Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6077
This work is licensed under a Creative Commons Attribution 4.0 International License.

Artigo de Pesquisa/Research Paper

Reconhecimento de Entidades Nomeadas em Português: Comparação de Modelos Pré-Treinados com Fine-Tuning

Named Entity Recognition in Portuguese: Comparison of Pre-Trained Models with Fine-Tuning

Guilherme Tapajós [Universidade do Estado do Amazonas | *gcta.snf21@uea.edu.br*]
Tiago de Melo [Universidade do Estado do Amazonas |*tmelo@uea.edu.br*]
Elloá B. Guedes [Universidade do Estado do Amazonas |*ebgcosta@uea.edu.br*]
Fábio Santos [Universidade do Estado do Amazonas |*fssilva@uea.edu.br*]

Grupo de Pesquisa Laboratório de Sistemas Inteligentes, Escola Superior de Tecnologia, Universidade do Estado do Amazonas, Av. Darcy Vargas, 1.200 - Parque Dez de Novembro, Manaus, AM, 69050-020, Brasil.

Resumo. O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa central do Processamento de Linguagem Natural (PLN), mas ainda limitada no português pela escassez de recursos. Este estudo analisa o fine-tuning de variantes base, largas e destiladas dos modelos BERT e RoBERTa, em configurações multilíngues e monolíngues, usando os conjuntos Harem, LeNER-Br e GeoCorpus. O XLM-RoBERTa-large obteve F1-scores de 83,8% no Harem e 92,3% no LeNER-Br, enquanto o BERT-large-cased alcançou 87,8% no GeoCorpus, superando as baselines em até cinco pontos percentuais. Modelos multilíngues mostraram melhor adaptabilidade e as versões destiladas mantiveram desempenho competitivo com menor custo computacional. Os resultados evidenciam que o ajuste fino de grandes modelos pré-treinados é uma estratégia eficaz para impulsionar o REN em português.

Abstract. Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP), yet its progress in Portuguese is still hindered by scarce resources. This study examines the fine-tuning of base, large and distilled variants of BERT and RoBERTa models, in both multilingual and monolingual configurations, using the Harem, LeNER-Br and GeoCorpus datasets. XLM-RoBERTa-large achieved F1 scores of 83.8% on Harem and 92.3% on LeNER-Br, while BERT-large-cased reached 87.8% on GeoCorpus, outperforming the baselines by up to five percentage points. Multilingual models showed greater adaptability, and distilled versions maintained competitive performance with lower computational cost. The results demonstrate that fine-tuning large pre-trained models is an effective strategy for advancing NER in Portuguese.

Palavras-chave: Reconhecimento de Entidades Nomeadas, PLN, Modelos Pré-treinados, Fine-tuning

Keywords: Named Entity Recognition, NLP, Pre-trained Models, Fine-tuning

Recebido/Received: 30 May 2025 • **Aceito/Accepted:** 23 June 2025 • **Publicado/Published:** 11 July 2025

1 Introdução

O Reconhecimento de Entidades Nomeadas (REN), exemplificado na Figura 1, é uma tarefa essencial no Processamento de Linguagem Natural (PLN), voltada à identificação automática de menções a entidades textuais, como pessoas, locais, organizações, entre outras. Desse modo, entende-se que REN é uma peça-chave para várias aplicações, incluindo sistemas de busca, análise de sentimentos e sumarização. Apesar da sua relevância, os avanços mais expressivos na área concentram-se na língua inglesa, onde há ampla disponibilidade de dados anotados e modelos específicos Pereira [2021]; de Almeida Neto and de Melo [2023]; Pires *et al*. [2019].

No contexto da língua portuguesa, ainda há uma notável carência de recursos linguísticos, especialmente em domínios técnicos e especializados. A escassez de *corpora* anotados e de modelos ajustados ao idioma limita a criação de soluções robustas, o que força o uso de alternativas, como tradução automática, que nem sempre capturam adequadamente as nuances morfossintáticas e culturais do português Souza *et al*. [2020]. Essa limitação representa um obstáculo para a ampliação do PLN na língua e reforça a necessidade de investigar abordagens que possibilitem adaptar modelos modernos a contextos com poucos recursos Hedderich *et al*. [2021].

