MLMs – Masked Language Models:

Modelos de Linguagem Mascarada

Bem-vindos a esta apresentação abrangente sobre Modelos de Linguagem Mascarada (MLMs), uma estrutura fundamental no Processamento de Linguagem Natural moderno. Exploraremos desde os conceitos básicos até aplicações avançadas, com foco especial nas pesquisas desenvolvidas em universidades brasileiras.

Introdução aos Modelos de Linguagem Mascarada (MLMs)

Os Modelos de Linguagem Mascarada (MLMs) representam uma abordagem revolucionária no Processamento de Linguagem Natural (PLN), caracterizando-se como modelos pré-treinados que aprendem representações contextuais bidirecionais a partir de texto não rotulado. Diferentemente dos modelos tradicionais que processam o texto sequencialmente, os MLMs consideram simultaneamente o contexto à esquerda e à direita de cada palavra.

Esta capacidade de compreensão bidirecional permite uma representação mais rica e completa do significado linguístico, resultando em desempenho superior em diversas tarefas de PLN.

Relevância para Automação e IA

- Formam a base de sistemas modernos de compreensão de linguagem
- Permitem avanços significativos em chatbots e assistentes virtuais
- Facilitam análise semântica em grande escala
- Viabilizam aplicações de extração de informações e classificação automática
- Fundamentam sistemas de respostas a perguntas e tradução automática

Os MLMs transformaram o panorama da IA linguística, estabelecendo novos patamares de desempenho e possibilitando aplicações anteriormente inviáveis.

Panorama da PLN no Brasil

Universidade de São Paulo (USP)

O Núcleo Interinstitucional de Linguística Computacional (NILC) da USP destaca-se pelo desenvolvimento do BERTimbau, um dos principais modelos pré-treinados para português brasileiro. Pesquisadores como Marcelo Finger e Sandra Aluísio lideram estudos em representação linguística, corpora e aplicações práticas de MLMs.

Universidade Federal de Minas Gerais (UFMG)

No Laboratório MINDS da UFMG, pesquisadores como Adriano Veloso e Wagner Meira Jr. focam-se em redes neurais para PLN, análise de sentimentos e detecção de desinformação utilizando MLMs. A universidade mantém colaborações com empresas de tecnologia para aplicações práticas.

Universidade Estadual de Campinas (UNICAMP)

O Instituto de Computação da UNICAMP, sob liderança de pesquisadores como Ariadne Carvalho e Anderson Rocha, desenvolve pesquisas em modelos multimodais, processamento semântico e aplicações de MLMs para análise de texto em português, com ênfase em aspectos computacionais e linguísticos.

O Brasil apresenta um ecossistema rico de pesquisa em PLN, com universidades federais e estaduais contribuindo significativamente para avanços no processamento da língua portuguesa. Estas instituições não apenas adaptam modelos internacionais para o contexto brasileiro, mas também desenvolvem abordagens inovadoras específicas para as particularidades linguísticas do português.

Evolução dos Modelos de Linguagem

1

1950-1990: Lógica Formal e Regras

Os primeiros sistemas de PLN baseavam-se em regras gramaticais explícitas e estruturas lógicas definidas manualmente. Estes sistemas, embora pioneiros, apresentavam limitações significativas na captura de nuances linguísticas.

2

1990-2010: Modelos Estatísticos

Surgimento de abordagens probabilísticas como n-gramas, modelos de Markov e classificadores estatísticos. Estas técnicas permitiram maior flexibilidade e robustez, mas ainda eram limitadas em termos de captura semântica.

3

2013: Word2Vec

Revolução com embeddings de palavras através do Word2Vec (Mikolov et al.), que permitiu representar palavras como vetores densos, capturando relações semânticas no espaço vetorial.

4

2014-2017: Embeddings Contextuais

ELMo e ULMFiT introduziram representações contextuais, onde o significado de uma palavra varia conforme o contexto em que aparece, uma evolução significativa em relação aos embeddings estáticos.

5

2017: Transformers

A arquitetura Transformer (Vaswani et al.) revolucionou o PLN com mecanismos de atenção, estabelecendo a base para os modelos de linguagem modernos.

6

2018-Presente: Era dos MLMs

BERT inaugura a era dos MLMs, seguido por RoBERTa, ALBERT, T5 e variantes multilíngues, estabelecendo novos paradigmas para compreensão linguística em PLN.

Esta evolução representa uma transição de abordagens baseadas em regras para sistemas de aprendizagem profunda, com capacidade crescente de capturar nuances semânticas e contextuais da linguagem humana. Os pesquisadores brasileiros têm acompanhado ativamente esta evolução, contribuindo com adaptações e inovações para o português.

O que são os MLMs?

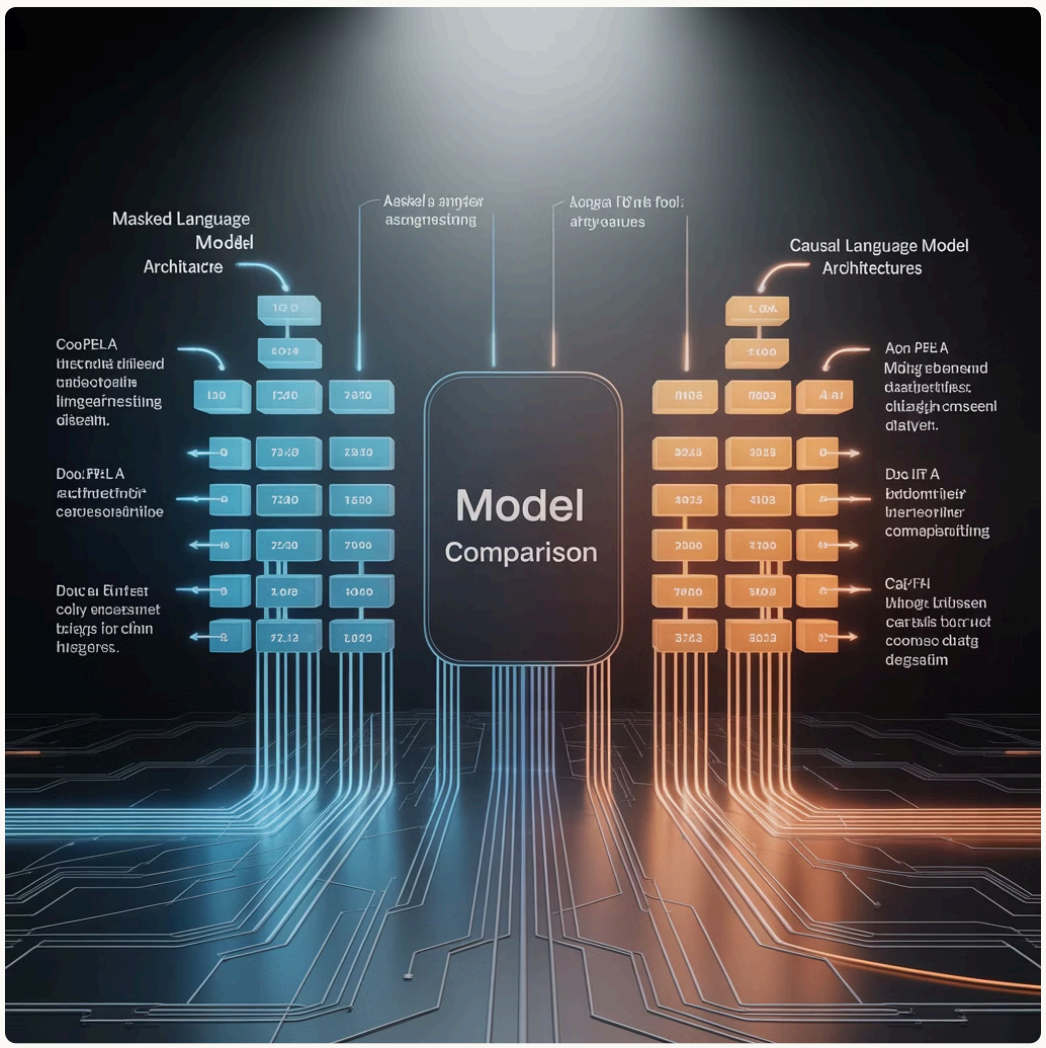
Conceito de Mascaramento de Tokens

Os Modelos de Linguagem Mascarada (MLMs) fundamentam-se na ideia de mascaramento aleatório de tokens em um texto durante o treinamento. O modelo é então treinado para prever estes tokens mascarados com base no contexto circundante. Este processo é conhecido como "masked language modeling" ou modelagem de linguagem mascarada.

Por exemplo, na frase "O gato [MASK] no tapete", o modelo deve utilizar o contexto para prever a palavra mascarada "dorme". Esta abordagem força o modelo a desenvolver uma compreensão profunda do contexto linguístico e das relações semânticas entre palavras.

A técnica permite aprendizagem bidirecional, considerando tanto o contexto à esquerda quanto à direita da palavra mascarada, resultando em representações linguísticas mais ricas e completas.

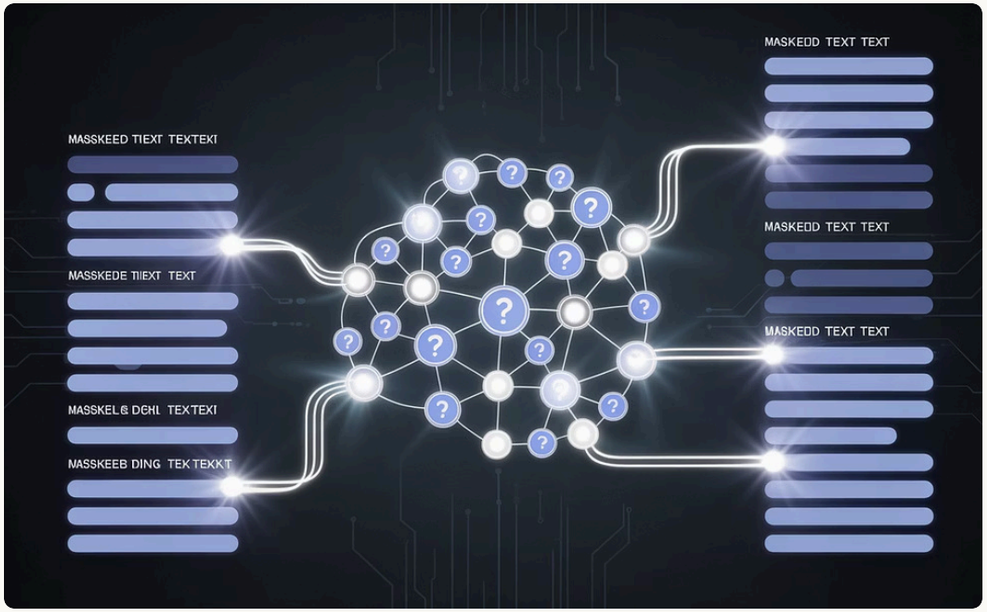
Contraste com Modelos de Linguagem Causal



Diferentemente dos MLMs, os modelos de linguagem causal (como o GPT) são treinados para prever a próxima palavra em uma sequência, considerando apenas o contexto à esquerda (palavras anteriores). Esta diferença fundamental resulta em comportamentos e aplicações distintas:

- MLMs são ideais para tarefas de compreensão (como classificação, extração de informações)
- Modelos causais destacam-se em tarefas de geração (como criação de texto, conversação)
- MLMs capturam contexto bidirecional, enquanto modelos causais focam em previsões unidirecionais

Porquê Mascarar Palavras?



Aprendizagem Auto-Supervisionada

O mascaramento permite aprendizagem auto-supervisionada, onde o modelo gera automaticamente seus próprios exemplos de treinamento a partir de texto não rotulado. Isto elimina a necessidade de dados rotulados manualmente, que são caros e demorados para produzir.

- Permite utilizar vastos corpora de texto sem anotação
- Reduz significativamente custos de preparação de dados
- Facilita o treinamento em novas línguas e domínios

Exemplo Intuitivo de Mascaramento

Na frase "A Universidade de São Paulo é uma [MASK] de excelência em pesquisa", o modelo aprende a prever "instituição" baseando-se no contexto. Este processo força o modelo a:

- Compreender relações semânticas entre palavras
- Capturar nuances sintáticas da língua
- Desenvolver representações contextuais robustas
- Adquirir conhecimento factual implícito nos textos

Pesquisadores da UFMG e USP demonstraram que modelos treinados com mascaramento apresentam desempenho superior em tarefas de compreensão de texto em português brasileiro, especialmente em contextos que exigem entendimento profundo de relações semânticas e conhecimento cultural específico.

Estudos realizados na UNICAMP e UFRJ confirmam que esta abordagem permite ao modelo capturar particularidades linguísticas do português que seriam difíceis de modelar através de regras explícitas ou abordagens tradicionais.

Fundamentos Matemáticos de MLMs

Função Objetivo: Probabilidade Condicional

A função objetivo central dos MLMs é maximizar a probabilidade condicional de tokens mascarados dado o contexto circundante. Formalmente, para uma sequência de tokens $X = [x_1, x_2, ..., x_n]$ onde alguns tokens são mascarados, o modelo aprende a estimar:

$$P(x_m|x_1, x_2, ..., x_{m-1}, x_{m+1}, ..., x_n)$$

Onde x_m representa um token mascarado. Diferentemente dos modelos causais que estimam $P(x_t|x_{<t})$, os MLMs consideram tanto o contexto à esquerda quanto à direita.

Formalização da Função de Perda

A função de perda típica é a entropia cruzada (cross-entropy) calculada apenas sobre os tokens mascarados:

$$L = -\sum_{i \in M} \log P(x_i|X_{\setminus i})$$

Onde:

- M é o conjunto de índices dos tokens mascarados
- $X_{\setminus i}$ representa a sequência completa exceto o token na posição i

Esta formulação permite que o modelo se concentre exclusivamente em prever os tokens mascarados, otimizando a capacidade de inferir informações a partir do contexto.

Pesquisas da UFMG e UNICAMP propuseram variações desta função de perda para melhor adaptação às características do português brasileiro, considerando aspectos morfológicos complexos como conjugação verbal e concordância nominal.

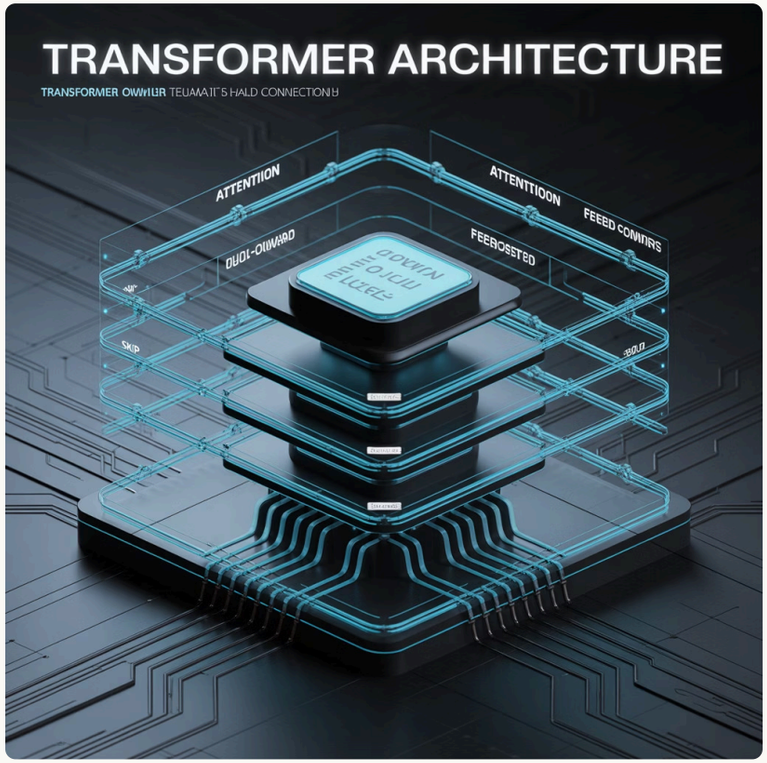
Visão Geral dos Transformers

Arquitetura Transformer

A arquitetura Transformer, introduzida por Vaswani et al. (2017), revolucionou o PLN e serve como fundamento para os MLMs. Seus componentes principais incluem:

- **Embeddings Posicionais**: Capturam a posição de cada token na sequência
- **Camadas de Atenção Multi-Cabeça**: Permitem ao modelo focar em diferentes partes do contexto simultaneamente
- **Feed-Forward Networks**: Processam representações para cada token individualmente
- **Layer Normalization**: Estabiliza o treinamento normalizando as ativações
- **Conexões Residuais**: Facilitam o treinamento de redes profundas evitando o problema do desvanecimento do gradiente

A capacidade de processamento paralelo dos Transformers, em contraste com as redes recorrentes sequenciais, permite treinamento mais eficiente e captura de dependências de longo alcance no texto.



Papel dos Transformers nos MLMs

Nos MLMs, utiliza-se tipicamente apenas a parte do codificador (encoder) da arquitetura Transformer original. Esta escolha arquitetural permite:

- Processamento bidirecional do contexto
- Representações contextuais ricas para cada token
- Captura eficiente de relações sintáticas e semânticas

Pesquisadores da USP e UFABC têm explorado adaptações da arquitetura Transformer para melhor capturar particularidades do português brasileiro, como acentuação, contrações e estruturas verbais complexas.

Mecanismo de Atenção

Autoatenção: Conceito e Cálculo

O mecanismo de atenção é o componente central dos Transformers e, consequentemente, dos MLMs. A autoatenção permite que cada token "preste atenção" a todos os outros tokens da sequência, ponderando sua importância relativa para a representação contextual.

