importante notar que os benchmarks relevantes para BERTimbau são as traduções para o português do dataset SQuAD (Stanford Question Answering Dataset).

* **SQuAD-BR v1.1:** Uma tradução do SQuAD v1.1 para o português foi realizada pelo grupo Deep Learning Brasil, utilizando tradução automática (Google Translate) e correções via crowdsourcing.8 Este dataset é referido como "SQuAD-BR". O BERTimbau foi rapidamente aplicado a esta tarefa através de *fine-tuning*:
  + Um trabalho inicial de Pierre Guillou, fazendo *fine-tuning* do BERTimbau Base, alcançou F1=82.50 e Exact Match (EM)=70.49.16
  + Posteriormente, Silva et al. 8 relataram ter melhorado esses resultados, estabelecendo um novo estado da arte no SQuAD-BR v1.1 ao fazer *fine-tuning* do BERTimbau Large. Embora os scores exatos não estejam nos trechos fornecidos, a reivindicação de SOTA é explícita.8 A rápida adoção e os subsequentes recordes de SOTA no SQuAD-BR demonstram o poder capacitador do lançamento de modelos fundacionais fortes para idiomas específicos. Assim que um benchmark relevante (SQuAD-BR) se tornou disponível 8, a comunidade pôde rapidamente alavancar o BERTimbau 8 para alcançar resultados de ponta, validando o impacto pretendido do recurso.
* **SQuAD-BR v2.0:** Existe uma tradução para o português do SQuAD v2.0.28 No entanto, os resultados de desempenho específicos do BERTimbau nesta versão são menos documentados nos materiais fornecidos em comparação com a v1.1. O snippet 30 mostra resultados para um BERT base genérico no SQuAD v2.0 em inglês (F1=78.08), não para o BERTimbau no SQuAD-BR v2.0. O snippet 27 menciona QA no SQuAD 2.0, mas não fornece resultados do BERTimbau.
* **FaQuAD:** Outro dataset de QA extrativa em português é o FaQuAD.31 Os snippets 21 mostram resultados de *fine-tuning* do BERTimbau em uma tarefa derivada chamada "FaQuAD-NLI" (Inferência de Linguagem Natural, não a tarefa original de QA). Nesses testes, BERTimbau-Large alcançou 92.26 e BERTimbau-Base ('small') 93.07 de acurácia, comparando-se favoravelmente a modelos menores customizados (TTL). É crucial notar que estes são resultados de NLI, não de QA extrativa no FaQuAD.

### **D. Desempenho em GLUE/SuperGLUE**

A publicação original do BERTimbau 2 e sua documentação associada 1 **não** reportam resultados nos benchmarks padrão em inglês GLUE (General Language Understanding Evaluation) ou SuperGLUE. Isso é esperado, dado que o modelo e sua avaliação foram focados em demonstrar utilidade para tarefas e conjuntos de dados em português brasileiro (ASSIN 2, MiniHAREM, SQuAD-BR).1

A ausência de scores GLUE/SuperGLUE não deve ser interpretada como uma deficiência do modelo, mas sim como um reflexo do seu escopo direcionado (português brasileiro) e das práticas de avaliação comuns na época para modelos não-ingleses. Avaliar um modelo português nesses benchmarks exigiria capacidades robustas de transferência interlingual ou avaliação em versões traduzidas, o que não era o foco principal do trabalho original. Isso ressalta a importância de recursos de avaliação específicos para cada idioma e as limitações de depender exclusivamente de leaderboards dominados pelo inglês para avaliar as capacidades de modelos em diferentes línguas.