Este trabalho investiga o *fine-tuning* de variantes BERT e RoBERTa aplicadas à REN em português, considerando diferentes tamanhos de modelo (base, destilado e largo), tempos de treinamento, composição multilíngue e capitalização. Os experimentos foram conduzidos sobre os conjuntos Harem Santos *et al*. [2006], LeNER-Br de Araujo *et al*. [2018] e GeoCorpus Amaral [2017], com análise comparativa dos resultados frente às *baselines* propostas. Os modelos XLM-RoBERTa-large e BERT-large-cased obtiveram os melhores F1-scores, os quais superaram as abordagens propostas nos trabalhos que contêm as *baselines*. Os resultados também evidenciaram que os modelos multilíngues apresentaram maior adaptabilidade a diferentes domínios, e que os modelos destilados mantiveram desempenho competitivo em classes frequentes, mesmo com menor custo computacional.

A análise mostra que modelos largos capturam melhor nuances contextuais, enquanto os multilíngues demonstram maior adaptabilidade a diferentes domínios. Modelos destilados, por sua vez, mantêm desempenho competitivo em classes frequentes, sendo vantajosos em cenários com res-

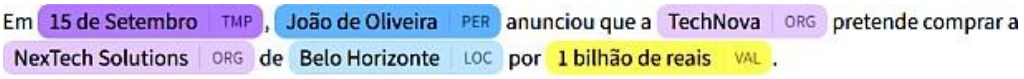

Figura 1. Exemplo de Tarefa de REN.

trição de recursos computacionais. A investigação também destaca o impacto da capitalização, com modelos *cased* superando versões *uncased* em entidades sensíveis à grafia. Por fim, este trabalho busca oferecer subsídios para pesquisas futuras que explorem o *fine-tuning* de modelos pré-treinados como alternativa viável frente à escassez de recursos para o português.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia. A Seção 4 detalha os experimentos e resultados. Por fim, a Seção 5 apresenta as conclusões e direções futuras.

2 Trabalhos Relacionados

Diversos estudos recentes têm investigado o REN em português em diferentes contextos. Modelos BERT foram aplicados a textos legislativos Albuquerque *et al*. [2022], enquanto documentos históricos foram explorados com Transformers Santos *et al*. [2024]. Um *corpus* voltado ao mercado financeiro foi desenvolvido em Zerbinati *et al*. [2024], e o BERTimbau Souza *et al*. [2020] foi utilizado para REN em textos médicos históricos Zilio *et al*. [2024]. Ademais, o contexto acadêmico foi abordado em Matos *et al*. [2024]. Embora relevantes, esses trabalhos geralmente focam em um único domínio ou modelo. Este estudo amplia essa perspectiva ao comparar sistematicamente variantes base, largas e destiladas de BERT e RoBERTa, em versões multilíngues e monolíngues, oferecendo uma visão mais abrangente sobre a aplicação de modelos pré-treinados em REN no português ajustados a diferentes *datasets*.

Como *baselines*, destacam-se a combinação de Word2Vec com BERT e BiLSTM-CRF aplicada aos conjuntos Harem e LeNER-Br da Silva and de Oliveira [2022], bem como o *fine-tuning* do modelo BERTimbau ao *dataset* GeoCorpus Rodrigues *et al*. [2022]. Estas abordagens são utilizadas como referência para avaliar o desempenho dos modelos ajustados neste trabalho.

3 Metodologia

Esta seção apresenta os procedimentos metodológicos adotados no estudo, desde a preparação dos dados até a avaliação dos modelos. As etapas detalhadas a seguir incluem: Conjunto de Dados, Pré-processamento, *Fine-tuning* dos Modelos e Métricas de Avaliação.

3.1 Conjuntos de Dados

Para esta pesquisa, foram utilizados três conjuntos de dados em português, previamente empregados em estudos de REN, devido ao seu grande volume e diversidade temática, garantindo sua adequação aos objetivos deste estudo. Os *datasets* são apresentados com mais detalhes a seguir.

3.1.1 Harem

O Harem[1] é um *corpus* em português com cerca de 93 mil palavras, distribuídas em 129 textos de diferentes gêneros literários e anotado com 10 classes de Entidades Nomeadas (EN), conforme a Tabela 1 Santos *et al*. [2006]. A divisão dos dados segue a proposta em Souza *et al*. [2020], com 4.505 exemplos de treino (55%), 237 de validação (3%) e 3.393 de teste (42%). O conjunto apresenta forte desbalanceamento, tornando a tarefa mais desafiadora.