Formalmente, a atenção é calculada como:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

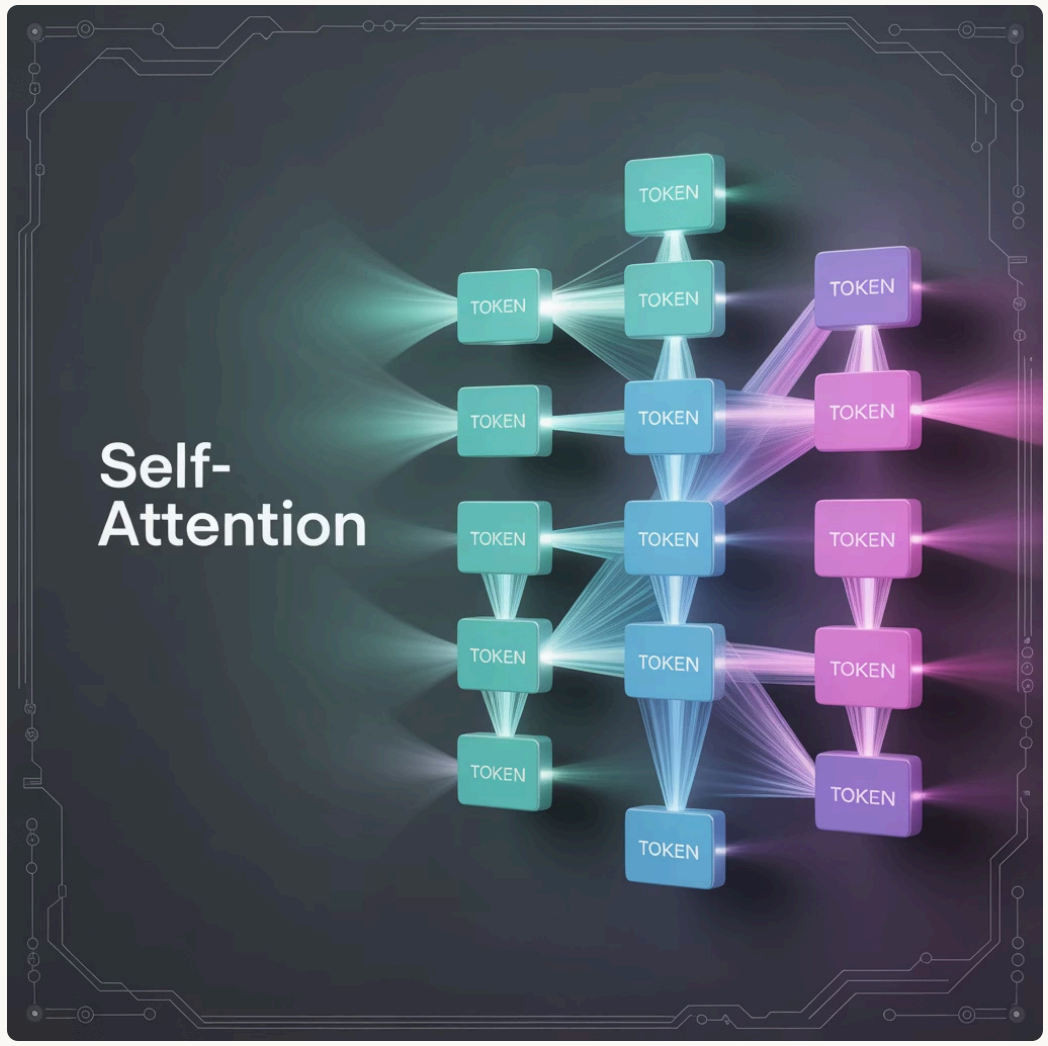
Onde Q (query), K (key) e V (value) são projeções lineares dos embeddings de entrada, e d_k é a dimensão das chaves.

A atenção multi-cabeça permite ao modelo focar em diferentes aspectos do contexto simultaneamente, através de múltiplas projeções paralelas:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

Onde cada $head_i$ é calculado com diferentes projeções lineares dos mesmos Q, K e V.

Exemplos Visuais

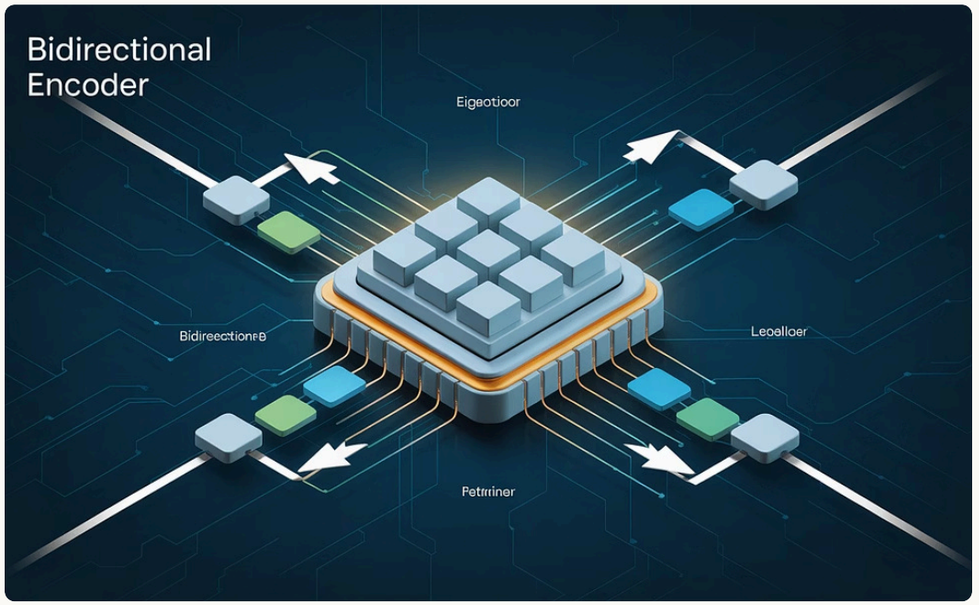


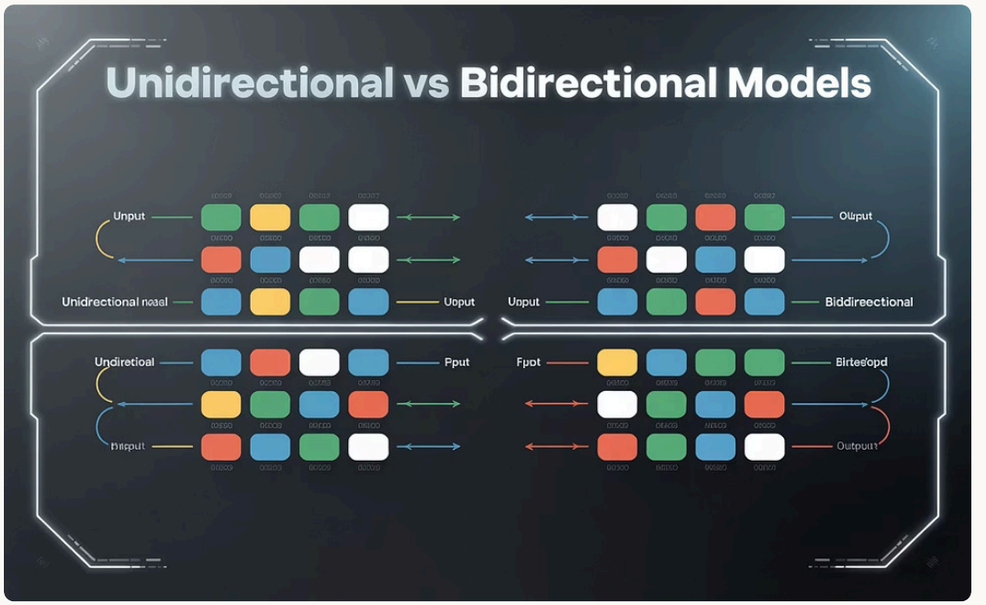
O diagrama ilustra como as palavras interagem no mecanismo de atenção. A intensidade das conexões representa o peso da atenção, mostrando como cada token influencia e é influenciado por outros tokens da sequência.

Em MLMs como BERT, esta atenção bidirecional é crucial para a compreensão contextual completa, permitindo que o modelo considere tanto o contexto precedente quanto o subsequente ao prever tokens mascarados.

Estudos conduzidos na UFRJ e UFPE demonstraram que os padrões de atenção em textos portugueses apresentam características distintas devido a estruturas sintáticas próprias da língua, como a ordem variável de sujeito-verbo-objeto e o uso frequente de pronomes clíticos.

Codificador Bidirecional nos MLMs





Contextualização Bidirecional

O codificador bidirecional permite que os MLMs considerem simultaneamente o contexto à esquerda e à direita de cada token. Esta capacidade é fundamental para capturar o significado completo de palavras ambíguas ou polissêmicas.

Por exemplo, na frase "O banco estava cheio", a palavra "banco" pode referir-se a uma instituição financeira ou a um assento. Apenas com contexto bidirecional completo o modelo pode resolver adequadamente esta ambiguidade.

Exemplos de Arquitetura Bidirecional

A bidireccionalidade é implementada através da atenção completa (não mascarada) nas camadas do Transformer, permitindo que cada token interaja diretamente com todos os outros tokens da sequência.

Estudos da USP e UNICAMP demonstraram que esta bidireccionalidade é especialmente vantajosa para o português, uma língua com ordem relativamente livre de palavras e rica em concordâncias de longa distância.

> "A capacidade de processamento bidirecional é um dos principais diferenciais dos MLMs em relação aos modelos autoregressivos, permitindo representações contextuais mais ricas e precisas." — Rodrigo Nogueira, pesquisador da UFRJ

Experimentos realizados por pesquisadores da UFMG comprovaram que, para tarefas de análise de sentimento em português brasileiro, modelos bidirecionais superam consistentemente suas contrapartes unidirecionais, com ganhos de desempenho de até 15% em datasets desafiadores.

Detalhes do Mascaramento de Tokens

Estratégia de Mascaramento Padrão

O procedimento típico de mascaramento em MLMs, exemplificado pelo BERT, envolve a seleção aleatória de 15% dos tokens em cada sequência para mascaramento. Estes tokens selecionados são então tratados da seguinte forma:

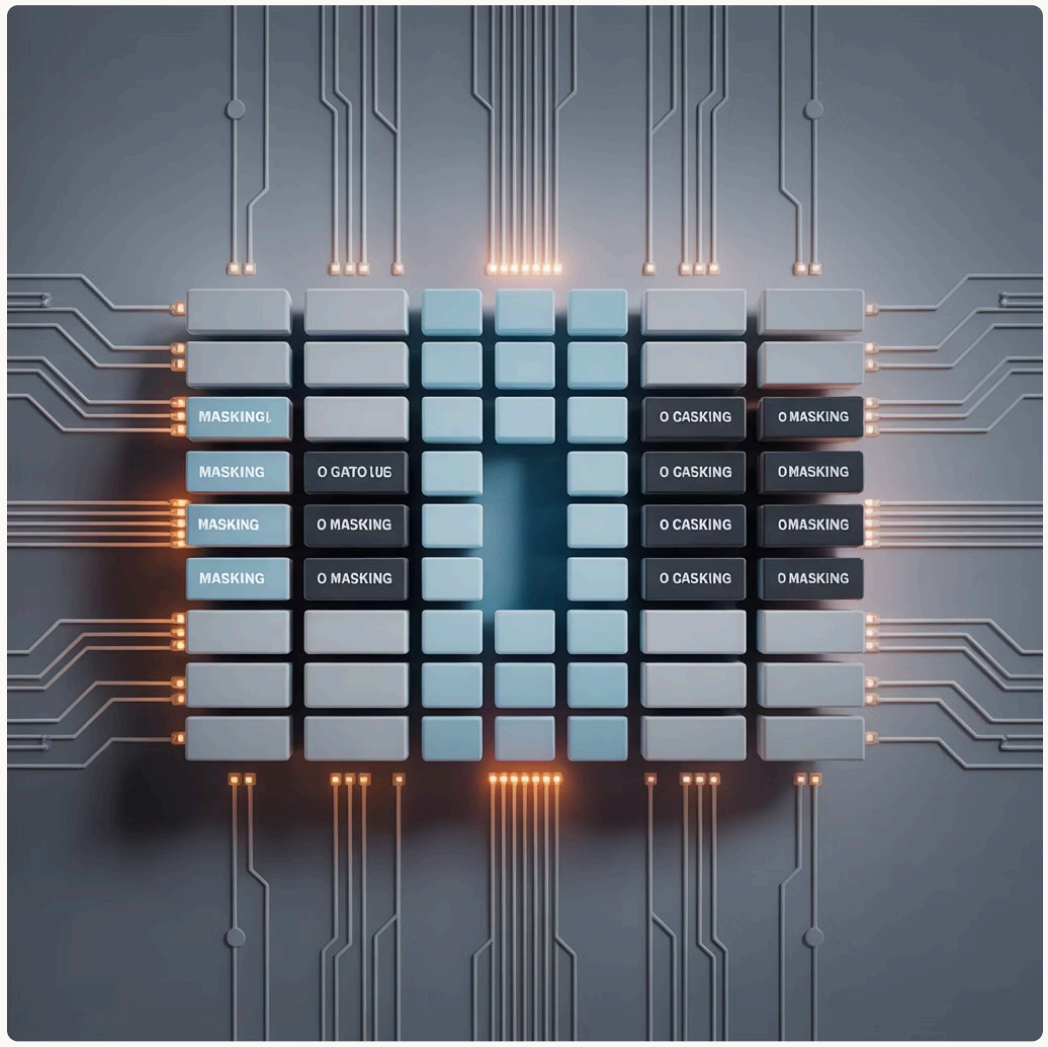
- **80% dos casos**: Substituição pelo token especial [MASK]
- **10% dos casos**: Substituição por um token aleatório
- **10% dos casos**: Manutenção do token original

Esta estratégia variada ajuda a reduzir a discrepância entre treinamento e inferência, já que o token [MASK] não aparece em textos naturais durante a utilização do modelo.

Estratégias Alternativas de Mascaramento

Pesquisadores têm proposto diversas variações da estratégia de mascaramento padrão para melhorar o desempenho dos MLMs:

- **Mascaramento de Span**: Mascarar sequências contíguas de tokens (SpanBERT)
- **Mascaramento Dinâmico**: Gerar novos padrões de mascaramento a cada época (RoBERTa)
- **Mascaramento Linguisticamente Informado**: Priorizar entidades nomeadas e termos importantes
- **Mascaramento Estrutural**: Considerar a estrutura sintática na seleção de tokens



Estudos conduzidos na UFPE e UFABC avaliaram estratégias de mascaramento específicas para o português brasileiro, considerando particularidades morfológicas como contrações, pronomes clíticos e conjugações verbais complexas. Resultados indicam que o mascaramento de unidades morfológicas completas, em vez de subpalavras arbitrárias, melhora o desempenho em tarefas específicas.

Processo de Treinamento dos MLMs

Pré-processamento Textual

Preparação de corpus através de normalização, tokenização e segmentação. Para o português brasileiro, este processo deve considerar acentuação, caracteres especiais e variações ortográficas. Pesquisadores da USP desenvolveram normalizadores específicos para lidar com particularidades do português coloquial encontrado em redes sociais.

Tokenização e Vocabulário

Definição do vocabulário de subpalavras utilizando algoritmos como WordPiece, BPE ou SentencePiece. O tamanho típico varia entre 30K e 50K tokens. Estudos da UNICAMP demonstraram que vocabulários específicos para português requerem ajustes para capturar eficientemente morfologia verbal rica.

Preparação de Dados de Treinamento

Formatação dos dados com mascaramento de tokens conforme a estratégia escolhida. Aplicação de técnicas de amostragem para lidar com desbalanceamento de domínios textuais. Pesquisadores da UFMG desenvolveram técnicas para balancear corpora de português brasileiro entre registros formais e informais.

Treinamento do Modelo

Treinamento com otimizadores como Adam, learning rate com aquecimento e decaimento, regularização via dropout e weight decay. Projetos da UFRJ estabeleceram parâmetros ótimos de treinamento para modelos de português considerando restrições computacionais típicas.

Avaliação e Refinamento

Avaliação em tarefas de validação, ajuste de hiperparâmetros e técnicas de regularização. Pesquisadores da UFF desenvolveram benchmarks específicos para avaliar compreensão de nuances culturais brasileiras em modelos linguísticos.

A qualidade e diversidade do corpus de treinamento são fatores críticos para o desempenho dos MLMs. No contexto brasileiro, grupos de pesquisa têm se dedicado à compilação de corpora abrangentes que representem adequadamente a variabilidade linguística do português brasileiro, incluindo variações regionais e registros formais e informais.

Implementação: Workflow de Treino

Pipeline Típico em Python/Transformers

```
from transformers import BertTokenizer, BertForMaskedLM
from transformers import DataCollatorForLanguageModeling
from transformers import Trainer, TrainingArguments

# Carregar tokenizador
tokenizer = BertTokenizer.from_pretrained("neuralmind/bert-base-portuguese-cased")

# Preparar dataset
def tokenize_function(examples):
    return tokenizer(examples["text"],
                     truncation=True,
                     padding="max_length",
                     max_length=128)

tokenized_datasets = datasets.map(tokenize_function, batched=True)

# Configurar data collator para mascaramento
data_collator = DataCollatorForLanguageModeling(
    tokenizer=tokenizer, mlm=True, mlm_probability=0.15
)

# Configurar treinamento
training_args = TrainingArguments(
    output_dir="./results",
    per_device_train_batch_size=16,
    num_train_epochs=3,
    save_steps=10_000,
    save_total_limit=2,
)

# Inicializar modelo
model = BertForMaskedLM.from_pretrained("neuralmind/bert-base-portuguese-cased")

# Treinar modelo
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    data_collator=data_collator,
)

trainer.train()
```

Uso de Bases Públicas Brasileiras

Pesquisadores brasileiros têm utilizado diversas fontes de dados para treinamento de MLMs em português:

- **BrWaC**: Corpus web brasileiro com 3,5 bilhões de tokens, compilado pela UFSC
- **ptt5**: Dataset baseado em conteúdo web português e brasileiro, desenvolvido pela USP
- **OSCAR**: Subconjunto português do corpus multilíngue filtrado
- **Portal de Notícias da USP**: Artigos jornalísticos sobre ciência e academia
- **Corpus Brasileiro**: Compilado pela UNICAMP com textos de diversos gêneros
- **CORPES**: Corpus de Português Escrito em Setores, da UFMG

Estas bases proporcionam uma representação diversificada do português brasileiro em diferentes registros e domínios, contribuindo para a robustez dos modelos treinados.

Arquiteturas Clássicas: BERT

Bidirectional Encoder Representations from Transformers

BERT, proposto por Devlin et al. (2018), foi o primeiro MLM a ganhar ampla adoção e estabeleceu o paradigma para modelos subsequentes. Suas principais características incluem:

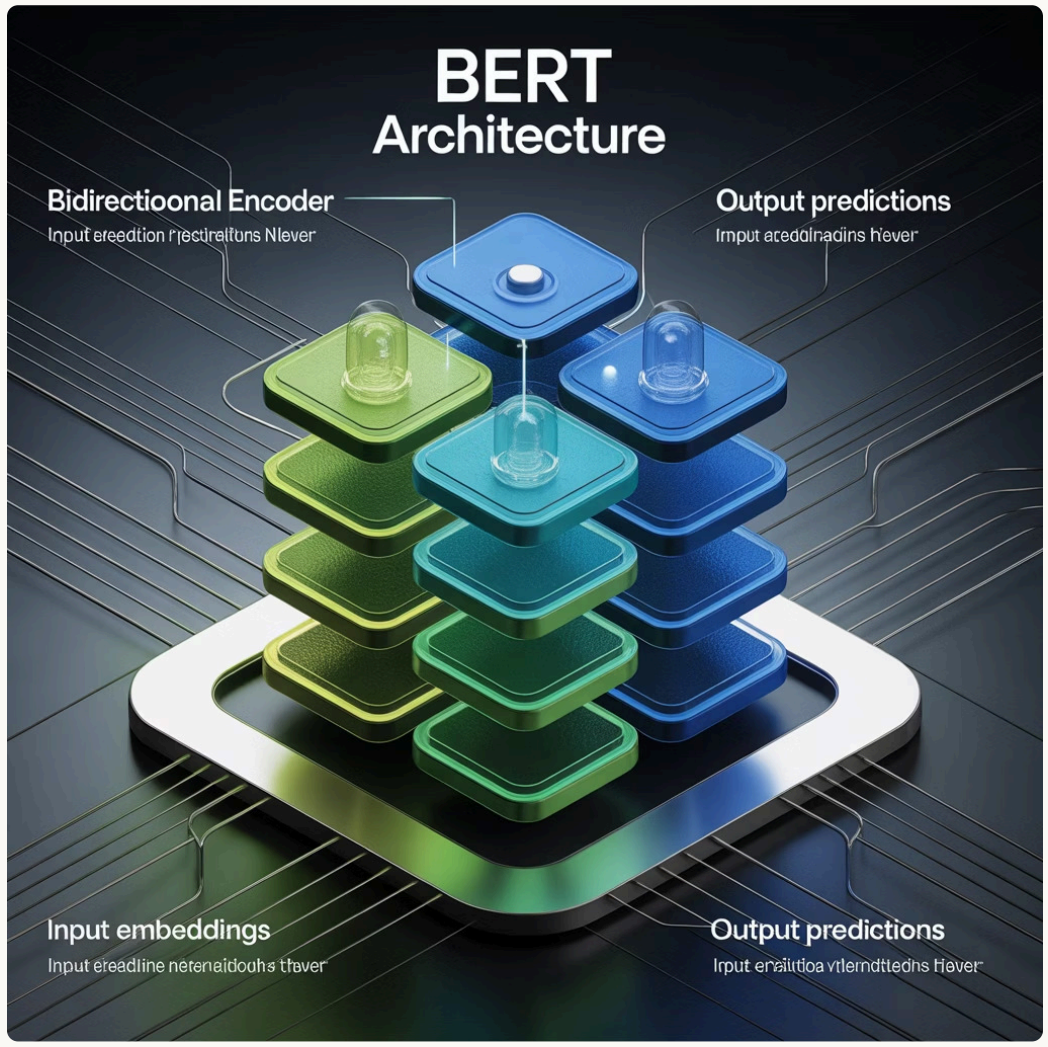
- Arquitetura baseada no codificador Transformer
- Pré-treinamento com mascaramento de tokens e predição de próxima sentença
- Duas variantes principais: BERT-Base (110M parâmetros) e BERT-Large (340M parâmetros)
- Representações contextuais profundas para cada token de entrada
- Capacidade de capturar relações semânticas e sintáticas complexas

O BERT estabeleceu novos recordes em diversos benchmarks de PLN quando foi lançado, demonstrando a eficácia da abordagem de MLM para compreensão de linguagem.