### **E. Outros Benchmarks e Tarefas**

A versatilidade do BERTimbau é evidenciada por seu uso como *baseline* ou através de *fine-tuning* em diversas outras tarefas e domínios do PLN em português:

* Adaptações para NER em contextos específicos.6
* Rotulação de Papéis Semânticos (Semantic Role Labeling - SRL).14
* Detecção de Notícias Falsas (Fake News Detection).13
* Classificação de Texto em diversos contextos.21
* PLN Jurídico.7
* Extração de Eventos.17
* Filtragem de Textos Gerados.18
* Análise de Sentimentos (comparado com BERTaú).36

Comparações em trabalhos mais recentes, como a avaliação contra modelos TTL 21, mostram que, embora BERTimbau (especialmente a versão Large) permaneça um *baseline* forte e competitivo em várias tarefas de classificação (IMDB, FaQuAD-NLI, HateBr, Assin2, AgNews), novas arquiteturas (como as baseadas em Llama 21) ou modelos treinados em conjuntos de dados diferentes ou maiores 37 podem igualar ou superar seu desempenho em tarefas específicas. Isso reflete a rápida evolução do campo de PLN e posiciona o BERTimbau como um modelo fundacional importante, mas representativo de um ponto específico no desenvolvimento contínuo de LLMs para o português.5

## **V. Especificações Técnicas Consolidadas**

A tabela a seguir sumariza as principais especificações técnicas e resultados de benchmark para as variantes Base e Large do BERTimbau, com base nas informações extraídas das fontes analisadas.

| **Modelo** | **Corpus Principal** | **Objetivos** | **Tokenizador** | **Tamanho do Vocabulário** | **Casing** | **Camadas (L)** | **Tamanho Oculto (H)** | **Cabeças de Atenção (A)** | **Parâmetros Totais** | **Passos Totais** | **Taxa de Aprendizagem (Pico)** | **Tamanho do Lote (Global)** | **Comp. Sequência (Treinamento)** | **Hardware** | **Tempo de Treinamento** | **Comp. Máx. Sequência (Contexto)** | **STS (ASSIN 2 - Pearson)** | **RTE (ASSIN 2 - F1)** | **NER (MiniHAREM 5 classes - F1)** | **NER (MiniHAREM 10 classes - F1)** | **QA (SQuAD-BR v1.1 - F1)** | **NLI (FaQuAD-NLI - Acurácia)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BERTimbau-Base | BrWaC (+ Wikipedia para vocabulário) 1 | MLM (WWM), NSP 1 | WordPiece (Vocabulário customizado PT-BR) 8 | ~30.000 8 | Cased 1 | 12 4 | 768 4 | 12 4 | ~110 Milhões 4 | 1.000.000 1 | 1×10−4 (10k warmup, linear decay) 23 | 128 23 | 512 23 | TPU v3-8 1 | ~4 dias 23 | 512 23 | 0.836 1 | 89.2 1 | 83.1 1 | 77.6 1 | 82.50 (fine-tuning por Guillou 16) | 93.07 (reportado como 'small' 21) |
| BERTimbau-Large | BrWaC (+ Wikipedia para vocabulário) 1 | MLM (WWM), NSP 1 | WordPiece (Vocabulário customizado PT-BR) 8 | ~30.000 8 | Cased 1 | 24 4 | 1024 4 | 16 4 | ~340-355 Milhões 4 | 1.000.000 1 | Não especificado (prov. similar); Fine-tuning LR: 1×10−5 23 | 256 (primeiros 900k passos), 128 (últimos 100k passos) 23 | 128 (primeiros 900k passos), 512 (últimos 100k passos) 23 | TPU v3-8 1 | ~7 dias 23 | 512 23 | 0.852 1 | 90.0 1 | 83.7 1 | 78.5 1 | SOTA (fine-tuning por Silva et al. 8)¹ | 92.26 21 |

¹ *Scores exatos para o SOTA de Silva et al. no SQuAD-BR v1.1 com BERTimbau-Large não foram encontrados nos snippets, mas a reivindicação de SOTA é clara.*

## **VI. Conclusões**

Este relatório detalhou a metodologia de treinamento, configuração e desempenho dos modelos BERTimbau Base e Large, desenvolvidos especificamente para o português brasileiro. A análise das fontes disponíveis confirma que BERTimbau foi pré-treinado no extenso corpus BrWaC, utilizando os objetivos MLM (com Whole Word Masking) e NSP, e um vocabulário WordPiece customizado de ~30k tokens otimizado para o idioma. O treinamento foi realizado em hardware TPU v3-8, exigindo recursos computacionais significativos.