3.1.2 LeNER-Br

O LeNER-Br[2] é um conjunto de dados em português destinado à tarefa de REN em documentos jurídicos. Composto por textos de legislação e casos jurídicos, possui 70 documentos legais de tribunais brasileiros, como o Supremo Tribunal Federal (STF) e Tribunal de Contas da União (TCU). A divisão dos dados, conforme disposto na Tabela 2, respeitou a partição definida pelos proponentes desse conjunto de dados, com 7.828 exemplos para treino (75%), 1.177 para validação (11%) e 1.390 para teste (14%) de Araujo *et al*. [2018].

3.1.3 GeoCorpus

O GeoCorpus[3] é um *corpus* aberto de textos técnicos de geociências — incluindo artigos, dissertações, teses e relatórios — de uma empresa multinacional de petróleo, focado nas Bacias Sedimentares Brasileiras Amaral [2017]. O *dataset* contém mais de 6 mil sentenças e cerca de 5,5 mil EN em 14 classes Rodrigues *et al*. [2022]. Os dados são divididos em 4.404 exemplos de treino (74%), 328 de validação (5%) e 1.253 de teste (21%). A tabela 3 ilustra as especificações do *dataset*.

3.2 Pré-Processamento

O pré-processamento envolveu a tokenização e codificação dos dados conforme o modelo: AutoTokenizer foi usado para variantes BERT, enquanto o RobertaTokenizerFast, mais eficiente para tokenização *Byte Pair Encoding* (BPE), foi aplicado aos modelos RoBERTa Wang *et al*. [2022]. As anotações seguiram o esquema BIO, marcando início (B), continuação (I) ou ausência (O) de entidades. *Tokens* especiais foram ignorados no cálculo da perda ao receberem o rótulo -100. Os *tokens* foram vetorizados com *embeddings* contextuais dos modelos pré-treinados, adaptados ao português via *fine-tuning*, permitindo ao modelo capturar relações entre *tokens* e entidades.

3.3 *Fine-Tuning* dos Modelos

Foram ajustados modelos, obtidos no *Hugging Face*[4] para os três conjuntos (Harem, LeNER-Br e GeoCorpus), incluindo variantes base — XLM-RoBERTa-base (XLMR-B), BERT-base-multilingual-uncased (BERT-BMU) e *cased* (BERT-BMC); destiladas — DistilRoBERTa (DistilR), DistilBERT-

[1] https://github.com/jneto04/ner-pt
[2] https://huggingface.co/datasets/peluz/lener_br
[3] https://github.com/jneto04/geocorpus
[4] https://huggingface.co/

Tabela 1. Especificações por classe do conjunto de dados Harem.

Classe	Descrição	Exemplos	Treino	Validação	Teste
ABS	Abstração	Justiça, Felicidade	698	37	613
ACO	Acontecimento	Guerra, Conferência	403	15	205
COI	Coisa	Livro, Carro	162	16	241
LOC	Localização	Brasil, Rio de Janeiro	1.790	95	1261
OBR	Obra	Dom Quixote, Bíblia	638	74	542
ORG	Organização	ONU, Google	1.580	116	1.212
OTR	Outro	Eclipse, Cometa	59	8	31
PER	Pessoa	João, Maria	1.814	87	1.534
TMP	Tempo	2023, Segunda-feira	708	52	594
VAL	Valor	100 dólares, 50 reais	867	43	599

Tabela 2. Especificações por classe do conjunto de dados *LeNER*-Br.

Classe	Descrição	Exemplos	Treino	Validação	Teste
JURISPRUDENCIA	Jurisprudência	Súmula 11	3.990	766	683
LEGISLACAO	Legislação	Lei Maria da Penha	13.058	2.628	2.688
LOCAL	Localização	São Paulo, Brasil	1.432	259	147
ORGANIZACAO	Organização	Supremo Tribunal Federal	6.674	1.611	1.370
PESSOA	Pessoa	João Silva	4.619	901	742
TEMPO	Tempo	15 de agosto de 2023	2.354	554	271

Tabela 3. Especificações por classe do conjunto de dados GeoCorpus.