Inovação e Impacto do BERT

O impacto do BERT no campo de PLN foi revolucionário por várias razões:

- Demonstrou a superioridade de modelos bidirecionais sobre unidirecionais para tarefas de compreensão
- Estabeleceu o paradigma de pré-treinar e depois fazer fine-tuning para tarefas específicas
- Viabilizou transferência de conhecimento linguístico entre diferentes domínios e tarefas
- Inspirou dezenas de variantes e aprimoramentos na arquitetura original



No Brasil, o BERTimbau, desenvolvido pela Neuralmind em colaboração com pesquisadores da USP e UNESP, foi o primeiro modelo BERT treinado especificamente para o português brasileiro, alcançando resultados superiores a modelos multilíngues em benchmarks nacionais.

Arquiteturas Otimizadas: RoBERTa, ALBERT, T5

RoBERTa

Desenvolvido pela Facebook AI, o RoBERTa (Robustly Optimized BERT Approach) introduziu melhorias significativas no treinamento:

- Remoção da tarefa de predição de próxima sentença
- Sequências de treinamento mais longas
- Mascaramento dinâmico em vez de estático
- Batches maiores e mais passos de treinamento

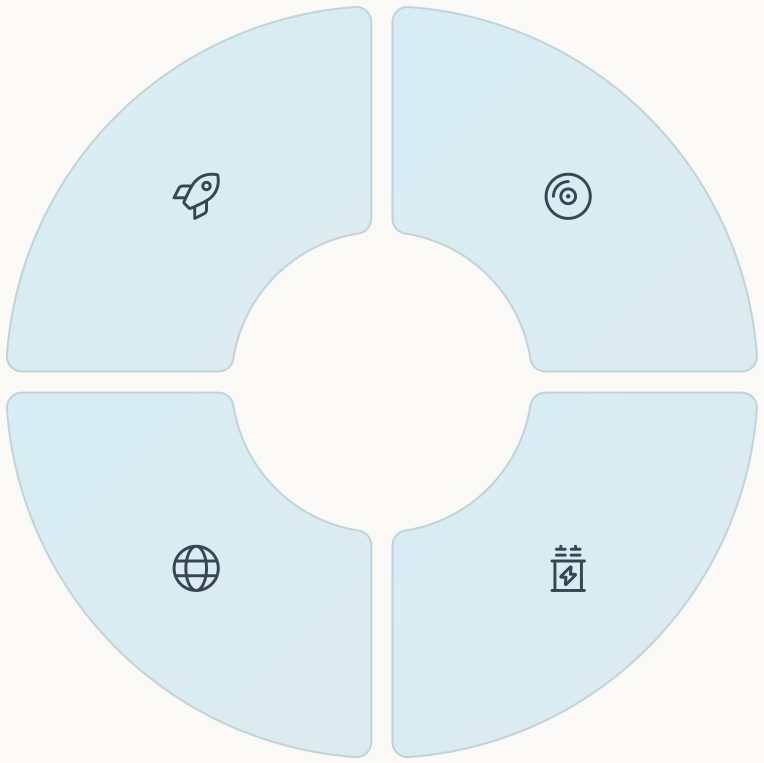
Pesquisadores da UFPE adaptaram as técnicas do RoBERTa para o português, observando ganhos de até 5% em tarefas de classificação de texto.

Modelos Multilíngues

Variantes multilíngues como mBERT e XLM-R foram treinadas em múltiplos idiomas simultaneamente:

- Vocabulário compartilhado entre línguas
- Transferência de conhecimento cross-lingual
- Cobertura de línguas com poucos recursos

A UFRJ conduziu estudos comparativos sobre a eficácia destes modelos para o português em contraste com modelos monolíngues.



ALBERT

O ALBERT (A Lite BERT) focou na redução do tamanho do modelo mantendo o desempenho:

- Fatoração de embeddings para reduzir parâmetros
- Compartilhamento de parâmetros entre camadas
- Substituição da NSP por modelagem de coerência de sentença

A UFABC implementou versões do ALBERT para português, viabilizando o uso em dispositivos com recursos limitados.

T5

O Text-to-Text Transfer Transformer (T5) reformula todas as tarefas de PLN como problemas de texto para texto:

- Arquitetura encoder-decoder completa
- Abordagem unificada para múltiplas tarefas
- Treinamento em escala sem precedentes

Pesquisadores da USP e UNICAMP colaboraram na adaptação do T5 para português brasileiro, desenvolvendo o PTT5.

Estas arquiteturas otimizadas demonstram a rápida evolução do campo, com cada variante abordando limitações específicas dos modelos anteriores ou expandindo suas capacidades. No contexto brasileiro, pesquisadores têm adaptado e avaliado estas arquiteturas para o português, contribuindo com inovações específicas para as características da língua.

MLMs em Português: Pesquisas Nacionais

BERTimbau (USP/UFRJ/UNESP)

O BERTimbau representa um marco significativo no desenvolvimento de MLMs para o português brasileiro. Desenvolvido pela NeuralMind em colaboração com pesquisadores da USP, UFRJ e UNESP, este modelo foi treinado em um corpus substancial de texto em português brasileiro.

- Treinado em 4.6 bilhões de tokens de texto em português
- Disponível nas versões base (110M parâmetros) e large (335M parâmetros)
- Supera consistentemente o mBERT em benchmarks brasileiros
- Incorpora conhecimento linguístico específico do português brasileiro

O BERTimbau estabeleceu novos padrões de desempenho em tarefas como análise de sentimento, reconhecimento de entidades nomeadas e classificação de texto em português.

Outros Modelos Nacionais Relevantes

- **PTT5**: Adaptação do T5 para português brasileiro, desenvolvido por pesquisadores da USP e UFF
- **BrBERT**: Modelo BERT treinado em corpus web brasileiro pela UFMG
- **AlbertPT**: Implementação do ALBERT para português pela UFABC
- **GPorTuguese**: Adaptação de modelos GPT para português pela UFPE
- **RoBERTa-PT**: Versão do RoBERTa otimizada para português pela UNICAMP

Um aspecto notável das pesquisas nacionais em MLMs é a preocupação com particularidades linguísticas do português brasileiro, como contração de preposições com artigos, colocação pronominal, concordância verbal complexa e variabilidade regional. Estudos da UNICAMP e USP têm avaliado sistematicamente como diferentes arquiteturas de MLMs capturam estas características linguísticas específicas.

Exemplos Práticos de Mascaramento

Exemplo de Frase Mascarada e Predição

> Texto Original: "A Universidade de São Paulo é uma instituição de ensino superior pública brasileira, mantida pelo estado de São Paulo."
>
> Texto Mascarado: "A Universidade de São Paulo é uma [MASK] de ensino superior [MASK] brasileira, mantida pelo [MASK] de São Paulo."

Predições de um MLM como o BERTimbau:

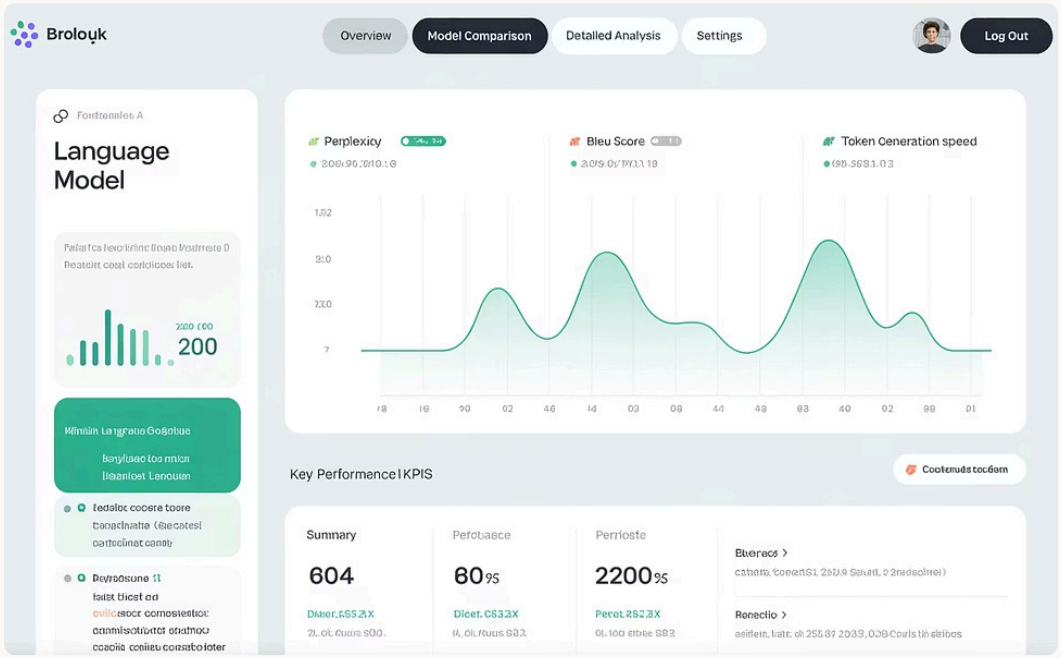
- Para o primeiro [MASK]: "instituição" (95%), "entidade" (3%), "escola" (1%)
- Para o segundo [MASK]: "pública" (87%), "privada" (8%), "federal" (3%)
- Para o terceiro [MASK]: "estado" (92%), "governo" (6%), "município" (1%)

Este exemplo demonstra como o modelo utiliza o contexto bidirecionalmente para prever com alta precisão os tokens mascarados, mesmo quando múltiplas palavras estão ausentes.

Código Exemplo em Python

```
from transformers import AutoModelForMaskedLM, AutoTokenizer
import torch

# Carregar modelo e tokenizador
model_name = "neuralmind/bert-base-portuguese-cased"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForMaskedLM.from_pretrained(model_name)

# Definir texto com mascaramento
text = "A Universidade de São Paulo é uma [MASK] de ensino."

# Tokenizar entrada
inputs = tokenizer(text, return_tensors="pt")

# Identificar posição do token [MASK]
mask_token_index = torch.where(inputs["input_ids"] ==
 tokenizer.mask_token_id)[1]

# Obter predições
outputs = model(**inputs)
predictions = outputs.logits

# Obter os 5 tokens mais prováveis para cada [MASK]
probs, indices = torch.topk(
 torch.softmax(predictions[0, mask_token_index], dim=-1), k=5
)

# Mostrar resultados
for i, (prob, index) in enumerate(zip(probs[0], indices[0])):
 token = tokenizer.convert_ids_to_tokens([index])[0]
 print(f"{token}: {prob.item():.2%}")
```

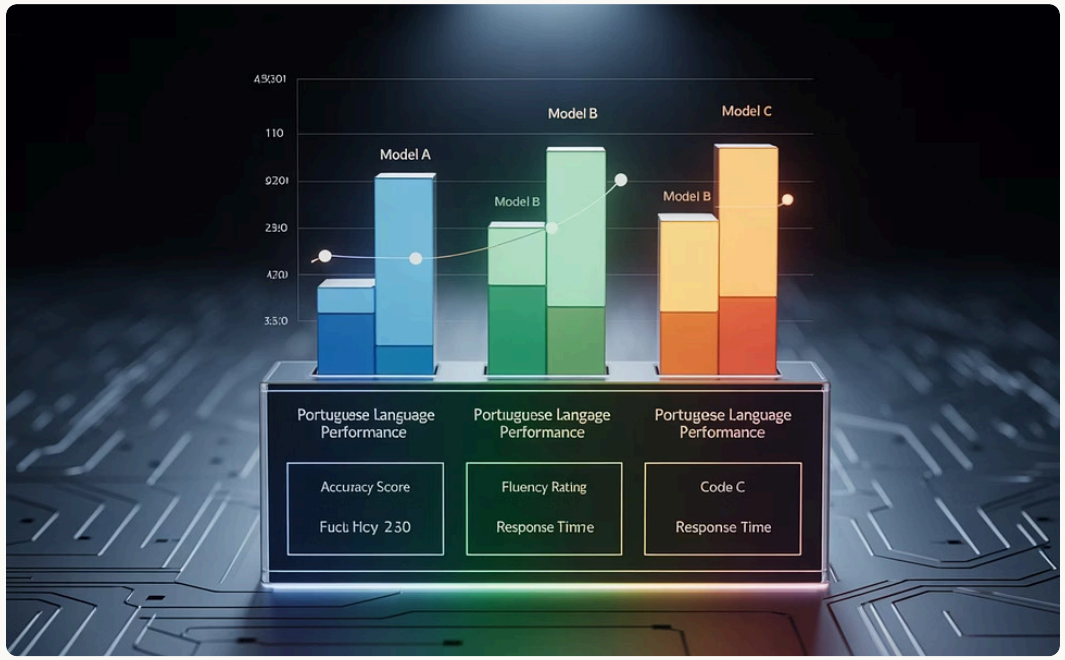
Avaliação e Métricas de Desempenho



Métricas Principais

A avaliação de MLMs envolve diversas métricas específicas para monitorar o desempenho durante o treinamento e a qualidade final do modelo:

- **Accuracy**: Percentagem de tokens mascarados corretamente preditos
- **Perplexity**: Medida da incerteza do modelo ao prever tokens mascarados (valores menores são melhores)
- **Loss**: Valor da função de perda durante o treinamento, geralmente entropia cruzada
- **MLM F1**: Equilíbrio entre precisão e cobertura nas predições de tokens mascarados



Benchmarks Nacionais e Internacionais

Para avaliar o desempenho em tarefas específicas após o fine-tuning, diversos benchmarks são utilizados:

- **ASSIN**: Corpus para avaliação de inferência textual em português brasileiro, desenvolvido pela USP
- **FaQCS**: Dataset de classificação de perguntas em português, da UFMG
- **Brazilian Portuguese Corpora (BPC)**: Conjunto de datasets diversos para avaliação de tarefas variadas, compilado pela UFPE
- **WikiQA-PT**: Dataset de perguntas e respostas em português da UNICAMP
- **GLUE-PT**: Adaptação do benchmark GLUE para português brasileiro, pela UFRJ

Estudos comparativos realizados por pesquisadores da USP e UFMG demonstraram que modelos específicos para português brasileiro, como o BERTimbau, superam consistentemente modelos multilíngues em benchmarks nacionais, com ganhos de 3-15% dependendo da tarefa. Estas avaliações sistemáticas são essenciais para direcionar o desenvolvimento de novos modelos e aprimorar os existentes.

Um desafio particular na avaliação de MLMs para português é a escassez de benchmarks amplos e padronizados, comparáveis ao GLUE em inglês. Pesquisadores da UNICAMP e UFRJ têm trabalhado no desenvolvimento de recursos de avaliação mais abrangentes específicos para o português brasileiro.

Ajuste Fino (Fine-tuning) em MLMs

Seleção de Dados

Identificação e preparação de datasets rotulados específicos para a tarefa alvo. Para o português brasileiro, pesquisadores da UFMG compilaram datasets para classificação de textos jurídicos, enquanto a USP desenvolveu corpora para análise de sentimento em reviews de produtos.

Adaptação da Arquitetura

Adição de camadas específicas para a tarefa (classificação, span prediction, etc.) sobre as representações do MLM. Estudos da UNICAMP demonstraram que diferentes configurações de camadas finais têm impacto significativo no desempenho para tarefas em português.

Treinamento Supervisionado

Ajuste dos parâmetros do modelo usando dados rotulados com objetivos específicos da tarefa. Pesquisadores da UFRJ estabeleceram parâmetros ótimos de learning rate e batch size para fine-tuning em datasets em português brasileiro.

Avaliação Específica

Utilização de métricas apropriadas para a tarefa (acurácia, F1, BLEU, etc.). A UFPE desenvolveu frameworks de avaliação específicos para aplicações em português, considerando particularidades linguísticas e culturais.

Adaptação para Tarefas Específicas

O processo de fine-tuning permite adaptar MLMs pré-treinados para diversas tarefas de PLN:

- **Classificação de Texto**: Adiciona-se uma camada de classificação sobre o token [CLS]
- **Question Answering**: Treina-se o modelo para identificar spans de texto que contêm respostas
- **NER**: Adaptação para identificar e classificar entidades nomeadas
- **Análise de Sentimento**: Fine-tuning para detectar polaridade e emoções em textos

Técnicas Avançadas de Fine-tuning

- **Gradual Unfreezing**: Descongelamento progressivo das camadas durante o treinamento
- **Discriminative Fine-tuning**: Taxas de aprendizado diferentes para camadas distintas
- **Layer-wise Learning Rate Decay**: Redução progressiva das taxas em direção às camadas iniciais
- **Parameter-Efficient Fine-tuning**: Técnicas como adapters, prefix-tuning e LoRA

Pesquisadores da UFABC demonstraram que estas técnicas avançadas podem reduzir até 70% o tempo de fine-tuning para aplicações em português brasileiro, mantendo níveis comparáveis de desempenho.

Transferência de Aprendizagem

Princípios da Transferência de Aprendizagem

A transferência de aprendizagem é um conceito fundamental para o sucesso dos MLMs, permitindo que conhecimento adquirido durante o pré-treinamento seja aplicado a novas tarefas com quantidade limitada de dados rotulados.