Os resultados de benchmark apresentados na publicação original demonstram inequivocamente que BERTimbau (ambas as variantes) superou o desempenho do mBERT e de abordagens anteriores em tarefas chave de PLN para o português, como STS, RTE e NER nos datasets ASSIN 2 e MiniHAREM. Subsequentemente, o modelo foi utilizado para alcançar o estado da arte na tarefa de Resposta a Perguntas no dataset SQuAD-BR v1.1. Esses resultados validam a eficácia da abordagem de pré-treinamento monolingue e específica para o idioma.

O lançamento público e de código aberto do BERTimbau teve um impacto considerável na comunidade de PLN em português, servindo como um catalisador para pesquisas e desenvolvimentos subsequentes.7 Ele se estabeleceu como um *baseline* robusto e ainda é frequentemente utilizado para comparação.5

É importante reconhecer as limitações contextuais: o treinamento exigiu recursos computacionais avançados, potencialmente limitando a replicabilidade 1, e a avaliação original concentrou-se em benchmarks específicos do português, não incluindo métricas de benchmarks multilingues ou ingleses como GLUE/SuperGLUE. Além disso, o campo de PLN está em constante evolução, e modelos mais recentes, utilizando arquiteturas diferentes ou treinados em conjuntos de dados ainda maiores, podem apresentar desempenho superior em certas tarefas.9

Em suma, BERTimbau representa um marco fundamental para o Processamento de Linguagem Natural em português brasileiro. Sua criação e disponibilização não apenas forneceram uma ferramenta de alto desempenho para a comunidade, mas também impulsionaram o progresso da área, estabelecendo um padrão de qualidade e demonstrando o valor de recursos linguísticos dedicados.