Classe	Descrição	Exemplo	Treino	Validação	Teste
EON	Eon	Fanerozoico	245	59	99
ERA	Era	Paleozoica	275	55	106
IDA	Idade	Asseliano	590	56	180
OTR	Outro	oncoides	865	56	241
PRD	Período	Rhyaciano	502	65	132
bacSED	Bacia Sedimentar	Bacia de Irecê	475	51	170
ctxGBAC	Contexto Geológico de Bacia	sistema tafrogênico	470	64	146
sedCARB	Rocha Sedimentar Carbonática	calcarenitos	264	61	109
sedORGN	Rocha Sedimentar Orgânica	Carvão	62	49	52
sedQUIM	Rocha Sedimentar Química	sucessões evaporíticas	58	53	53
sedSLCT	Rocha Sedimentar Siliciclástica	diamictito	737	65	182
uniESTG	Unidade Litoestratigráfica	Supergrupo Espinhaço	945	88	254

uncased (DistilBU) e DistilBERT-multilingual-cased (DistilBMC); e largas — BERT-large-cased (BERT-LC), *uncased* (BERT-LU) e XLM-RoBERTa-large (XLMR-L). Buscaram-se versões largas e destiladas equivalentes aos modelos base, mas nem todas estão disponíveis como multilíngues. Os modelos multilíngues usados foram XLMR-B, XLMR-L, BERT-BMU, BERT-BMC e DistilBMC.

O *fine-tuning* foi realizado no Google Colab (GPU T4), com Transformers 4.41.1 e PyTorch 2.1.2, taxa de aprendizagem 2e-5, 10 épocas, otimizador Adam e escalonamento linear. Usaram-se lotes de 16 para modelos base/destilados e 2 para os largos, devido à limitação de memória.

3.4 Métricas de Avaliação

O desempenho dos modelos foi avaliado com as métricas de precisão (P), revocação (R) e F1-Score (F1), calculadas segundo o padrão CoNLL Tjong Kim Sang and De Meulder [2003], amplamente adotado em tarefas de REN. O F1-Score foi obtido no modo micro, em que P e R são calculadas globalmente para todas as classes antes da média F1, garantindo uma avaliação uniforme e comparável. Considerando A como o conjunto de entidades corretamente reconhecidas e B como o conjunto de entidades identificadas pelo modelo, as métricas são definidas pelas Equações (1)-(3).

$$P = \frac{|A \cap B|}{|B|}, \quad (1)$$

$$R = \frac{|A \cap B|}{|A|}, \quad (2)$$

$$F_1 = \frac{2 \times (P \times R)}{P + R} \quad (3)$$

4 Experimentos

Esta seção apresenta os resultados do *fine-tuning* dos modelos para REN em português. A análise é dividida em três partes: (i) comparação dos modelos quanto ao tamanho, capacidade multilíngue, tempo de treinamento e capitalização; (ii) avaliação por classe de entidade em cada conjunto de dados; e (iii) comparação dos melhores modelos com as *baselines* adotadas.

4.1 Análise Geral

A Tabela 4 apresenta os resultados globais após o *fine-tuning*, incluindo os tempos de treinamento em minutos. Observa-se

que modelos largos, como o XLMR-L e o BERT-LC, alcançaram os melhores F1-scores nos três conjuntos, com maiores custos computacionais. Modelos destilados, como o DistilBMC, apresentaram tempos significativamente menores e desempenho competitivo, especialmente no LeNER-Br. Esse padrão reforça a busca por um equilíbrio entre eficiência e desempenho em modelos compactos Sanh *et al*. [2019]; Jiao *et al*. [2020].

Tabela 4. Resultados gerais após *fine-tuning*.