Este processo envolve:

- Utilização de representações genéricas aprendidas durante o pré-treinamento
- Adaptação destas representações para tarefas específicas
- Preservação de conhecimento linguístico profundo, como sintaxe e semântica
- Redução significativa da necessidade de dados rotulados para novas tarefas

Pesquisadores da USP e UNICAMP demonstraram que MLMs pré-treinados em português brasileiro podem alcançar resultados competitivos em tarefas específicas com apenas 10-20% dos dados rotulados que seriam necessários para treinar modelos do zero.

Transferência entre Domínios e Tarefas

A transferência de conhecimento pode ocorrer em diferentes níveis:

- **Entre Domínios**: Por exemplo, do domínio geral para texto jurídico ou médico
- **Entre Tarefas**: Como de classificação para extração de informação
- **Entre Línguas**: De modelos multilíngues para aplicações específicas em português

Estudos da UFMG avaliaram a eficácia da transferência entre diferentes domínios textuais em português brasileiro, encontrando padrões de transferência distintos dos observados em estudos com textos em inglês, particularmente em domínios técnicos como jurídico e médico.

> "A transferência de aprendizagem é particularmente valiosa no contexto brasileiro, onde a escassez de dados rotulados em diversos domínios especializados representa um desafio significativo para o desenvolvimento de aplicações de PLN." — Sandra Aluísio, pesquisadora da USP

Casos de Aplicação em PLN

Análise de Sentimentos

MLMs têm revolucionado a análise de sentimentos em português brasileiro, permitindo capturar nuances contextuais e expressões idiomáticas. O Banco Central do Brasil utiliza modelos baseados em BERTimbau para monitorar sentimentos em notícias econômicas, conforme estudo publicado pela USP.

Pesquisadores da UFMG desenvolveram um sistema que analisa sentimentos em redes sociais sobre serviços públicos, auxiliando gestores a identificar áreas problemáticas.

Respostas Automáticas

Sistemas de atendimento automatizado baseados em MLMs têm sido implementados em diversos setores no Brasil. A UNICAMP, em parceria com empresas de telecomunicações, desenvolveu assistentes virtuais capazes de compreender e responder a consultas complexas de clientes.

O setor bancário brasileiro tem sido pioneiro na adoção destas tecnologias, com instituições como Itaú e Bradesco implementando chatbots avançados baseados em MLMs para serviços ao cliente.

Tradução Automática

MLMs têm contribuído significativamente para avanços em tradução automática envolvendo o português. Pesquisadores da UFRJ desenvolveram modelos específicos para tradução português-inglês que superam sistemas tradicionais em métricas como BLEU e TER.

Um projeto colaborativo entre UFPE e USP criou um sistema especializado para tradução de textos jurídicos e documentos oficiais, mantendo terminologia consistente e estilo formal apropriado.

Estas aplicações demonstram como os MLMs estão sendo adaptados para necessidades específicas do contexto brasileiro, considerando particularidades linguísticas, culturais e de domínio. A capacidade destes modelos de capturar nuances contextuais os torna particularmente adequados para tarefas que exigem compreensão semântica profunda.

Estudos conduzidos pela UFABC e UFF demonstram que o desempenho de MLMs específicos para português brasileiro nestas aplicações supera consistentemente tanto modelos tradicionais quanto adaptações de modelos multilíngues, justificando o investimento em desenvolvimento de recursos linguísticos nacionais.

Outras Aplicações Práticas

Detecção de Fake News (USP, UFMG, UFF)

A proliferação de desinformação representa um desafio significativo no contexto brasileiro. Pesquisadores têm explorado MLMs para desenvolver sistemas eficazes de detecção de notícias falsas:

- O grupo de Ciência de Dados da USP desenvolveu o "FakeCheck-BR", um sistema baseado em BERTimbau que analisa padrões linguísticos indicativos de desinformação em português brasileiro
- Pesquisadores da UFMG criaram um modelo que integra análise de credibilidade de fontes com avaliação de conteúdo, alcançando 87% de precisão em datasets nacionais
- A UFF estabeleceu uma parceria com o Tribunal Superior Eleitoral para implementar ferramentas de monitoramento de desinformação eleitoral baseadas em MLMs

Estas iniciativas demonstram o potencial dos MLMs para aplicações de alto impacto social, com capacidade de adaptação às particularidades do contexto informacional brasileiro.

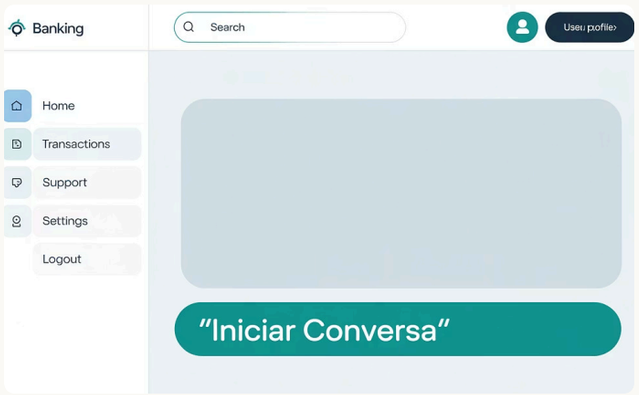
Resumo e Geração de Texto (UFABC, UFPE)

Sistemas de resumo automático e geração de texto baseados em MLMs têm encontrado aplicações diversas:

- O projeto "SumBR" da UFABC utiliza modelos baseados em T5 adaptados para português para gerar resumos de artigos científicos e documentos técnicos
- Pesquisadores da UFPE desenvolveram um sistema de geração de relatórios automatizados para dados governamentais, facilitando a interpretação de estatísticas públicas
- A UNICAMP implementou um modelo para simplificação textual que adapta conteúdos complexos para diferentes níveis de letramento, com aplicações educacionais

A adaptação de MLMs para estas aplicações específicas envolve desafios particulares relacionados às características do português brasileiro, como a riqueza verbal e a estrutura sintática flexível. Soluções desenvolvidas por pesquisadores brasileiros frequentemente incorporam conhecimento linguístico específico para superar estas dificuldades.

MLMs em Ambientes Industriais



Atendimento Automático

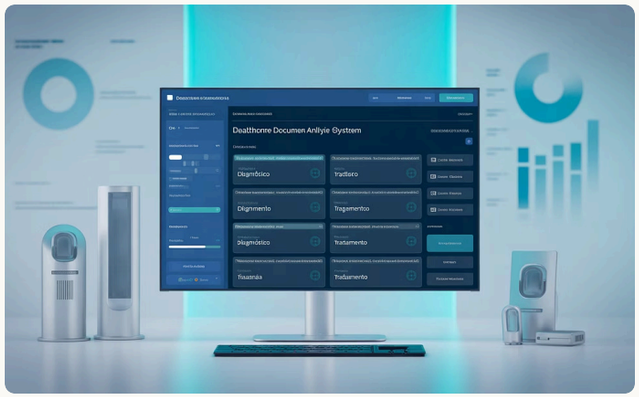
O setor de serviços brasileiro tem adotado MLMs para melhorar a experiência de atendimento ao cliente:

- Grandes varejistas como Magazine Luiza e B2W implementaram assistentes virtuais baseados em MLMs capazes de compreender consultas complexas em linguagem natural
- Estudos da UFPE demonstram redução de 40% no tempo médio de atendimento e aumento de 35% na satisfação do cliente após implementação destas soluções
- Empresas de telecomunicações como Vivo e Claro utilizam MLMs para triagem e resolução automática de problemas técnicos comuns

Setor Bancário

Instituições financeiras brasileiras lideram a adoção de MLMs em diversos processos:

- Análise automatizada de documentos para concessão de crédito, reduzindo o tempo de processamento em até 80%
- Detecção de fraudes através da análise semântica de transações e comunicações
- Assistentes virtuais para planejamento financeiro personalizado
- Análise de sentimento em notícias para prever impactos em investimentos



Saúde

O setor de saúde tem explorado MLMs para aplicações especializadas:

- Extração e estruturação de informações de prontuários médicos, conforme pesquisa da UNICAMP
- Classificação automatizada de relatos de eventos adversos para farmacovigilância
- Sistemas de resposta a perguntas sobre medicamentos e protocolos para profissionais de saúde
- Análise de tendências em dados epidemiológicos textuais

A aplicação industrial de MLMs no Brasil enfrenta desafios específicos relacionados à integração com sistemas legados, questões regulatórias e adaptação para domínios especializados. Colaborações entre universidades e empresas, como exemplificado por parcerias entre UFMG e Itaú ou USP e Hospital Albert Einstein, têm sido fundamentais para superar estas barreiras.

Limitações dos MLMs

Viés dos Dados e Fairness

MLMs treinados em grandes corpora de texto tendem a absorver e potencialmente amplificar vieses presentes nos dados de treinamento:

- Estudos da USP identificaram vieses de gênero em representações de profissões em modelos treinados em português brasileiro
- Pesquisadores da UFMG documentaram vieses raciais e regionais em modelos pré-treinados, refletindo desigualdades sociais presentes nos corpora
- A UFRJ desenvolveu métricas específicas para quantificar vieses em modelos linguísticos no contexto brasileiro

Estas questões são particularmente relevantes no Brasil, onde desigualdades históricas podem ser reforçadas por sistemas automatizados se não forem adequadamente mitigadas durante o desenvolvimento e aplicação dos modelos.

Iniciativas como o "Observatório de IA Ética" da UNICAMP visam estabelecer diretrizes para desenvolvimento responsável de MLMs no contexto brasileiro.

Dificuldades com Contextos Longos

MLMs tradicionais como BERT enfrentam limitações significativas ao processar textos longos:

- Restrição típica de 512 tokens, insuficiente para muitos documentos reais
- Decaimento da atenção em tokens distantes, prejudicando a compreensão de referências de longa distância
- Dificuldade em capturar estrutura argumentativa global de documentos extensos

Estas limitações são particularmente relevantes para aplicações jurídicas e acadêmicas no Brasil, onde documentos tendem a ser extensos e formais.

Pesquisadores da UFABC e USP têm explorado técnicas como segmentação hierárquica e atenção esparsa para superar estas limitações em aplicações específicas para o português brasileiro.

Reconhecer e abordar estas limitações é essencial para o desenvolvimento responsável e eficaz de MLMs no contexto brasileiro. Pesquisadores nacionais têm contribuído ativamente para a literatura sobre mitigação de vieses e processamento de documentos longos, com soluções adaptadas às particularidades linguísticas e socioculturais do Brasil.

Considerações Éticas

Implicações para Privacidade

MLMs treinados em dados públicos podem inadvertidamente memorizar informações sensíveis presentes nos corpora de treinamento. Pesquisadores da UFPE identificaram casos onde modelos reproduziam dados pessoais de brasileiros presentes em documentos de treinamento.

A Lei Geral de Proteção de Dados (LGPD) brasileira impõe restrições específicas ao uso de dados pessoais para treinamento de modelos de IA. O grupo de Direito Digital da USP desenvolveu diretrizes para conformidade de projetos de MLM com a legislação nacional.

Questões de Segurança

MLMs podem ser vulneráveis a ataques adversariais e manipulação. Estudos da UNICAMP demonstraram que modelos em português podem ser induzidos a gerar conteúdo inapropriado ou falso através de prompt engineering malicioso.

Pesquisadores da UFMG desenvolveram técnicas de avaliação de robustez específicas para MLMs em português brasileiro, considerando particularidades linguísticas como variação dialetal e expressões idiomáticas regionais que podem ser exploradas em ataques.

Pesquisas Nacionais em Ética de IA

O Brasil tem contribuído significativamente para discussões éticas sobre MLMs e IA linguística. O Centro de Ética em IA da USP coordena pesquisas interdisciplinares sobre impactos sociais de tecnologias de linguagem no contexto brasileiro.

A UFRJ estabeleceu o Observatório de Tecnologias Linguísticas, que monitora aplicações de MLMs e avalia seus impactos em diferentes setores da sociedade brasileira, com atenção especial a questões de inclusão e acessibilidade.

A intersecção entre tecnologias linguísticas avançadas e o contexto sociocultural brasileiro apresenta desafios éticos particulares. A presença de desigualdades estruturais, diversidade linguística regional e peculiaridades do sistema jurídico nacional exigem abordagens específicas para o desenvolvimento e implementação responsável de MLMs no Brasil.

Iniciativas como o "Framework Ético para PLN" da UNICAMP buscam estabelecer princípios e práticas adaptados à realidade brasileira, considerando aspectos como representatividade linguística, neutralidade cultural e transparência algorítmica.

Otimizando Treinamento de MLMs

Técnicas Modernas de Paralelização

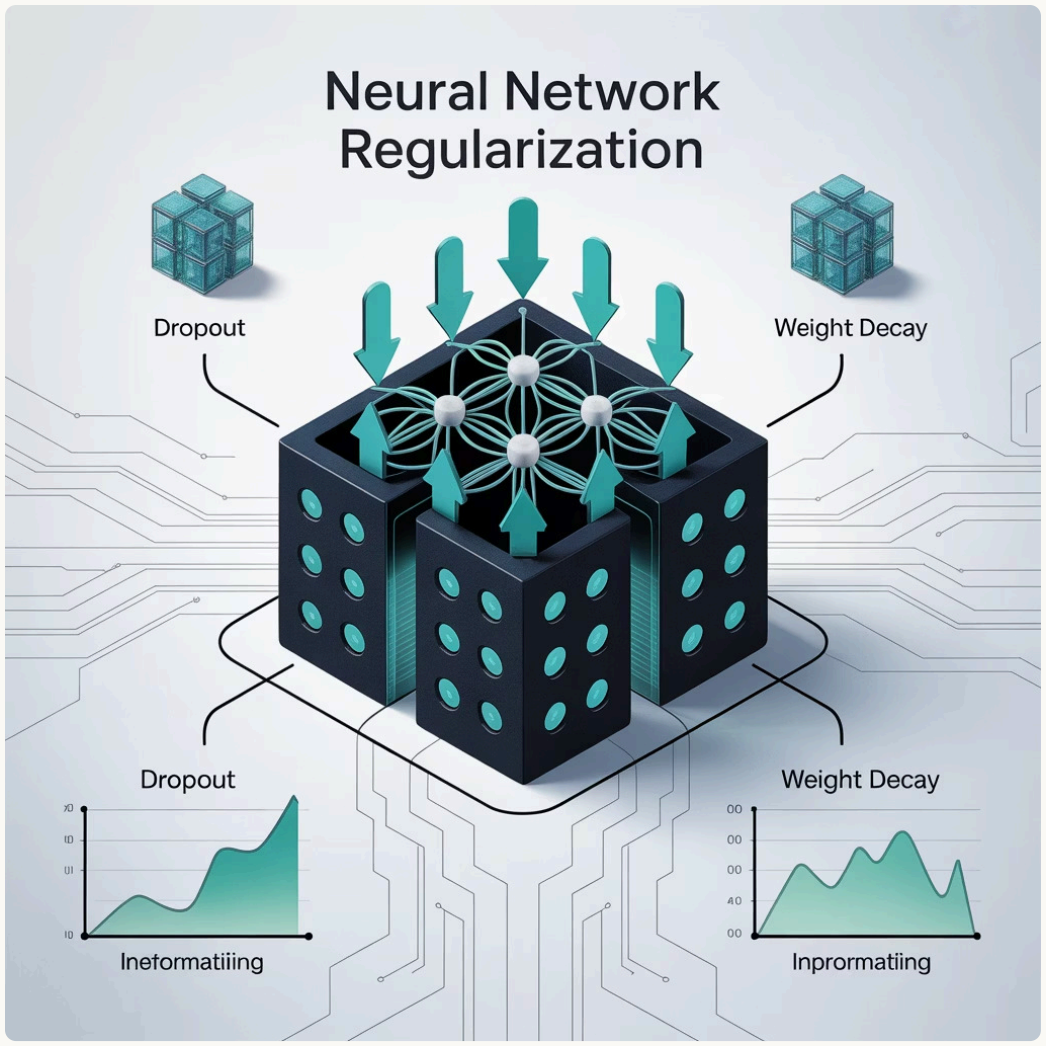
O treinamento eficiente de MLMs requer estratégias avançadas de paralelização para gerenciar modelos cada vez maiores:

- **Paralelismo de Dados**: Distribuição de batches entre múltiplos dispositivos, reduzindo tempo de processamento
- **Paralelismo de Modelo**: Particionamento do modelo entre dispositivos, permitindo modelos maiores que a memória de um único dispositivo
- **Paralelismo de Pipeline**: Divisão de camadas do modelo entre dispositivos, otimizando utilização de hardware

Pesquisadores da USP e UFABC adaptaram estas técnicas para infraestruturas disponíveis em universidades brasileiras, estabelecendo pipelines otimizados para treinamento de MLMs em português com recursos computacionais limitados.

Um estudo da UNICAMP demonstrou ganhos de eficiência de até 60% ao implementar estratégias de paralelização adaptadas para clusters heterogêneos típicos de universidades brasileiras.

Estratégias de Regularização



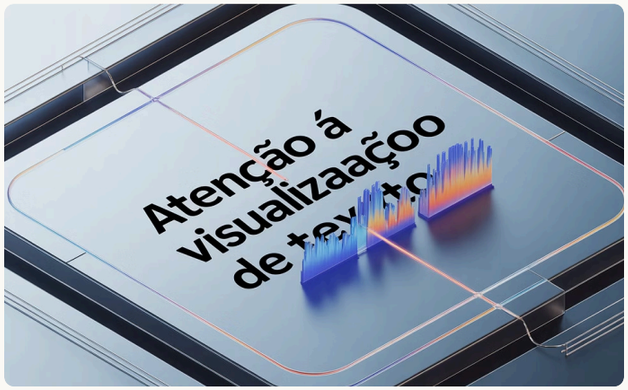
Para prevenir overfitting e melhorar generalização, diversas técnicas de regularização são aplicadas:

- **Dropout**: Desativação aleatória de neurônios durante treinamento
- **Weight Decay**: Penalização de pesos grandes para favorecer modelos mais simples
- **Early Stopping**: Interrupção do treinamento quando performance em validação começa a degradar
- **Gradient Clipping**: Limitação da magnitude dos gradientes para estabilizar treinamento

Pesquisadores da UFMG estabeleceram parâmetros ótimos de regularização para corpora em português brasileiro, considerando características como maior variabilidade morfológica e sintática em comparação com inglês.

A otimização de treinamento é particularmente relevante no contexto brasileiro, onde recursos computacionais para pesquisa em IA frequentemente são mais limitados em comparação com centros internacionais. Estratégias inovadoras desenvolvidas por pesquisadores nacionais têm permitido avanços significativos mesmo com estas restrições.