#### Referências citadas

1. neuralmind-ai/portuguese-bert: Portuguese pre-trained ... - GitHub, acessado em abril 29, 2025, <https://github.com/neuralmind-ai/portuguese-bert>
2. (PDF) BERTimbau: Pretrained BERT Models for Brazilian Portuguese - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/345395208_BERTimbau_Pretrained_BERT_Models_for_Brazilian_Portuguese>
3. BERTimbau: Pretrained BERT Models for Brazilian Portuguese | springerprofessional.de, acessado em abril 29, 2025, <https://www.springerprofessional.de/en/bertimbau-pretrained-bert-models-for-brazilian-portuguese/18488758>
4. Transformers for Brazilian Portuguese Text Classification of News Using Deep Learning and Natural Language Processing Models Bas, acessado em abril 29, 2025, <https://www.iiis.org/CDs2022/CD2022Summer/papers/SA702PC.pdf>
5. PeLLE: Encoder-based language models for Brazilian Portuguese based on open data, acessado em abril 29, 2025, <https://arxiv.org/html/2402.19204v1>
6. marquesafonso/bertimbau-large-ner-selective - Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/marquesafonso/bertimbau-large-ner-selective>
7. rufimelo/Legal-BERTimbau-base - Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/rufimelo/Legal-BERTimbau-base>
8. New State-of-the-Art for Question Answering on Portuguese SQuAD v1.1, acessado em abril 29, 2025, <https://sol.sbc.org.br/index.php/kdmile/article/download/24974/24795>
9. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis | Request PDF - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/374537805_BERT_models_for_Brazilian_Portuguese_Pretraining_evaluation_and_tokenization_analysis>
10. View of BERTugues: A Novel BERT Transformer Model Pre-trained for Brazilian Portuguese, acessado em abril 29, 2025, <https://ojs.uel.br/revistas/uel/index.php/semexatas/article/view/50630/51525>
11. arXiv:2201.03382v1 [cs.CL] 10 Jan 2022, acessado em abril 29, 2025, <https://arxiv.org/pdf/2201.03382>
12. Open Sentence Embeddings for Portuguese with the Serafim PT\* encoders family - arXiv, acessado em abril 29, 2025, <https://arxiv.org/html/2407.19527v1>
13. Talendar/br\_fake\_news\_detection: Fake news detection on texts written in Brazilian Portuguese. - GitHub, acessado em abril 29, 2025, <https://github.com/Talendar/br_fake_news_detection>
14. asofiaoliveira/srl\_bert\_pt: Portuguese BERT and XLM-R models fine-tuned in semantic role labeling. - GitHub, acessado em abril 29, 2025, <https://github.com/asofiaoliveira/srl_bert_pt>
15. Example doesn't work (documentation outdated? ) · Issue #33 · neuralmind-ai/portuguese-bert - GitHub, acessado em abril 29, 2025, <https://github.com/neuralmind-ai/portuguese-bert/issues/33>
16. pierreguillou/bert-base-cased-squad-v1.1-portuguese - Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese>
17. arXiv:2408.16932v1 [cs.CL] 29 Aug 2024, acessado em abril 29, 2025, <https://arxiv.org/pdf/2408.16932>
18. Tucano - Nicholas Kluge Corrêa, acessado em abril 29, 2025, <https://nkluge-correa.github.io/Tucano/>
19. Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models? - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2024.propor-1.31.pdf>
20. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language - arXiv, acessado em abril 29, 2025, <https://arxiv.org/pdf/2406.10806>
21. nicholasKluge/TeenyTinyLlama-160m-IMDB - Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/nicholasKluge/TeenyTinyLlama-160m-IMDB>
22. cabrita: closing the gap for foreign languages - arXiv, acessado em abril 29, 2025, <https://arxiv.org/pdf/2308.11878>
23. core.ac.uk, acessado em abril 29, 2025, <https://core.ac.uk/download/427533232.pdf>
24. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study - MDPI, acessado em abril 29, 2025, <https://www.mdpi.com/1424-8220/22/21/8184>
25. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study - PMC, acessado em abril 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9655936/>
26. neuralmind/bert-large-portuguese-cased · Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/neuralmind/bert-large-portuguese-cased>
27. New State-of-the-Art for Question Answering on Portuguese SQuAD v1.1 | Semantic Scholar, acessado em abril 29, 2025, <https://www.semanticscholar.org/paper/New-State-of-the-Art-for-Question-Answering-on-v1.1-Silva-Laterza/500c5980255d9183098a90d52ea6f92b7d6b910f>
28. ajdavidl/Portuguese-NLP: List of resources and tools developed with focus on Portuguese. - GitHub, acessado em abril 29, 2025, <https://github.com/ajdavidl/Portuguese-NLP>
29. piEsposito/br-quad-2.0: Stanford Question Answering Dataset (SQuAD) 2.0 translated to Brazilian Portuguese (PT-BR) language. - GitHub, acessado em abril 29, 2025, <https://github.com/piEsposito/br-quad-2.0>
30. QA squadv2.0 Bertbase - NGC Catalog - NVIDIA, acessado em abril 29, 2025, <https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/qa_squadv2_0_bertbase>
31. Evaluation of Question Answer Generation for Portuguese: Insights and Datasets - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2024.findings-emnlp.306.pdf>
32. README.md - unicamp-dl/cross-lingual-analysis - GitHub, acessado em abril 29, 2025, <https://github.com/unicamp-dl/cross-lingual-analysis/blob/main/README.md>
33. TeenyTinyLlama/Evaluation/EVAL.md at main · Nkluge-correa, acessado em abril 29, 2025, <https://github.com/Nkluge-correa/TeenyTinyLlama/blob/main/Evaluation/EVAL.md>
34. (PDF) Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/364764601_Towards_Transfer_Learning_Techniques-BERT_DistilBERT_BERTimbau_and_DistilBERTimbau_for_Automatic_Text_Classification_from_Different_Languages_A_Case_Study>
35. felipemaiapolo/legalnlp: LegalNLP - Natural Language Processing Methods for the Brazilian Legal Language - GitHub, acessado em abril 29, 2025, <https://github.com/felipemaiapolo/legalnlp>
36. BERTaú: Itaú BERT for digital customer service - arXiv, acessado em abril 29, 2025, <https://arxiv.org/pdf/2101.12015>
37. Paper page - Tucano: Advancing Neural Text Generation for Portuguese - Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/papers/2411.07854>
38. Sabiá: Portuguese Large Language Models | Request PDF - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/374641537_Sabia_Portuguese_Large_Language_Models>