Corpus	Modelo	Precisão	Revocação	F1-Score	Tempo
Harem	XLMR-B	79,80%	81,44%	80,61%	12:56
	BERT-BMU	78,33%	75,89%	77,09%	9:36
	BERT-BMC	78,78%	80,08%	79,42%	22:34
	DistilR	65,76%	68,51%	67,11%	3:56
	DistilBU	65,56%	63,24%	64,38%	3:04
	DistilBMC	74,55%	77,80%	76,14%	2:57
	XLMR-L	**83,44%**	**84,12%**	**83,78%**	106:48
	BERT-LU	68,95%	64,42%	66,61%	25:09
	BERT-LC	72,41%	74,23%	73,31%	56:55
LeNER-Br	XLMR-B	73,97%	92,12%	82,05%	12:43
	BERT-BMU	86,78%	87,58%	87,18%	29:03
	BERT-BMC	84,57%	84,75%	84,66%	31:58
	DistilR	80,13%	84,30%	82,16%	6:19
	DistilBU	78,46%	78,10%	78,28%	14:29
	DistilBMC	79,60%	78,48%	79,03%	15:29
	XLMR-L	**91,66%**	**92,89%**	**92,27%**	38:43
	BERT-LU	81,95%	85,34%	83,61%	24:31
	BERT-LC	80,46%	82,98%	81,70%	166:48
GeoCorpus	XLMR-B	83,57%	80,23%	81,87%	22:56
	BERT-BMU	80,40%	78,63%	79,51%	15:03
	BERT-BMC	80,46%	86,81%	83,52%	7:16
	DistilR	78,53%	84,41%	81,36%	6:04
	DistilBU	77,16%	84,55%	80,69%	5:25
	DistilBMC	80,79%	88,68%	84,55%	5:11
	XLMR-L	84,46%	89,70%	87,00%	57:56
	BERT-LU	81,71%	88,06%	84,76%	82:48
	BERT-LC	**86,80%**	**88,72%**	**87,75%**	58:17

Comparando os modelos, o modelo XLMR-L superou o melhor destilado em 10% no Harem e 12,3% no LeNER-Br. No GeoCorpus, o BERT-LC teve um ganho de 3,2% sobre o DistilBMC, reforçando a superioridade dos modelos largos em domínios técnicos Devlin *et al*. [2019]; Liu *et al*. [2019]. Modelos base, como o XLMR-B e BERT-BMC, entregam bom custo-benefício, enquanto os destilados são adequados para restrições computacionais Sanh *et al*. [2019].

Modelos multilíngues (XLMR-B, XLMR-L, BERT-BMU, BERT-BMC, DistilBMC) superaram os monolíngues nos três corpora: Harem (77,9% vs. 75,4%), LeNER-Br (83,9% vs. 79,1%) e GeoCorpus (83,3% vs. 82,7%). Essa vantagem evidencia sua adaptabilidade a diferentes domínios e estilos de texto, sendo resultado do pré-treinamento com dados de múltiplos idiomas. Mesmo sem especialização, mostram-se robustos em tarefas de REN em português, sendo uma alternativa eficiente diante da escassez de recursos anotados.

A comparação entre modelos *cased* e *uncased* revelou desempenho superior dos primeiros, sobretudo em *corpora* formais. O BERT-BMC superou o BERT-BMU em todos os conjuntos, com até 4,0% de diferença. O mesmo ocorreu entre os modelos largos, com vantagem de até 2,7% para o BERT-LC. Esses resultados reforçam a importância da capitalização para distinguir EN, especialmente nomes próprios.

4.2 Análise Específica

As tabelas desta seção apresentam os resultados por classe de entidade nomeada para os modelos avaliados, organizados nas versões base, larga e destilada. Para facilitar a leitura dos comentários, utilizam-se as seguintes abreviações: XLMR-B, XLMR-L e DistilR referem-se às versões base, larga e destilada do XLM-RoBERTa; BERT-BMU, BERT-LU e DistilBU às versões do BERT multilíngue *uncased*; e BERT-BMC, BERT-LC e DistilBMC às versões do BERT multilíngue *cased*.

4.2.1 Harem

A Tabela 5 mostra que o XLMR-L superou a versão base em PER, LOC e principalmente OBR (+78%). O BERT-LC teve melhor desempenho que o *uncased* em ORG (+12,8%) e OBR (+18,3%), destacando a importância da capitalização. Em classes desafiadoras, como OTR e ACO, o XLMR-L apresentou um ganho de 40,4%, enquanto o DistilBMC se destacou em ACO (+54,8% em relação ao BERT-BMC), mostrando que modelos expostos a múltiplos domínios podem superar limitações arquiteturais.

4.2.2 LeNER-Br

A Tabela 6 revela que as classes PESSOA e TEMPO tiveram os melhores desempenhos, destacando-se o XLMR-B em PESSOA (+3,6%) e DistilBU em TEMPO (+0,3%) em relação ao BERT-LC. O XLMR-L apresentou ganhos significativos em JURISPRUDÊNCIA (+62,6%) e LOCAL (+26,6%) em comparação aos destilados. Em ORGANIZAÇÃO, a capitalização beneficiou o BERT-LC, com 4,7% de vantagem sobre a versão *uncased*. Os resultados sugerem que modelos maiores lidam melhor com padrões jurídicos, enquanto os destilados enfrentam mais dificuldade nesses domínios.