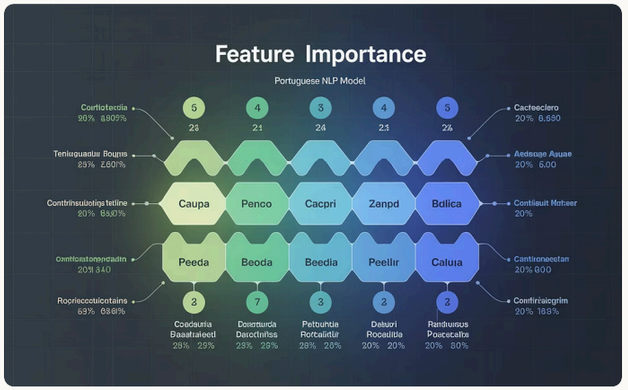
Interpretação e Explainability em MLMs



Visualização de Atenção

Ferramentas como BERTViz, adaptadas por pesquisadores da USP para modelos em português, permitem visualizar padrões de atenção entre tokens. Estudos com textos jurídicos brasileiros revelaram que MLMs desenvolvem atenção especializada para termos técnicos e construções verbais específicas do contexto legal.

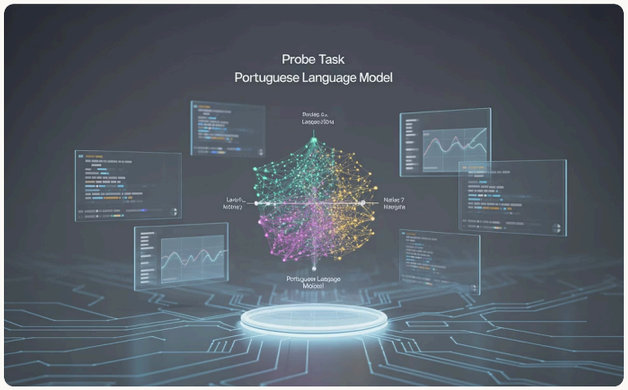
A UNICAMP desenvolveu interfaces interativas que permitem a não-especialistas explorar e compreender decisões de MLMs em aplicações práticas como análise de sentimento em reviews de produtos.



Análise de Importância de Tokens

Técnicas como LIME e SHAP, adaptadas para português pela UFMG, identificam quais tokens mais influenciam predições específicas. Um estudo com classificação de documentos acadêmicos revelou diferenças significativas nos padrões de decisão entre modelos pré-treinados em português brasileiro versus adaptações de modelos multilíngues.

Pesquisadores da UFRJ desenvolveram métricas específicas para quantificar a confiabilidade de explicações em MLMs considerando particularidades linguísticas do português.



Probing Tasks

Tarefas de sondagem (probing tasks) desenvolvidas pela UFPE permitem avaliar quais propriedades linguísticas são capturadas em diferentes camadas dos MLMs. Análises revelaram que informações morfossintáticas do português, como conjugação verbal e concordância, são codificadas nas camadas intermediárias, enquanto relações semânticas aparecem em camadas superiores.

A UFF adaptou frameworks de probing para avaliar o conhecimento cultural brasileiro implícito em MLMs, testando compreensão de referências regionais e eventos históricos nacionais.

A interpretabilidade de MLMs é crucial para aplicações em setores regulados como saúde, finanças e direito, particularmente relevantes no contexto brasileiro. Pesquisadores nacionais têm contribuído significativamente para o desenvolvimento de técnicas de explainability adaptadas às particularidades linguísticas e contextuais do português brasileiro.

Um desafio específico no contexto brasileiro é a interpretação de modelos que processam linguagem coloquial com variações regionais significativas. Estudos da UFABC demonstraram que técnicas de interpretação convencionais podem falhar ao explicar decisões envolvendo expressões idiomáticas e gírias regionais.

Riscos e Robustez

Adversarial Attacks em PLN

MLMs são vulneráveis a ataques adversariais que podem comprometer seu desempenho e segurança:

- **Ataques de Substituição**: Troca de palavras por sinônimos ou caracteres visualmente similares para enganar o modelo
- **Ataques de Inserção**: Adição de palavras irrelevantes mas influentes para manipular predições
- **Prompt Injection**: Introdução de instruções maliciosas que subvertem o comportamento pretendido
- **Data Poisoning**: Contaminação dos dados de treinamento para introduzir vulnerabilidades específicas

Pesquisadores da UFMG identificaram vulnerabilidades específicas em MLMs para português relacionadas à rica morfologia da língua, onde pequenas modificações em sufixos podem alterar drasticamente predições.

Propostas para Defesa (USP/UFMG)

Pesquisadores brasileiros têm desenvolvido técnicas específicas para aumentar a robustez de MLMs contra ataques:

- O laboratório de Segurança Computacional da USP desenvolveu métodos de detecção de textos adversariais baseados em análise de coerência semântica
- A UFMG propôs técnicas de augmentação de dados para treinamento que incorporam exemplos adversariais, aumentando robustez
- A UNICAMP implementou uma camada de defesa baseada em análise morfossintática específica para o português
- A UFRJ criou benchmarks para avaliação sistemática de robustez contra diferentes tipos de ataques em português brasileiro

> "A robustez de modelos linguísticos no contexto brasileiro apresenta desafios particulares devido à variabilidade dialetal e riqueza morfológica do português. Sistemas de defesa eficazes precisam considerar estas particularidades linguísticas." — Thiago Pardo, pesquisador da USP

A crescente adoção de MLMs em aplicações críticas como sistemas bancários, jurídicos e de saúde no Brasil torna a questão da robustez particularmente relevante. Iniciativas colaborativas entre academia e indústria buscam estabelecer padrões e práticas para desenvolvimento e implantação segura destas tecnologias no contexto nacional.

O Papel do Pré-Processamento

Limpeza de Dados

A qualidade do pré-processamento impacta significativamente o desempenho final dos MLMs. Pesquisadores da USP desenvolveram pipelines específicos para textos em português brasileiro, considerando particularidades como:

- Normalização de acentuação e caracteres especiais
- Tratamento de abreviações e contrações comuns
- Remoção de artefatos específicos de fontes web brasileiras

Tokenização

A escolha do algoritmo de tokenização e a construção do vocabulário são cruciais para MLMs em português. Estudos da UNICAMP compararam diferentes abordagens:

- WordPiece: eficiente para morfologia verbal complexa
- BPE: melhor para neologismos e termos técnicos
- SentencePiece: vantajoso para textos com variações dialetais

Normalização

Técnicas de normalização específicas para português brasileiro foram desenvolvidas pela UFPE:

- Padronização de variantes ortográficas regionais
- Tratamento de estrangeirismos comuns no português brasileiro
- Normalização de entidades nomeadas com características locais

Enriquecimento

Pesquisadores da UFMG exploraram técnicas de enriquecimento de dados para melhorar a robustez e cobertura dos modelos:

- Augmentação com variantes dialetais brasileiras
- Incorporação de conhecimento linguístico específico
- Balanceamento de registros formais e informais

O pré-processamento adequado é particularmente crítico para o português brasileiro devido à sua rica morfologia, variabilidade dialetal e presença de construções linguísticas específicas. Pesquisadores da UFRJ demonstraram que a adaptação de técnicas de pré-processamento para características específicas do português pode melhorar o desempenho final dos MLMs em até 12% em tarefas como análise de sentimento e classificação de textos.

Um desafio particular no contexto brasileiro é o tratamento adequado do português coloquial usado em redes sociais, com abreviações, gírias e construções não-padrão. A UFABC desenvolveu normalizadores específicos para este tipo de conteúdo, melhorando significativamente o desempenho de MLMs em análise de opinião online.

Arquiteturas Emergentes

Modelos Multimodais (Texto + Imagem)

A integração de múltiplas modalidades representa uma fronteira promissora para MLMs, expandindo suas capacidades além do texto:

- Pesquisadores da UNICAMP desenvolveram o "VisualBERTpt", adaptando a arquitetura VisualBERT para português brasileiro
- O laboratório de IA da USP implementou uma versão do CLIP com legendas em português, demonstrando melhor desempenho em datasets visuais brasileiros
- A UFMG explorou aplicações de modelos multimodais para análise de mídias sociais brasileiras, integrando texto e imagens

Estas arquiteturas multimodais abrem possibilidades para aplicações como acessibilidade digital, comércio visual e educação interativa, particularmente relevantes no contexto brasileiro.

Transformers de Última Geração

Avanços recentes em arquiteturas Transformer têm sido adaptados para o português brasileiro:

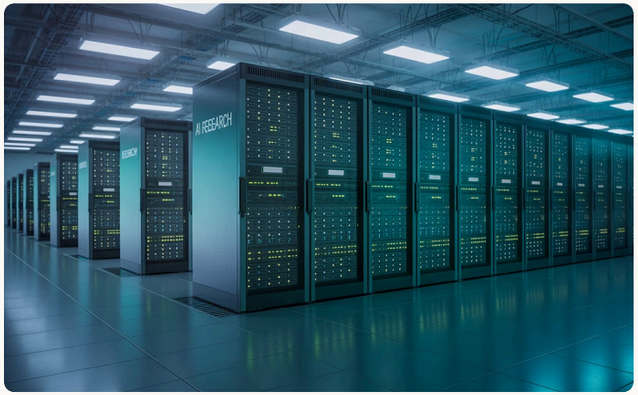
- **Transformers Eficientes**: Pesquisadores da UFABC implementaram versões do Longformer e Reformer otimizadas para processamento de documentos longos em português
- **Sparse Transformers**: A UFRJ explorou padrões de atenção esparsa adaptados para estruturas sintáticas do português
- **Parameter-Efficient Transformers**: A USP desenvolveu técnicas de adaptação eficiente para domínios específicos como jurídico e médico

Estas arquiteturas avançadas abordam limitações específicas dos MLMs tradicionais, como processamento de documentos longos e adaptação eficiente para domínios específicos, desafios particularmente relevantes para aplicações em português brasileiro.

Um estudo comparativo da UNICAMP demonstrou que arquiteturas eficientes como o Performer apresentam vantagens significativas para o processamento de textos jurídicos brasileiros, que tendem a ser extensos e formais.

A pesquisa em arquiteturas emergentes no Brasil frequentemente foca na adaptação de modelos avançados para contextos com restrições computacionais, refletindo a realidade de muitas instituições nacionais. Esta abordagem tem resultado em contribuições significativas para o campo de MLMs eficientes e otimizados.

Infraestrutura para Treino de MLMs



Clusters GPU em Universidades Brasileiras

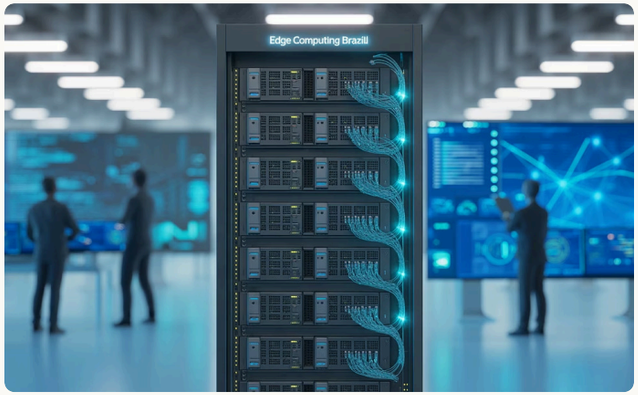
Diversas universidades brasileiras têm investido em infraestrutura computacional para viabilizar pesquisas em MLMs:

- O Centro de Computação Científica e Matemática Aplicada (CCMA) da USP mantém um cluster com 40 GPUs NVIDIA V100, utilizado para treinamento de modelos como o BERTimbau
- O Laboratório Nacional de Computação Científica (LNCC) disponibiliza o supercomputador Santos Dumont para pesquisadores de diversas instituições
- A UFMG estabeleceu o Centro de IA com infraestrutura dedicada para pesquisa em modelos linguísticos

Colaborações e Recursos Compartilhados

Diante das limitações de recursos, pesquisadores brasileiros têm adotado abordagens colaborativas:

- A rede ComputaBR conecta recursos computacionais de múltiplas instituições para projetos colaborativos
- Parcerias com empresas como Google e Microsoft têm proporcionado acesso a créditos em plataformas de cloud computing
- Consórcios interinstitucionais como o C4AI (Centro de IA) facilitam o compartilhamento de recursos e conhecimento



Computação de Borda para Implantação

Para viabilizar a aplicação de MLMs em contextos com conectividade limitada, comum em regiões remotas do Brasil:

- A UFPE desenvolveu técnicas de destilação de conhecimento para criar versões compactas de MLMs capazes de rodar em dispositivos de borda
- Pesquisadores da UFABC exploraram quantização e pruning para otimizar modelos para hardware restrito
- A UNICAMP implementou soluções de MLMs offline para aplicações em áreas rurais com conectividade instável

A infraestrutura computacional representa um desafio significativo para pesquisas em MLMs no Brasil, exigindo abordagens inovadoras para maximizar recursos limitados. Estas restrições têm estimulado contribuições significativas em áreas como eficiência computacional, paralelização e otimização de modelos, com potencial impacto global.

Análise Comparativa: BERT vs GPT

Diferenças de Treinamento e Inferência

Objetivo de Treinamento	Prever tokens mascarados com contexto bidirecional	Prever próximo token com contexto unidirecional
Arquitetura Base	Apenas encoder Transformer	Apenas decoder Transformer
Atenção	Bidirecional (completa)	Unidirecional (mascarada)
Treinamento	Paralelizável (tokens independentes)	Sequencial (dependência autoregressiva)
Paradigma de Uso	Fine-tuning para tarefas específicas	Geração de texto via prompting

Pesquisadores da USP compararam sistematicamente o desempenho de modelos baseados em BERT e GPT em tarefas em português brasileiro, encontrando padrões de superioridade dependentes da natureza da tarefa.

Contextos de Uso Recomendados

1 MLMs (BERT) são recomendados para:

- Classificação de textos (ex: análise de sentimento, categorização)
- Extração de informações e reconhecimento de entidades
- Respostas a perguntas em contextos definidos
- Tarefas que exigem compreensão semântica profunda

2 Modelos Causais (GPT) são preferíveis para:

- Geração de texto criativo ou fluente
- Completamento de frases e parágrafos
- Diálogo interativo e chatbots conversacionais
- Tarefas que envolvem produção de linguagem natural

Estudos da UFMG e UNICAMP demonstraram que abordagens híbridas, combinando as forças de ambos os tipos de modelos, apresentam resultados promissores para tarefas complexas em português brasileiro, como resposta a perguntas abertas e sumarização extrativa-abstrativa.

Papers Chave em Repositórios Brasileiros

Universidade de São Paulo (USP)

O repositório da USP contém contribuições fundamentais para MLMs em português:

- "BERTimbau: Pretrained BERT Models for Brazilian Portuguese" (Souza et al., 2020) - documentando o desenvolvimento do primeiro BERT específico para português brasileiro
- "Avaliação de Embeddings Contextuais para o Português" (Rodrigues et al., 2019) - comparação sistemática de representações contextuais
- "PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data" (Carmo et al., 2020) - adaptação do T5 para português

Universidade Federal de Minas Gerais (UFMG)

A UFMG contribuiu significativamente com pesquisas sobre aplicações e avaliação de MLMs:

- "Análise de Viés em Modelos Pré-treinados para o Português" (Pereira et al., 2021) - estudo sobre vieses em MLMs
- "BERTtimiza: Otimização de Hiperparâmetros para Fine-tuning de BERT em Português" (Silva et al., 2020) - técnicas específicas para ajuste fino
- "Representações Contextuais para Classificação de Textos Jurídicos" (Oliveira et al., 2022) - aplicações em domínio específico

Universidade Estadual de Campinas (UNICAMP)

O repositório da UNICAMP contém importantes contribuições metodológicas:

- "Explorando Arquiteturas Eficientes de Transformer para o Português" (Almeida et al., 2021) - análise de eficiência computacional
- "Avaliação de Robustez de MLMs em Tarefas de NLI para o Português" (Santos et al., 2022) - testes de robustez e adversariais
- "Modelos Multimodais Texto-Imagem para o Português Brasileiro" (Lima et al., 2023) - expansão para processamento multimodal

Estes repositórios acadêmicos representam um valioso recurso para pesquisadores interessados em MLMs para o português brasileiro, documentando não apenas resultados finais mas também metodologias, desafios e lições aprendidas no processo de adaptação e desenvolvimento destas tecnologias para o contexto linguístico brasileiro.

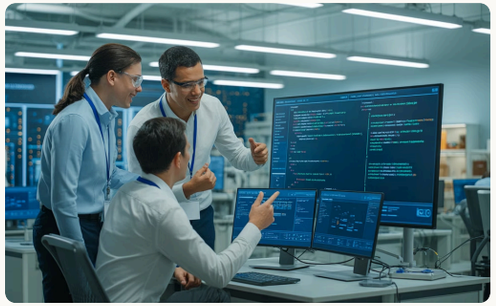
Muitos destes trabalhos destacam particularidades do desenvolvimento de MLMs para o português, como a necessidade de vocabulários específicos para capturar eficientemente a morfologia rica da língua, estratégias de avaliação que considerem variação dialetal, e técnicas de pré-processamento adaptadas para características específicas do português brasileiro.

Projetos Acadêmicos em Universidades Federais

UFABC: Modelos Compactos e Eficientes

A Universidade Federal do ABC destaca-se por pesquisas em modelos linguísticos otimizados:

- Projeto "LightBERT-PT": desenvolvimento de versões compactas de BERT para português
- Iniciativa "MLMs na Borda": adaptação de modelos para dispositivos com recursos limitados
- Laboratório de Processamento de Linguagem Natural: técnicas de destilação e quantização



UFPE: Aplicações Regionais e Sociais

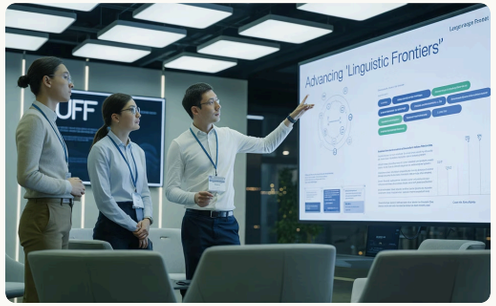
A Universidade Federal de Pernambuco tem focado em aplicações com impacto social:

- Centro de Informática: adaptação de MLMs para português nordestino
- Projeto "NLP para Inclusão": desenvolvimento de ferramentas de acessibilidade baseadas em MLMs
- Observatório de Mídias Sociais: análise de discurso em redes sociais brasileiras

UFRJ: Avaliação e Benchmarking

A Universidade Federal do Rio de Janeiro contribui com frameworks de avaliação:

- PESC/COPPE: desenvolvimento do benchmark GLUE-PT para português
- Laboratório de PLN: metodologias para avaliação de robustez e fairness
- Projeto "Tradutor Contextual": modelos para tradução específica de domínio



UFF: Aplicações em Saúde e Educação

A Universidade Federal Fluminense foca em aplicações específicas:

- Instituto de Computação: MLMs para processamento de registros médicos
- Projeto "NLP Educacional": sistemas tutoriais inteligentes baseados em MLMs
- Laboratório de Inteligência Computacional: modelos para análise de dados de saúde pública

Estes projetos acadêmicos não apenas avançam o estado da arte em MLMs para o português brasileiro, mas também formam uma nova geração de pesquisadores especializados nesta área. A diversidade de abordagens e aplicações reflete a riqueza do ecossistema de pesquisa em PLN no Brasil, com cada instituição contribuindo com perspectivas únicas baseadas em suas áreas de expertise e contextos regionais.