4.2.3 GeoCorpus

Na Tabela 7, sedORGN e EPC se destacaram como as classes com melhor desempenho. O BERT-LC superou o XLMR-B em EPC (+2,6%) e ERA (+8,0%). Em ctxGBAC e sedCARB, os modelos largos também obtiveram ganhos relevantes, chegando a 5,5%. Já os destilados se aproximaram dos modelos base em classes recorrentes, como EPC (–0,8%), mas registraram quedas mais acentuadas em categorias técnicas, como ERA (–3,7%) e ctxGBAC (–11,3%), evidenciando a superioridade dos modelos largos no tratamento de vocabulário especializado.

4.3 Análise Comparativa às *Baselines*

A Tabela 8 apresenta os melhores modelos ajustados em comparação com as *baselines* de cada conjunto. No Harem, o XLMR-L superou a *baseline* com Word2Vec e $\text{BERT}_{\text{large}}$da Silva and de Oliveira [2022] em 5,4% no F1-score, destacando sua capacidade de lidar com textos diversos. No GeoCorpus, o BERT-LC obteve 4,9% a mais que o BERTimbau, mesmo sem ser especializado no domínio, reforçando a eficácia dos modelos largos Beltagy *et al*. [2019]. No LeNER-Br, o XLMR-L apresentou desempenho praticamente equivalente ao da *baseline*, evidenciando a robustez dos modelos multilíngues Conneau *et al*. [2020]; Wolf *et al*. [2020] mesmo em contextos jurídicos.

Tabela 5. Pontuações F1 para modelos base, largos e destilados no Harem.

Classe	XLM-RoBERTa			BERT Uncased			BERT Cased		
	Base	Largo	Destilado	Base	Largo	Destilado	Base	Largo	Destilado
ABS	75,0	**80,41**	56,34	62,5	58,57	58,17	80,0	51,61	58,92
ACO	63,6	66,67	0,00	**92,3**	41,18	30,45	52,2	21,05	80,78
COI	89,4	81,48	65,23	71,4	75,00	54,89	80,9	68,57	79,31
LOC	83,7	**84,52**	78,56	77,7	78,64	68,22	82,4	67,19	79,67
OBR	46,4	**82,61**	48,12	64,7	50,63	49,78	66,7	60,00	67,45
ORG	83,8	82,98	64,37	76,6	78,77	67,19	79,2	63,00	75,83
OTR	11,8	**52,17**	0,00	33,3	25,00	0,00	27,3	11,11	12,56
PER	90,4	**91,49**	74,82	87,2	80,17	67,34	88,5	76,19	81,12
TMP	95,4	93,94	87,45	95,4	85,71	75,23	90,6	87,80	92,67
VAL	69,2	72,73	69,34	**89,2**	76,32	73,56	86,6	82,86	87,12
Média Geral	80,6	**83,78**	67,12	77,0	66,61	64,38	80,0	73,31	76,14

Tabela 6. Pontuações F1 para Modelos Base, Largos e Destilados no LeNER-Br.

Classe	XLM-RoBERTa			BERT Uncased			BERT Cased		
	Base	Largo	Destilado	Base	Largo	Destilado	Base	Largo	Destilado
JURISPRUDÊNCIA	50,1	**81,47**	72,46	66,0	67,16	55,19	59,0	61,03	51,74
LEGISLAÇÃO	86,6	**87,73**	78,25	84,6	80,78	78,11	82,8	83,36	76,42
LOCAL	81,8	**85,54**	67,57	79,2	64,59	65,30	74,9	74,69	61,74
ORGANIZAÇÃO	88,3	**89,54**	82,82	87,2	80,36	77,38	85,3	84,08	79,10
PESSOA	**99,6**	95,80	83,07	98,1	91,61	86,17	93,1	96,68	91,55
TEMPO	96,5	**98,06**	96,11	97,0	95,18	96,39	96,3	96,53	94,57
Média Geral	82,0	**92,27**	82,16	87,1	83,61	79,03	84,6	81,7	79,03

Tabela 7. Pontuações F1 para Modelos Base, Largos e Destilados no GeoCorpus.

Classe	XLM-RoBERTa			BERT Uncased			BERT Cased		
	Base	Largo	Destilado	Base	Largo	Destilado	Base	Largo	Destilado
EON	91,3	**92,8**	88,1	91,4	88,0	92,3	90,3	91,6	91,0
EPC	95,8	96,2	97,0	94,9	97,9	94,5	94,3	**98,3**	93,3
ERA	80,8	82,4	79,8	82,5	85,7	84,0	82,7	**87,3**	81,1
IDA	91,9	93,0	94,3	91,8	94,6	92,5	92,1	**95,5**	92,9
OTR	89,7	**90,6**	57,3	89,9	62,2	**90,6**	64,3	90,1	89,5
PRD	92,1	92,6	90,5	90,8	**93,5**	91,5	90,3	93,4	90,4
bacSED	74,5	75,2	70,8	76,9	65,5	**77,9**	74,5	73,3	75,2
ctxGBAC	68,2	**69,0**	60,5	61,9	64,9	62,5	60,6	66,0	61,3
sedCARB	75,6	76,3	78,0	81,8	**86,3**	82,5	72,4	79,7	73,0
sedORGN	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
sedQUIM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
sedSLCT	86,6	87,2	89,4	85,8	89,7	86,5	92,7	90,7	**93,5**
uniESTG	85,4	**86,0**	79,3	77,3	82,6	78,5	78,3	83,9	78,8
Média Geral	81,87	87,0	81,36	79,51	85,0	78,28	83,4	**87,75**	84,55

Tabela 8. Comparação entre os melhores modelos ajustados e as *baselines*.

Dataset	Melhor Modelo	*Baseline*	F1-Score (%)
Harem	XLMR-L	Word2Vec + $\text{BERT}_{\text{large}}$	83,8 vs 78,4
LeNER-Br	XLMR-L	Word2Vec + $\text{BERT}_{\text{large}}$	92,27 vs 92,7
GeoCorpus	BERT-LC	BERTimbau	87,8 vs 82,9

5 Considerações Finais

Este trabalho investigou o *fine-tuning* de modelos BERT e RoBERTa aplicados ao Reconhecimento de Entidades Nomeadas (REN) em português, avaliando variantes base, largas e destiladas, nas versões multilíngues e monolíngues. Os experimentos com os conjuntos Harem, LeNER-Br e GeoCorpus mostraram que modelos largos, como o XLM-RoBERTa-large e o BERT-large-cased, alcançaram os melhores resultados, superando as *baselines* tradicionais em até 5% de F1-score. O XLM-RoBERTa-large atingiu 83,8% no Harem e 92,3% no LeNER-Br, enquanto o BERT-large-cased obteve 87,8% no GeoCorpus.

Modelos multilíngues apresentaram melhor desempenho que os monolíngues nos três conjuntos: Harem (77,9% vs. 75,4%), LeNER-Br (83,9% vs. 79,1%) e GeoCorpus (83,3% vs. 82,7%), demonstrando maior adaptabilidade a diferentes domínios. Modelos destilados mantiveram desempenho competitivo com tempos de treinamento reduzidos, sendo viáveis em cenários com restrições computacionais. A capitalização também influenciou os resultados, com os modelos *cased* apresentando ganhos de até 4%.

Como trabalhos futuros, pretende-se investigar modelos alternativos de grande escala, como Qwen e DeepSeek, avaliando seu desempenho em tarefas de REN em português por meio de técnicas de *fine-tuning* e abordagens *zero-shot*. Além disso, considera-se a aplicação desses modelos em novos domínios, como os contextos clínico e científico, com o objetivo de ampliar a cobertura temática e analisar sua efetividade em diferentes cenários de uso.

Declarações complementares

Contribuições dos autores

Guilherme Tapajós foi responsável pela pesquisa de conjuntos de dados em língua portuguesa e de modelos *open-source*, bem como pelo *fine-tuning* destes e pela redação do manuscrito. Tiago de Melo foi responsável pela concepção da proposta do trabalho e da trilha de elaboração deste. Além disso, supervisionou a pesquisa e forneceu orientações. Elloá Guedes e Fábio Santos atuaram em coorientação, revisão, edição e, baseados nos resultados trazidos, em discussões de incrementos ao artigo. Todos os autores revisaram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

O script utilizado para o fine-tuning dos modelos está disponível em: `https://www.kaggle.com/code/guilhermetapajs/fine-tuning-ner-trabalho`. Além disso, todos os modelos treinados podem ser acessados e baixados por meio da página do Hugging Face: `https://huggingface.co/GuiTap`.