Benchmarking de Modelos em Português

Resultados Práticos em Bancos de Dados Nacionais

A avaliação sistemática de MLMs em datasets específicos para português brasileiro revela padrões interessantes de desempenho:

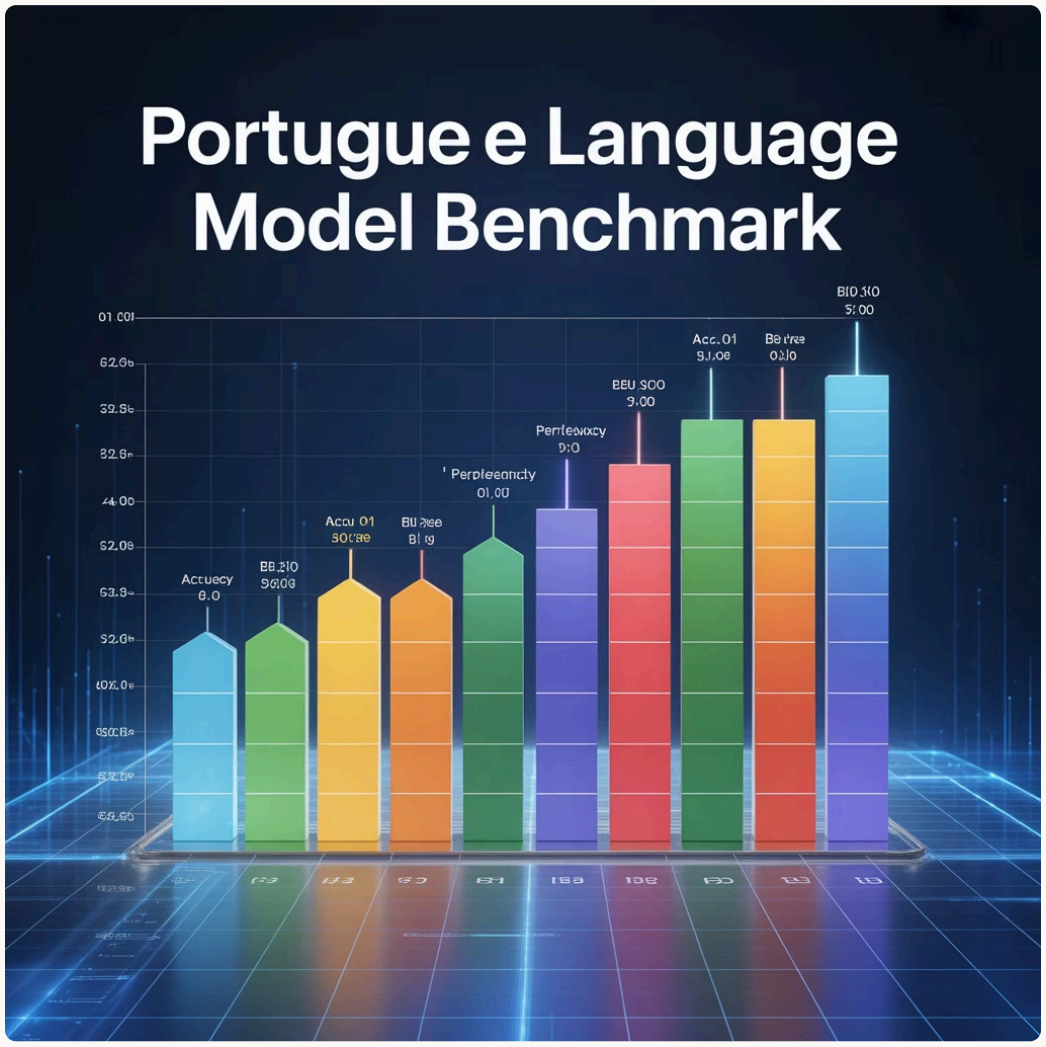
mBERT	82.3%	78.5%	71.2%
BERTimbau-base	86.7%	83.4%	75.8%
BERTimbau-large	88.9%	85.2%	77.5%
PTT5-base	87.3%	82.8%	76.3%
XLM-R	85.1%	81.7%	74.9%

Estes resultados, compilados por pesquisadores da UFRJ e USP, demonstram consistentemente a vantagem de modelos especificamente treinados para português brasileiro (como BERTimbau) sobre alternativas multilíngues em tarefas diversas.

Open Datasets do Brasil

Diversos conjuntos de dados públicos têm sido desenvolvidos para facilitar a avaliação de MLMs em português brasileiro:

- **ASSIN**: Avaliação de Similaridade Semântica e Inferência Textual, desenvolvido pela USP
- **FaQCS**: Dataset para classificação de perguntas em português, pela UFMG
- **CorPop**: Corpus de linguagem popular brasileira em redes sociais, compilado pela UFPE
- **LeNER-Br**: Dataset para reconhecimento de entidades nomeadas, pela UNICAMP
- **TweetSentBR**: Corpus de tweets em português com anotação de sentimento, pela UFF
- **OlimpoBR**: Dataset para detecção de discurso de ódio em português, pela UFABC



A disponibilidade crescente destes recursos específicos para português brasileiro tem impulsionado tanto o desenvolvimento quanto a avaliação rigorosa de MLMs adaptados para esta língua. Um desafio contínuo, identificado por pesquisadores da UNICAMP, é a necessidade de datasets que representem adequadamente a diversidade regional e socioeconômica do português brasileiro.

Customizando um MLM para Domínio Específico

Análise de Requisitos do Domínio

O processo de adaptação começa com uma análise detalhada das características linguísticas específicas do domínio alvo. Pesquisadores da USP desenvolveram metodologias para identificação de particularidades terminológicas e estruturais em textos jurídicos brasileiros, incluindo latinismos e construções sintáticas específicas do direito nacional.

Compilação de Corpus Especializado

A aquisição e curadoria de dados representativos é crucial. A UFMG estabeleceu protocolos para compilação de corpora jurídicos a partir de repositórios como o STF e STJ, enquanto a UNICAMP desenvolveu metodologias para textos médicos brasileiros, considerando questões de privacidade e terminologia específica do SUS.

Adaptação do Vocabulário

O ajuste do vocabulário para incluir terminologia especializada é fundamental. Pesquisadores da UFRJ implementaram técnicas para integração eficiente de termos técnicos no vocabulário de tokenização, preservando cobertura de linguagem geral enquanto acomoda terminologia específica de engenharia e ciências.

Continued Pre-training

O treinamento continuado em dados do domínio específico permite adaptação eficiente. A UFPE estabeleceu parâmetros ótimos para continued pre-training em diferentes domínios, enquanto a UFABC desenvolveu técnicas de treinamento eficientes para recursos computacionais limitados.

Avaliação em Tarefas Específicas

A validação do modelo adaptado em tarefas representativas do domínio é essencial. A UFF desenvolveu benchmarks específicos para avaliação de MLMs em textos educacionais brasileiros, enquanto a USP criou frameworks de avaliação para domínio jurídico.

A customização de MLMs para domínios específicos tem se mostrado particularmente valiosa no contexto brasileiro, onde terminologias técnicas e estruturas linguísticas especializadas podem divergir significativamente de padrões internacionais. Por exemplo, o vocabulário jurídico brasileiro contém não apenas termos em português, mas também expressões latinas e construções específicas do sistema legal nacional.

Estudos da UNICAMP demonstraram que modelos customizados para domínios específicos podem superar modelos genéricos em até 25% em tarefas especializadas, justificando o investimento em adaptação mesmo com recursos computacionais limitados.

Mitos e Verdades dos MLMs

Mito: MLMs Compreendem Linguagem como Humanos

Embora MLMs demonstrem capacidades impressionantes, sua "compreensão" difere fundamentalmente da cognição humana. Pesquisadores da USP demonstraram que BERTimbau, apesar de alto desempenho em tarefas de PLN, falha em capturar nuances pragmáticas e pressuposições culturais específicas do contexto brasileiro.

Experimentos da UNICAMP revelaram que MLMs podem atingir altas pontuações em benchmarks enquanto ainda falham em tarefas que exigem raciocínio causal ou compreensão de convenções sociais brasileiras não explicitamente mencionadas no texto.

Verdade: MLMs Capturam Padrões Linguísticos Complexos

MLMs efetivamente capturam padrões estatísticos sofisticados em dados linguísticos. A UFMG comprovou que modelos como BERTimbau identificam corretamente dependências sintáticas de longa distância em português, incluindo concordância entre elementos separados por múltiplas orações.

Estudos da UFRJ demonstraram que MLMs pré-treinados em português brasileiro adquirem representações que refletem propriedades morfológicas complexas como conjugações verbais irregulares e formação de plural, mesmo para casos raros.

Mito: MLMs São Igualmente Eficientes em Todas as Tarefas

Existe uma variação significativa no desempenho de MLMs dependendo da tarefa. Pesquisadores da UFPE documentaram que MLMs para português brasileiro apresentam desempenho inconsistente entre diferentes tipos de tarefas, com particular dificuldade em inferência pragmática e resolução de correferência em textos complexos.

A UFABC identificou que mesmo MLMs robustos como BERTimbau-large podem apresentar degradação significativa quando confrontados com textos contendo dialetos regionais brasileiros ou linguagem altamente coloquial não representada nos dados de treinamento.

Verdade: Modelos Específicos para Português Superam Multilíngues

Evidências consistentes confirmam que modelos treinados especificamente para português brasileiro superam modelos multilíngues em tarefas nacionais. A USP documentou ganhos de 5-15% em diversas tarefas ao utilizar BERTimbau em comparação com mBERT, especialmente em domínios com terminologia específica brasileira.

Estudos comparativos da UNICAMP demonstraram que esta vantagem se mantém mesmo quando modelos multilíngues são fine-tuned com dados em português, sugerindo que o pré-treinamento específico para a língua captura nuances linguísticas fundamentais.

Compreender as capacidades reais e limitações dos MLMs é essencial para sua aplicação efetiva. Pesquisadores brasileiros têm contribuído significativamente para desmistificar estas tecnologias, proporcionando avaliações rigorosas e contextualmente relevantes para o cenário nacional.

Requisitos Computacionais

Estimativa de Tempo e Hardware para Treinar de Raiz

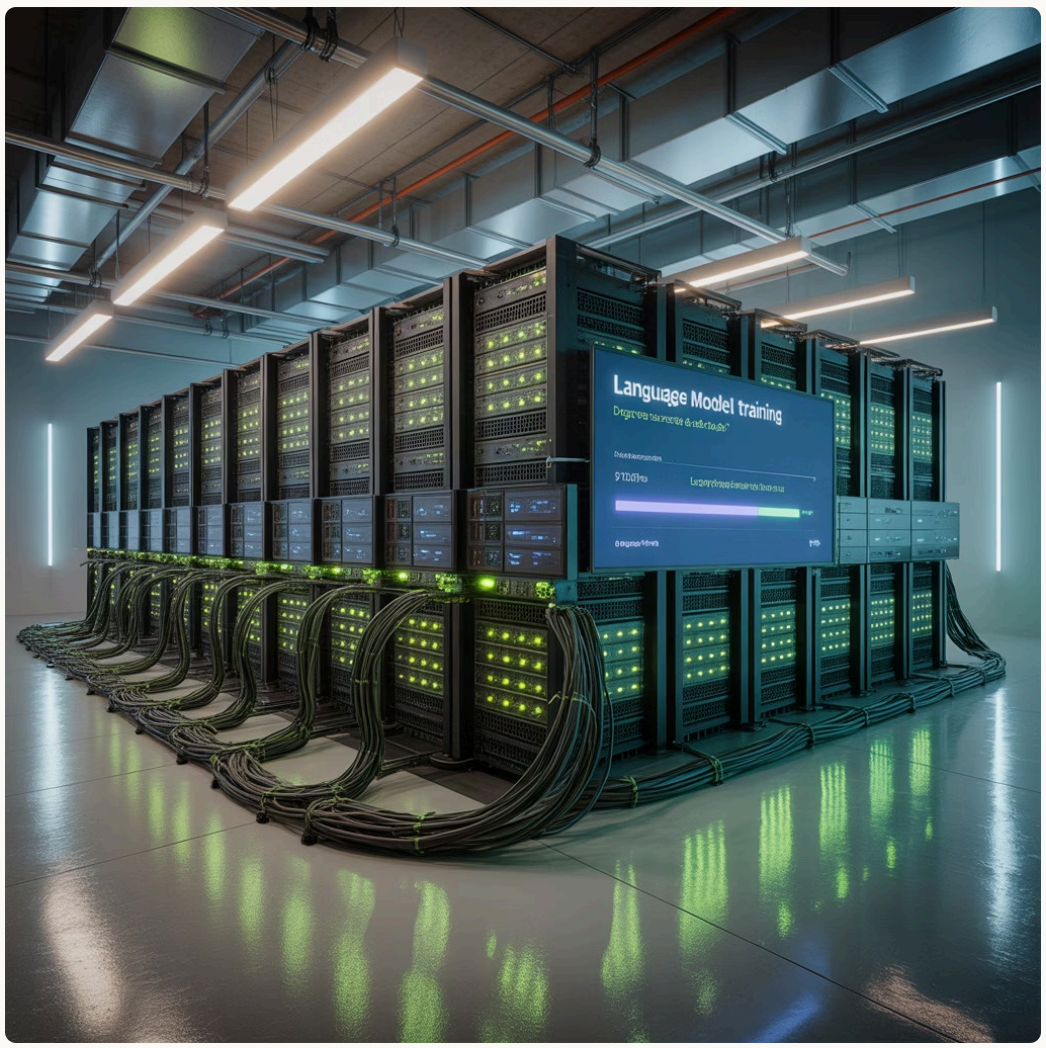
O treinamento de MLMs de raiz requer recursos computacionais substanciais. Pesquisadores da USP e UNICAMP documentaram os requisitos típicos para modelos em português brasileiro:

BERT-base PT (110M)	8x V100	3-4	15-20K
BERT-large PT (340M)	16x V100	5-7	40-60K
RoBERTa-base PT (125M)	8x V100	4-6	20-30K
T5-base PT (220M)	16x V100	7-10	50-80K
ALBERT-base PT (12M)	4x V100	2-3	8-12K

Alternativas Viáveis para Recursos Limitados

Reconhecendo as restrições de recursos enfrentadas por muitas instituições brasileiras, pesquisadores têm desenvolvido alternativas mais acessíveis:

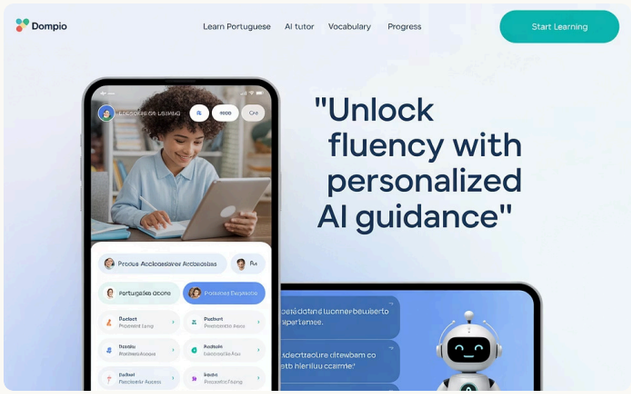
- **Continued Pre-training**: A UFABC demonstrou que adaptar modelos existentes com continued pre-training em dados específicos pode reduzir requisitos computacionais em até 90% com perda mínima de desempenho
- **Treinamento Distribuído**: A UFPE implementou protocolos para distribuir treinamento entre múltiplos clusters heterogêneos, comuns em universidades brasileiras
- **Arquiteturas Eficientes**: Pesquisadores da UFRJ adaptaram modelos como ALBERT para português, reduzindo significativamente requisitos de hardware
- **Pruning e Quantização**: A UFMG desenvolveu técnicas para reduzir tamanho e requisitos computacionais de modelos pré-treinados

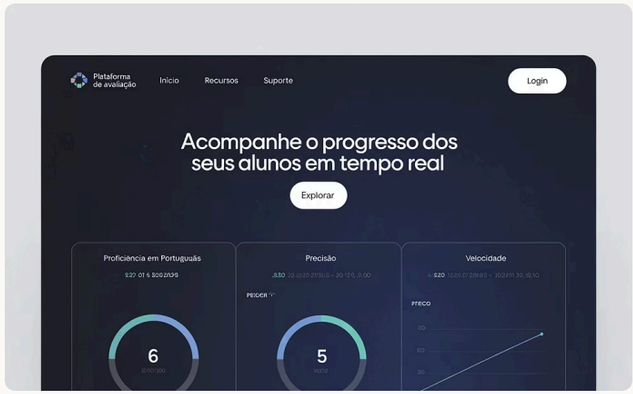


A questão dos recursos computacionais representa um desafio significativo para pesquisadores brasileiros em MLMs, estimulando inovações em eficiência e otimização. Colaborações entre universidades e parcerias com a indústria têm sido estratégias importantes para viabilizar projetos ambiciosos nesta área.

A experiência brasileira em desenvolvimento de MLMs com restrições de recursos tem produzido conhecimentos valiosos sobre otimização de modelos, com potencial aplicação em outros contextos onde recursos computacionais avançados são limitados.

MLMs para Ensino de Idiomas





Assistentes de Escrita e Gramática

MLMs têm revolucionado o desenvolvimento de ferramentas para aprendizagem de português. Pesquisadores da USP desenvolveram o "GramatiBot", um assistente baseado em BERTimbau que identifica e explica erros gramaticais contextualizados, especialmente útil para estudantes estrangeiros aprendendo português brasileiro.

A UNICAMP implementou um sistema de sugestões de escrita que considera variações regionais do português brasileiro, ajudando estudantes a compreender nuances dialetais e registros formais/informais.

Sistemas de Diálogo para Prática Conversacional

A prática conversacional é um componente crítico no aprendizado de idiomas. A UFPE desenvolveu "ConversaBR", um sistema de diálogo baseado em MLMs que simula conversações em português brasileiro sobre temas cotidianos, adaptando-se ao nível de proficiência do aprendiz.

Estudos realizados pela UFRJ demonstraram que estudantes utilizando sistemas conversacionais baseados em MLMs por 30 minutos diários apresentaram ganhos de fluência 40% superiores em comparação com métodos tradicionais.

Avaliação Automática de Proficiência

A avaliação precisa e consistente é fundamental no processo educacional. Pesquisadores da UFMG criaram um sistema baseado em MLMs para avaliar redações em português brasileiro, fornecendo feedback detalhado sobre coesão, coerência e adequação lexical.

A UFABC implementou uma ferramenta de avaliação de leitura que utiliza MLMs para gerar questões de compreensão personalizadas a partir de qualquer texto, adaptando-se automaticamente ao nível do estudante.

Estas aplicações educacionais dos MLMs têm encontrado adoção crescente em plataformas de ensino brasileiras, tanto para falantes nativos aprimorando habilidades de escrita formal quanto para estrangeiros aprendendo português como segunda língua. A capacidade dos MLMs de modelar nuances linguísticas específicas do português brasileiro representa uma vantagem significativa sobre abordagens tradicionais baseadas em regras.

Um desafio particular identificado por pesquisadores da UFF é a adaptação destes sistemas para diferentes variantes do português, considerando a rica diversidade linguística do Brasil. Projetos colaborativos entre UFPE e UFMG têm trabalhado no desenvolvimento de modelos que reconhecem e respeitam esta diversidade no contexto educacional.