Agradecimentos

Os autores agradecem o apoio material concedido pelo Laboratório de Sistemas Inteligentes (LSI) da Universidade do Estado do Amazonas (UEA). EBG agradece o apoio financeiro da CAPES por meio do Projeto COFECUB-PJ3126170P.

Referências

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language*. Springer, Cham. DOI: 10.1007/978-3-030-98305-5_1.

Amaral, D. O. F. (2017). *Reconhecimento de entidades nomeadas na área da Geologia: bacias sedimentares brasileiras*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil. Disponível em: `https://tede2.pucrs.br/tede2/handle/tede/8035`.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. DOI: 10.18653/v1/D19-1371.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Transactions of the Association for Computational Linguistics*, 8:63–77. Disponível em: `https://arxiv.org/abs/1911.02116`.

da Silva, M. G. and de Oliveira, H. T. A. (2022). Combining word embeddings for portuguese named entity recognition. In *Computational Processing of the Portuguese Language*, pages 198–208. Springer, Cham. DOI: 10.1007/978-3-030-98305-5_19.

de Almeida Neto, J. A. and de Melo, T. (2023). Exploring supervised learning models for multi-label text classification in brazilian restaurant reviews. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 126–140. SBC. DOI: 10.5753/eniac.2023.233843.

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language*, pages 313–323. Springer, Cham. DOI: 10.1007/978-3-319-99722-3_32.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 1:4171–4186. DOI: 10.18653/v1/N19-1423.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*. Disponível em: `https://arxiv.org/abs/2010.12309`.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2020). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*. Disponível em: `https://arxiv.org/abs/1909.10351`.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. DOI: 10.48550/arXiv.1907.11692.

Matos, E., Rodrigues, M., and Teixeira, A. (2024). Towards the automatic creation of ner systems for new domains. In *16th International Conference on Computational Processing of Portuguese Language (PROPOR 2024)*, pages 218–227. Disponível em: `https://aclanthology.org/2024.propor-1.22/`.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115. DOI: 10.1007/s10462-020-09870-1.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *ACL*. Disponível em: `https://arxiv.org/abs/1906.01502`.

Rodrigues, R. B. M., Privatto, P. I. M., de Sousa, G. J., Murari, R. P., Afonso, L. C. S., Papa, J. P., Pedronette, D. C. G., Guilherme, I. R., Perrout, S. R., and Riente, A. F. (2022). Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In *Computational Processing of the Portuguese Language*, pages 101–109. Springer, Cham. DOI: 10.1007/978-3-030-98305-5_10.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*. Disponível em: `https://arxiv.org/abs/1910.01108`.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). Harem: An advanced ner evaluation contest for portuguese. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. Disponível em: `https://aclanthology.org/L06-1027/`.

Santos, J., Vieira, R., Olival, F., Cameron, H., and Farrica, F. (2024). Named entity recognition specialised for portuguese 18th-century history research. In *16th International Conference on Computational Processing of Portuguese Language (PROPOR 2024)*, pages 117–126. Disponível em: `https://dspace.uevora.pt/rdpc/handle/10174/36583`.

Souza, F. C., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417. Springer, Cham. DOI: 10.1007/978-3-030-61377-8_28.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. Disponível em: `https://www.aclweb.org/anthology/W03-0419`.

Wang, Y., Tong, H., Zhu, Z., and Li, Y. (2022). Nested named entity recognition: A survey. *ACM Transactions on Knowledge Discovery from Data*. DOI: 10.1145/3522593.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Disponível em: `https://aclanthology.org/2020.emnlp-demos.6/`.

Zerbinati, M. M., Roman, N. T., and Di Felippo, A. (2024). A corpus of stock market tweets annotated with named entities. In *16th International Conference on Computational Processing of Portuguese Language (PROPOR 2024)*, pages 276–284. Disponível em: `https://aclanthology.org/2024.propor-1.28/`.

Zilio, L., Lazzari, R. R., and Finatto, M. J. B. (2024). Nlp for historical portuguese: Analysing 18th-century medical texts. In *16th International Conference on Computational Processing of Portuguese Language (PROPOR 2024)*, pages 76–85. Disponível em: `https://aclanthology.org/2024.propor-1.8/`.