Desafios Atuais na Pesquisa

Contexto Longo e Multi-turn Dialogue

O processamento eficiente de textos longos e manutenção de contexto em conversações extensas permanecem desafios significativos para MLMs:

- Pesquisadores da USP têm explorado adaptações de arquiteturas como Longformer e BigBird para processamento eficiente de documentos jurídicos brasileiros, que frequentemente excedem os limites contextuais de modelos padrão
- A UFRJ desenvolveu técnicas de sumarização hierárquica para preservar informações relevantes em diálogos extensos em português
- A UNICAMP propôs métodos de atenção esparsa adaptados para estruturas documentais específicas do contexto brasileiro, como processos judiciais e laudos médicos

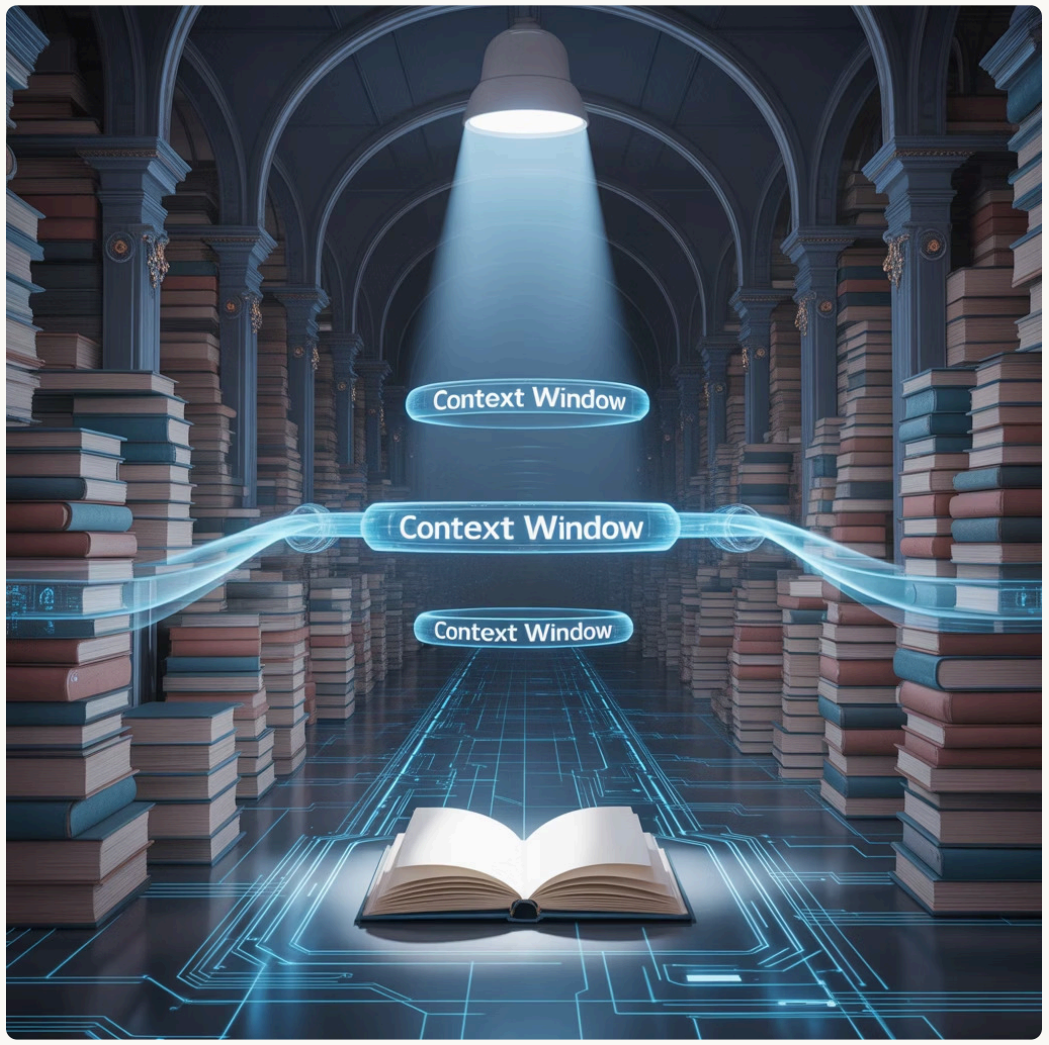
Estas limitações são particularmente relevantes no contexto brasileiro, onde documentos oficiais e processos administrativos tendem a ser extensos e formais, frequentemente ultrapassando as capacidades contextuais de MLMs padrão.

Eficiência Energética em ML

A sustentabilidade computacional tem emergido como uma preocupação crítica na pesquisa de MLMs:

- O Laboratório de Computação Sustentável da UFABC tem investigado o impacto ambiental do treinamento e inferência de MLMs no contexto brasileiro
- Pesquisadores da UFPE desenvolveram frameworks para medição e otimização do consumo energético em aplicações de PLN
- A UFMG implementou técnicas de quantização e pruning que reduzem significativamente o consumo energético durante inferência, com aplicações em dispositivos móveis
- A USP estabeleceu diretrizes para reporting de eficiência energética em pesquisas de PLN, promovendo transparência e conscientização

Esta questão é particularmente relevante no Brasil, onde o acesso a infraestrutura computacional avançada é limitado e considerações de custo energético podem determinar a viabilidade de implementações em larga escala.



> "O desenvolvimento de MLMs eficientes e sustentáveis não é apenas uma questão técnica, mas também de equidade no acesso a tecnologias linguísticas avançadas. Modelos que requerem infraestrutura massiva perpetuam desigualdades no desenvolvimento tecnológico global." — André Carvalho, pesquisador da USP

Perspectivas Futuras

Modelos Maiores e Mais Eficientes

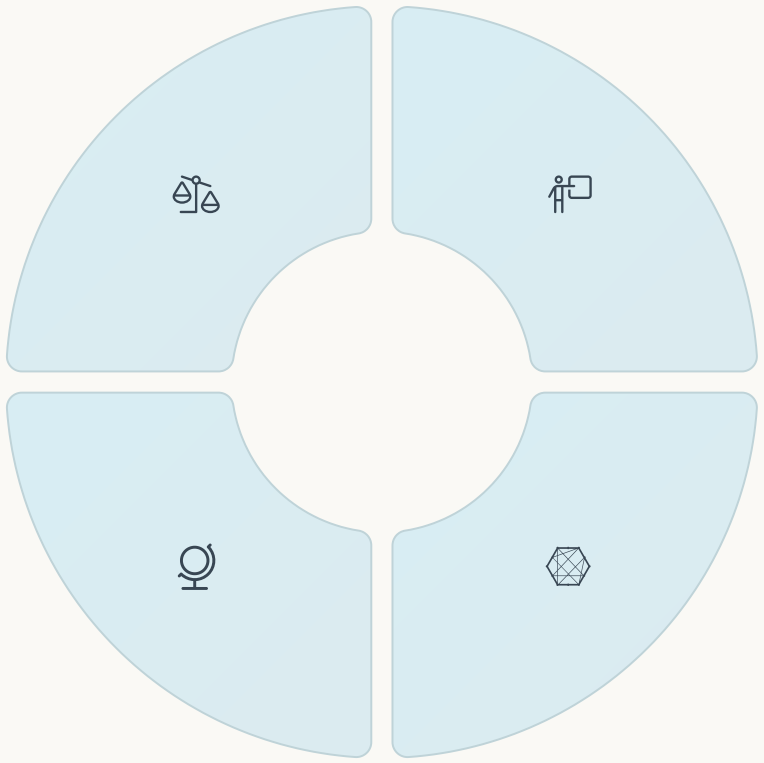
A tendência de escalamento continua, mas com foco crescente em eficiência:

- Pesquisadores da USP e UNICAMP estão colaborando no desenvolvimento de um modelo de grande escala especificamente para português brasileiro
- A UFABC lidera pesquisas em técnicas de scaling eficiente, como mixture-of-experts, para viabilizar modelos maiores com recursos limitados
- Estudos da UFMG investigam o trade-off entre tamanho e especificidade linguística em MLMs para português

Modelos Culturalmente Contextualizados

Cresce o foco em modelos que compreendam referências culturais específicas:

- A UFPE lidera iniciativas para incorporar conhecimento cultural brasileiro em MLMs
- Pesquisadores da UFF desenvolvem datasets para avaliar compreensão de referências culturais regionais
- A UFABC explora métodos para representar adequadamente diversidade linguística e cultural brasileira



Instruções Contextuais

A adaptação de modelos através de instruções em linguagem natural ganha relevância:

- A UFRJ desenvolve técnicas de instruction tuning específicas para português
- Pesquisadores da UFPE exploram a eficácia de prompts em português brasileiro para orientar comportamento de modelos
- A UFF investiga métodos para incorporar conhecimento cultural brasileiro através de instruções contextualizadas

Interface com LLMs

A integração entre MLMs e Large Language Models oferece novas possibilidades:

- A USP explora arquiteturas híbridas que combinam compreensão bidirecional dos MLMs com capacidades generativas dos LLMs
- Pesquisadores da UNICAMP desenvolvem métodos de transferência de conhecimento entre MLMs especializados e LLMs generalistas
- A UFMG investiga a aplicação de MLMs como verificadores de precisão factual para conteúdo gerado por LLMs

O futuro dos MLMs no contexto brasileiro aponta para modelos mais especializados e culturalmente conscientes, capazes de compreender nuances linguísticas regionais e referências contextuais específicas. Ao mesmo tempo, a pressão por eficiência computacional e sustentabilidade continuará impulsionando inovações em arquiteturas e métodos de treinamento.

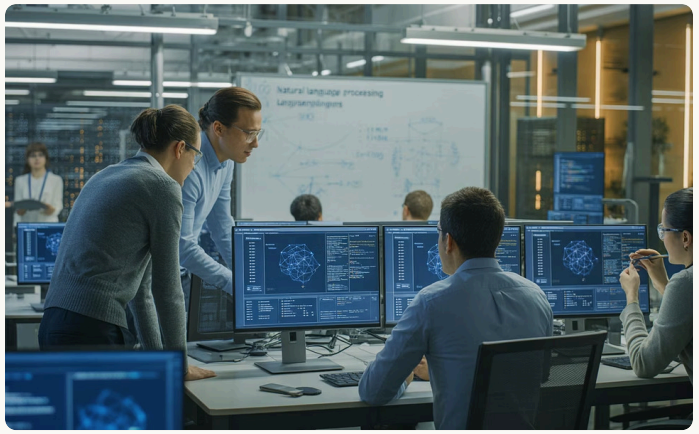
A colaboração entre instituições nacionais e intercâmbio com centros internacionais de pesquisa será fundamental para o avanço contínuo desta área no Brasil, permitindo o desenvolvimento de tecnologias linguísticas que atendam às necessidades específicas da sociedade brasileira.

Iniciativas Abertas e Comunidades de Pesquisa

Grupos de Pesquisa na USP

A Universidade de São Paulo abriga importantes centros de pesquisa em MLMs e PLN:

- O Núcleo Interinstitucional de Linguística Computacional (NILC), coordenado por Sandra Aluísio e Thiago Pardo, desenvolve recursos linguísticos e modelos para português brasileiro
- O Centro de Inteligência Artificial (C4AI), parceria com IBM e FAPESP, mantém linhas de pesquisa em processamento de linguagem natural
- O Grupo de Processamento de Linguagem Natural do IME-USP, liderado por Marcelo Finger, foca em aspectos formais e computacionais de MLMs



Laboratórios na UFMG

A Universidade Federal de Minas Gerais contribui significativamente através de vários grupos:

- O Laboratório MINDS (Machine Intelligence and Data Science), coordenado por Adriano Veloso e Wagner Meira Jr., desenvolve pesquisas em representações contextuais e aplicações de MLMs
- O Grupo de Processamento de Linguagem Natural, liderado por Renato Rocha Souza, investiga adaptações de MLMs para domínios específicos
- O Laboratório de Inteligência Computacional (LINC) mantém projetos em análise de sentimento e processamento multilíngue

Iniciativas na UNICAMP e UFRJ

Outras instituições mantêm centros de excelência em pesquisa sobre MLMs:

- O Instituto de Computação da UNICAMP, através do grupo liderado por Anderson Rocha e Ariadne Carvalho, foca em eficiência computacional e MLMs multimodais
- O PESC/COPPE da UFRJ, com pesquisadores como Leila Cristina Carneiro Bergamasco, desenvolve benchmarks e avaliação sistemática de modelos
- O Laboratório de PLN da UFRJ, coordenado por Celso Vieira, trabalha em aplicações para processamento de textos científicos e acadêmicos

Estas iniciativas frequentemente mantêm repositórios abertos e disponibilizam recursos como modelos pré-treinados, datasets e ferramentas para a comunidade. Exemplos incluem o BERTimbau (disponibilizado pelo NILC/USP), o corpus BrWaC (compilado pela UFSC em colaboração com outras instituições) e o benchmark GLUE-PT (desenvolvido pela UFRJ).

A colaboração interinstitucional tem sido um fator determinante para o avanço da pesquisa em MLMs no Brasil, permitindo a junção de recursos e expertise para enfrentar desafios que seriam intransponíveis para instituições isoladas. Redes como a BrAI (Brazilian Artificial Intelligence Research) facilitam esta integração entre grupos de pesquisa nacionais.

Ferramentas e Frameworks Populares

Plataformas e Bibliotecas para MLMs

Pesquisadores brasileiros utilizam e contribuem para diversos frameworks específicos para desenvolvimento e aplicação de MLMs:

- **HuggingFace Transformers**: A biblioteca mais utilizada para acesso a modelos pré-treinados e implementação de fine-tuning. Pesquisadores da USP e UNICAMP contribuíram com modelos específicos para português brasileiro no hub público
- **PyTorch**: Framework de deep learning preferido para pesquisa em MLMs no Brasil, com ampla adoção em universidades como UFMG, USP e UNICAMP
- **TensorFlow**: Utilizado especialmente em projetos com foco em implantação industrial, comum em colaborações entre academia e empresas
- **Fairseq**: Framework do Facebook AI adotado por pesquisadores da UFRJ e UFPE para experimentos com arquiteturas avançadas

Grupos de pesquisa brasileiros frequentemente desenvolvem extensões e adaptações destes frameworks para atender necessidades específicas do processamento de português brasileiro.

Ferramentas Desenvolvidas no Brasil

Além de utilizar frameworks internacionais, a comunidade brasileira tem desenvolvido ferramentas específicas:

- **NLPortBR**: Toolkit desenvolvido pela USP para pré-processamento específico de português brasileiro, incorporando normalizadores para variantes regionais e gírias
- **BERTEval**: Sistema da UFMG para avaliação sistemática e visualização de desempenho de MLMs em datasets brasileiros
- **PTViz**: Ferramenta da UNICAMP para visualização e interpretação de atenção em modelos como BERTimbau
- **CorpusBuilder**: Framework da UFRJ para compilação eficiente de corpora específicos para domínios técnicos em português

Estas ferramentas são frequentemente disponibilizadas como software de código aberto, facilitando adoção por outros pesquisadores e contribuindo para o ecossistema nacional de PLN.

A disponibilidade de ferramentas adaptadas para processamento de português brasileiro tem acelerado significativamente a pesquisa e aplicação de MLMs no contexto nacional. Muitas destas ferramentas endereçam desafios específicos como tokenização eficiente de contrações, tratamento de acentuação e processamento de estruturas sintáticas particulares do português.

Iniciativas de documentação em português e tutoriais adaptados ao contexto brasileiro, como os desenvolvidos pela UFABC e UFPE, têm facilitado a adoção destas tecnologias por novos pesquisadores e profissionais, contribuindo para a expansão da comunidade nacional de PLN.

Pipeline Completo: Do Dado Bruto ao Modelo

Coleta e Curadoria de Dados

O processo inicia com aquisição de corpus representativo. Pesquisadores da USP desenvolveram metodologias para coleta balanceada de textos em português brasileiro, considerando diversidade regional, registros formais/informais e domínios temáticos. A UFMG estabeleceu protocolos para filtragem de conteúdo problemático e balanceamento de fontes.

Pré-processamento e Normalização

A preparação adequada dos dados é crucial. A UNICAMP implementou pipelines específicos para normalização de texto em português brasileiro, tratando questões como variação ortográfica, contrações e caracteres especiais. A UFRJ desenvolveu ferramentas para segmentação eficiente e tokenização adaptada às particularidades morfológicas do português.

Tokenização e Construção de Vocabulário

A definição adequada de unidades lexicais impacta todo o processo. Pesquisadores da UFPE compararam algoritmos de tokenização para português, demonstrando vantagens de abordagens como BPE para capturar morfologia complexa. A UFABC otimizou parâmetros de vocabulário para balanço entre cobertura e eficiência computacional.

Configuração e Treinamento

O treinamento eficiente requer configuração cuidadosa. A USP estabeleceu parâmetros ótimos para treinamento de MLMs em português brasileiro, incluindo learning rates, tamanhos de batch e estratégias de mascaramento. A UFMG desenvolveu técnicas de monitoramento para detecção precoce de problemas como memorização excessiva ou underfitting.

Avaliação e Refinamento

A avaliação rigorosa guia melhorias incrementais. A UFRJ implementou protocolos de avaliação em múltiplas dimensões, incluindo desempenho em benchmarks, eficiência computacional e robustez a variações linguísticas. A UFF desenvolveu metodologias para identificação e correção de vieses específicos do contexto brasileiro.

Implantação e Monitoramento

A transição para ambientes produtivos requer considerações adicionais. A UFABC estabeleceu práticas para otimização de inferência em diferentes plataformas, desde servidores até dispositivos móveis. A UNICAMP desenvolveu frameworks para monitoramento contínuo de desempenho e detecção de degradação em aplicações reais.

Este pipeline completo reflete a maturidade crescente da pesquisa brasileira em MLMs, com contribuições significativas em cada etapa do processo. A experiência acumulada por instituições nacionais tem permitido o desenvolvimento de modelos cada vez mais sofisticados e adaptados para as particularidades do português brasileiro e contextos de aplicação locais.

Exercício Prático: Treinando BERT com Dados Locais

Passos Essenciais para Laboratório

Este exercício prático, desenvolvido por pesquisadores da USP e UFMG, proporciona uma introdução hands-on ao treinamento de MLMs com dados em português brasileiro:

1. **Preparação do Ambiente**: Configuração de ambiente Python com bibliotecas necessárias (PyTorch, Transformers, Datasets)
2. **Obtenção de Dados**: Download e preparação de corpus de treinamento (pode-se utilizar subconjuntos do BrWaC ou Corpus Brasileiro)
3. **Pré-processamento**: Limpeza, normalização e segmentação do texto em português
4. **Tokenização**: Utilização de tokenizador existente ou treinamento de novo tokenizador específico para o corpus
5. **Configuração do Modelo**: Definição de hiperparâmetros e arquitetura (pode-se iniciar de modelo existente como BERTimbau)
6. **Treinamento**: Execução do treinamento com monitoramento de métricas relevantes
7. **Avaliação**: Teste do modelo em tarefas específicas como classificação ou NER

```
# Exemplo de código para continued pre-training
import transformers
from transformers import BertTokenizer, BertForMaskedLM
from transformers import DataCollatorForLanguageModeling
from transformers import Trainer, TrainingArguments

# Carregar modelo e tokenizador pré-treinados
tokenizer = BertTokenizer.from_pretrained(
    "neuralmind/bert-base-portuguese-cased")
model = BertForMaskedLM.from_pretrained(
    "neuralmind/bert-base-portuguese-cased")

# Carregar e tokenizar dados
def tokenize_function(examples):
    return tokenizer(
        examples["text"],
        truncation=True,
        padding="max_length",
        max_length=128
    )

# Aplicar tokenização
tokenized_dataset = dataset.map(
    tokenize_function,
    batched=True,
    num_proc=4,
    remove_columns=["text"]
)

# Configurar data collator para mascaramento
data_collator = DataCollatorForLanguageModeling(
    tokenizer=tokenizer,
    mlm=True,
    mlm_probability=0.15
)

# Definir parâmetros de treinamento
training_args = TrainingArguments(
    output_dir="./results",
    overwrite_output_dir=True,
    num_train_epochs=3,
    per_device_train_batch_size=16,
    save_steps=10_000,
    save_total_limit=2,
)

# Inicializar trainer
trainer = Trainer(
    model=model,
    args=training_args,
    data_collator=data_collator,
    train_dataset=tokenized_dataset["train"],
)

# Iniciar treinamento
trainer.train()

# Salvar modelo
model.save_pretrained("./modelo-resultante")
tokenizer.save_pretrained("./modelo-resultante")
```

Este exercício foi aplicado com sucesso em disciplinas de pós-graduação na USP, UNICAMP e UFMG, proporcionando aos estudantes experiência prática com o ciclo completo de desenvolvimento de MLMs. Os participantes frequentemente adaptam o exercício para domínios específicos de interesse, como textos jurídicos, científicos ou literários em português brasileiro.

Recursos computacionais acessíveis, como Google Colab ou clusters universitários, são suficientes para este exercício quando se trabalha com subconjuntos menores de dados ou quando se realiza continued pre-training a partir de modelos existentes.

Analisando Resultados e Interpretando Erros

Técnicas de Troubleshooting

A análise sistemática de erros é fundamental para o aprimoramento de MLMs. Pesquisadores brasileiros desenvolveram metodologias específicas para diagnóstico de problemas comuns:

- **Análise de Perplexidade por Categoria**: A USP implementou ferramentas para avaliar perplexidade do modelo segmentada por categorias gramaticais, identificando classes de palavras problemáticas
- **Visualização de Atenção**: A UNICAMP adaptou ferramentas como BERTViz para análise de padrões de atenção em textos em português, revelando possíveis falhas de compreensão contextual
- **Probing Tasks**: A UFMG desenvolveu tarefas específicas para avaliar aspectos linguísticos como compreensão morfológica e sintática em português brasileiro
- **Análise de Erro por Similaridade**: A UFRJ criou métodos para agrupar erros similares, facilitando a identificação de padrões sistemáticos

Estas técnicas permitem identificar limitações específicas dos modelos e direcionar esforços de melhoria de forma eficiente.

Padrões Comuns de Erro em Português

Pesquisas conduzidas por instituições brasileiras identificaram padrões recorrentes de erro em MLMs para português:

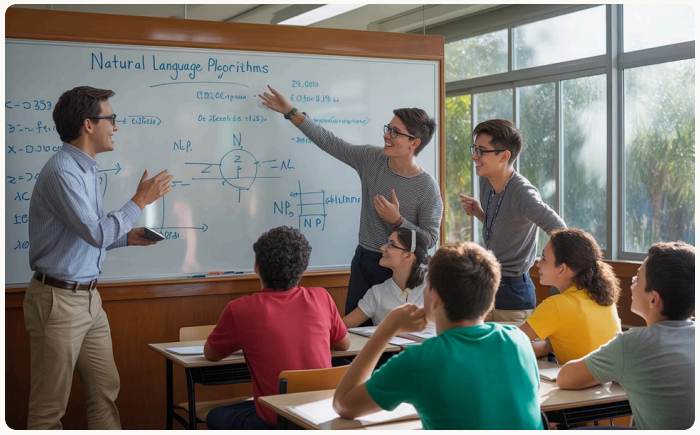
- **Ambiguidade Lexical**: Dificuldade com palavras polissêmicas frequentes em português, como "manga" (fruta/parte de roupa) e "vela" (iluminação/navegação)
- **Expressões Idiomáticas**: Falhas na compreensão de expressões como "dar com os burros n'água" ou "pegar o bonde andando"
- **Concordância de Longa Distância**: Problemas com concordância entre elementos separados por múltiplas orações subordinadas
- **Regionalismos**: Dificuldade com termos e construções específicas de variantes regionais do português brasileiro

O reconhecimento destes padrões tem orientado o desenvolvimento de datasets específicos para avaliação e estratégias de treinamento adaptadas.

A análise sistemática de erros não apenas aprimora modelos individuais, mas também contribui para a compreensão mais profunda de como MLMs representam e processam características específicas do português brasileiro. Pesquisadores da UFPE, por exemplo, utilizaram análise de erros para desenvolver insights sobre como modelos representam a rica morfologia verbal do português.

Estudos comparativos conduzidos pela USP e UNICAMP revelaram diferenças significativas nos padrões de erro entre modelos específicos para português brasileiro e adaptações de modelos multilíngues, reforçando a importância de recursos linguísticos específicos para a língua.

Projetos de Extensão em Universidades



Monitoria e Divulgação (USP/UFABC)

Iniciativas de extensão universitária têm democratizado o acesso ao conhecimento sobre MLMs:

- O programa "IA para Todos" da USP oferece oficinas sobre PLN e MLMs para estudantes de ensino médio de escolas públicas
- A UFABC mantém o projeto "Linguagem Computacional" que desenvolve materiais didáticos em português sobre MLMs acessíveis para público não-técnico
- Hackathons temáticos organizados pela USP promovem o desenvolvimento de aplicações baseadas em MLMs para resolver problemas locais

Transferência Tecnológica (UFMG/UFRJ)

A ponte entre pesquisa acadêmica e aplicações práticas é fortalecida por iniciativas como:

- O "NLP para PMEs" da UFMG oferece consultoria para pequenas e médias empresas interessadas em implementar soluções baseadas em MLMs
- O programa "Linguagem e Negócios" da UFRJ conecta pesquisadores a startups para desenvolvimento de aplicações inovadoras
- Workshops setoriais organizados pela UFMG disseminam conhecimento sobre aplicações de MLMs em áreas como saúde, jurídico e varejo

Preservação Linguística (UNICAMP/UFPE)

MLMs também contribuem para iniciativas de preservação cultural e linguística:

- O projeto "Línguas Brasileiras" da UNICAMP adapta técnicas de MLMs para documentação e processamento de línguas indígenas em risco
- A UFPE desenvolve recursos computacionais para variantes regionais do português brasileiro pouco representadas em corpora padrão
- Iniciativas colaborativas com comunidades tradicionais para registro e processamento de vocabulários específicos

Estes projetos de extensão demonstram como a pesquisa em MLMs transcende o ambiente acadêmico, impactando positivamente a sociedade através de educação, inovação empresarial e preservação cultural. Além disso, estas iniciativas frequentemente proporcionam feedback valioso para pesquisadores, identificando necessidades e desafios reais que podem orientar futuros desenvolvimentos.

A extensão universitária também desempenha papel fundamental na formação de uma nova geração de pesquisadores e profissionais, democratizando o acesso a conhecimentos avançados sobre PLN e MLMs e estimulando vocações em regiões e comunidades tradicionalmente sub-representadas neste campo.

Produção Científica Nacional (Exemplos)

Artigos e Monografias de Destaque: USP

A Universidade de São Paulo tem contribuído significativamente para a literatura sobre MLMs:

- "BERTimbau: A BERT-based Model for Brazilian Portuguese," por Souza et al. (2020), documentando o desenvolvimento e avaliação do primeiro modelo BERT específico para português brasileiro
- "Análise Semântica Distribucional em Português Brasileiro: Um Estudo Comparativo de Embeddings Contextuais," tese de doutorado de Fonseca (2021), comparando diferentes arquiteturas de MLMs para português
- "Adaptação de Modelos de Linguagem Mascarada para Domínios Técnicos em Português Brasileiro," dissertação de mestrado de Silva (2022), focando em adaptação para textos jurídicos e médicos

O NILC/USP mantém uma biblioteca digital que reúne publicações relacionadas a PLN em português, incluindo diversos trabalhos sobre MLMs desenvolvidos tanto na própria USP quanto em outras instituições brasileiras.

Contribuições da UFMG

Pesquisadores da Universidade Federal de Minas Gerais produziram trabalhos relevantes em aspectos específicos de MLMs:

- "Avaliação de Robustez de Modelos Pré-treinados para o Português Brasileiro," por Ferreira et al. (2022), investigando vulnerabilidades a ataques adversariais
- "Representações Contextuais para Detecção de Desinformação em Português," dissertação de mestrado de Almeida (2021), aplicando MLMs para combate a fake news
- "MLMs para Análise de Sentimento em Reviews de Produtos: Estudo Comparativo em Português Brasileiro," por Santos et al. (2023), comparando diferentes arquiteturas e estratégias de fine-tuning

O repositório institucional da UFMG oferece acesso a dezenas de trabalhos relacionados a MLMs, incluindo artigos, dissertações e teses desenvolvidas nos programas de pós-graduação em Ciência da Computação e Linguística.

A produção científica nacional sobre MLMs tem crescido significativamente nos últimos anos, com contribuições em diversos aspectos desde desenvolvimento de recursos linguísticos até aplicações inovadoras. Publicações brasileiras têm aparecido em conferências internacionais relevantes como ACL, EMNLP e NAACL, além de periódicos especializados.

Esta produção não apenas documenta avanços técnicos, mas também analisa criticamente as implicações sociais, éticas e culturais da aplicação destas tecnologias no contexto brasileiro, contribuindo para um desenvolvimento mais responsável e inclusivo de sistemas de PLN.

Competências Necessárias para Trabalhar com MLMs

Matemática

O desenvolvimento e aplicação de MLMs requer fundamentos sólidos em diversas áreas matemáticas:

- Álgebra Linear: essencial para compreender operações matriciais em redes neurais
- Cálculo Multivariável: fundamental para algoritmos de otimização e backpropagation
- Probabilidade e Estatística: base para modelagem estatística da linguagem
- Teoria da Informação: conceitos como entropia e divergência KL são amplamente utilizados

Cursos de graduação em matemática aplicada e estatística na USP e UNICAMP oferecem disciplinas específicas para aplicações em aprendizado de máquina e PLN.

Linguística

Conhecimentos linguísticos são fundamentais para o trabalho efetivo com MLMs:

- Morfologia e Sintaxe: compreensão da estrutura do português brasileiro
- Semântica e Pragmática: base para análise de significado e contexto
- Sociolinguística: entendimento de variações regionais e sociais da língua
- Linguística Computacional: interface entre linguística teórica e implementações

Programas interdisciplinares como o de Linguística Computacional da UFRJ e UFMG proporcionam formação específica nesta interseção.

Ciência da Computação

Habilidades técnicas específicas são necessárias para implementação e otimização:

- Programação (Python, PyTorch, TensorFlow): ferramentas essenciais para implementação
- Estruturas de Dados e Algoritmos: fundamentais para processamento eficiente
- Computação Paralela e Distribuída: necessária para treinamento de modelos grandes
- Cloud Computing: conhecimentos para utilização de infraestrutura escalável

Cursos de especialização em IA e PLN oferecidos por instituições como USP, UFPE e UFABC formam profissionais com estas competências específicas.

O desenvolvimento de profissionais capacitados para trabalhar com MLMs requer formação interdisciplinar, combinando conhecimentos técnicos com compreensão linguística e fundamentação matemática. Reconhecendo esta necessidade, universidades brasileiras têm desenvolvido programas específicos que integram estas áreas.

Além das competências técnicas, habilidades como pensamento crítico, ética aplicada e comunicação efetiva são cada vez mais valorizadas, refletindo a complexidade e o impacto social destas tecnologias. Iniciativas como o programa de ética em IA da USP e o observatório de tecnologias linguísticas da UFRJ contribuem para o desenvolvimento destas dimensões.

Desafios em Modelos Multilíngues

Modelos Cross-lingual Brasileiros

O desenvolvimento de modelos que integram português com outras línguas apresenta desafios e oportunidades específicas:

- Pesquisadores da USP investigaram transferência cross-lingual entre português e espanhol, demonstrando que proximidade linguística facilita compartilhamento de representações
- A UFRJ desenvolveu modelos especializados em português-inglês para aplicações de tradução técnica e científica
- A UFPE explorou a adaptação de modelos multilíngues para melhor desempenho em português, identificando estratégias de tokenização e treinamento que beneficiam particularmente línguas latinas
- A UFMG estudou a representação de conceitos culturais específicos em modelos multilíngues, identificando lacunas na transferência de conhecimento cultural

Estes estudos destacam tanto o potencial quanto as limitações de modelos multilíngues para aplicações em português brasileiro.

Particularidades e Desafios

A inclusão efetiva do português em modelos multilíngues enfrenta desafios específicos:

- **Desbalanceamento de Recursos**: A disponibilidade de dados em português frequentemente é menor que em inglês ou chinês, levando a representações menos robustas
- **Competição por Capacidade do Modelo**: Em arquiteturas com parâmetros compartilhados, línguas dominantes podem receber atenção desproporcional
- **Particularidades Linguísticas**: Aspectos como contrações, pronomes clíticos e concordância de gênero complexa em português requerem capacidade específica do modelo
- **Variação Dialetal**: Diferenças entre português brasileiro e europeu frequentemente não são adequadamente capturadas em modelos multilíngues

Pesquisadores da UNICAMP propuseram arquiteturas especializadas para modelos multilíngues que preservam melhor as particularidades de cada língua, utilizando mecanismos de atenção específicos por idioma que são ativados conforme necessário durante o processamento.

Estudos da UFF demonstraram que estratégias de balanceamento de corpus e técnicas de data augmentation podem melhorar significativamente a representação do português brasileiro em modelos multilíngues, reduzindo a disparidade de desempenho em relação a línguas com mais recursos.

> "O desenvolvimento de modelos verdadeiramente multilíngues, que tratam todas as línguas com equidade, é tanto um desafio técnico quanto uma questão de justiça linguística." — Marcos Zampieri, pesquisador colaborador da UFMG

Boas Práticas e Cuidados na Aplicação

Evitar Viés

MLMs absorvem e podem amplificar vieses presentes nos dados de treinamento. Pesquisadores da USP desenvolveram metodologias específicas para identificação e mitigação de vieses relacionados a gênero, raça e origem regional em modelos treinados em português brasileiro. A UFMG propôs técnicas de balanceamento de corpus para representação equitativa de diferentes variantes dialetais e registros linguísticos.

Prevenir Overfitting

O treinamento excessivo em dados limitados pode resultar em modelos que não generalizam adequadamente. A UNICAMP estabeleceu protocolos de validação cruzada específicos para datasets em português, considerando particularidades como variação regional. A UFPE desenvolveu técnicas de regularização adaptadas para características morfológicas do português brasileiro.

Garantir Uso Ético

Aplicações responsáveis de MLMs requerem consideração cuidadosa de implicações éticas. A UFRJ estabeleceu frameworks para avaliação de impacto social de aplicações de PLN no contexto brasileiro. A USP desenvolveu diretrizes específicas para aplicações em setores sensíveis como saúde, educação e segurança pública, considerando aspectos culturais e legais nacionais.

Documentar Limitações

Transparência sobre capacidades e restrições dos modelos é essencial. A UFABC propôs formatos de documentação para MLMs inspirados em model cards, adaptados para incluir informações específicas sobre desempenho em diferentes variantes do português. A UFMG desenvolveu metodologias para caracterização sistemática de domínios e registros linguísticos em que o modelo foi avaliado.

Considerações para o Contexto Brasileiro

Além das boas práticas gerais, o contexto brasileiro apresenta considerações específicas:

- Atenção à representatividade regional, evitando privilegiar variantes linguísticas de regiões específicas
- Consideração de aspectos culturais brasileiros que podem influenciar interpretação semântica
- Conformidade com a Lei Geral de Proteção de Dados (LGPD) brasileira
- Avaliação de impacto em comunidades tradicionalmente marginalizadas

Recursos para Desenvolvimento Responsável

Diversas instituições brasileiras disponibilizam recursos para apoiar aplicações éticas e responsáveis:

- O "Guia de PLN Ético" desenvolvido pela USP e UNICAMP
- O framework de avaliação de viés linguístico da UFMG
- A metodologia de auditoria de datasets da UFRJ
- Workshops e cursos de extensão sobre ética em IA oferecidos por diversas universidades federais

Referências Bibliográficas e Repositórios

Repositórios Universitários

- **USP**: https://www.teses.usp.br/ - Biblioteca Digital de Teses e Dissertações com extensa coleção em PLN e MLMs
- **UFMG**: https://repositorio.ufmg.br/ - Repositório Institucional com publicações sobre modelos linguísticos e aplicações
- **UNICAMP**: https://repositorio.unicamp.br/ - Acervo de pesquisas em computação linguística e PLN
- **UFRJ**: https://pantheon.ufrj.br/ - Repositório de publicações acadêmicas em PLN e ciência da computação
- **UFABC**: https://repositorio.ufabc.edu.br/ - Produções acadêmicas em IA e processamento de linguagem
- **UFPE**: https://repositorio.ufpe.br/ - Trabalhos em PLN com foco em aplicações regionais
- **UFF**: https://app.uff.br/riuff/ - Repositório institucional com pesquisas em linguística computacional

Recursos Adicionais

- **NILC/USP**: Repositório de Recursos - Corpora, léxicos e ferramentas para PLN em português
- **BERTimbau**: HuggingFace - Modelo BERT pré-treinado para português brasileiro
- **PTT5**: GitHub - Implementação do T5 para português brasileiro
- **BrWaC**: UFRGS - Corpus web brasileiro com bilhões de tokens
- **ASSIN**: GitHub - Dataset para avaliação de similaridade semântica e inferência textual
- **Portal NLP-PT**: NILC - Coletânea de recursos para PLN em português

Estes recursos representam uma amostra da riqueza de materiais disponíveis para pesquisa e desenvolvimento em MLMs para português brasileiro. A crescente disponibilidade de recursos abertos tem democratizado o acesso a estas tecnologias e estimulado inovação na comunidade brasileira de PLN.

As referências bibliográficas e repositórios listados constituem um ponto de partida valioso para pesquisadores, estudantes e profissionais interessados em MLMs para português brasileiro. A contínua expansão destes recursos, através de esforços colaborativos entre instituições nacionais, tem sido fundamental para o avanço da área no país.

O compartilhamento aberto de conhecimento, dados e implementações não apenas acelera o progresso científico, mas também promove equidade no desenvolvimento tecnológico, permitindo que instituições com recursos limitados participem ativamente da pesquisa de ponta em processamento de linguagem natural.

Referências complementares

Além das referências listadas acima, recomenda-se consultar os repositórios digitais das principais universidades brasileiras que mantêm acervos atualizados de teses, dissertações e artigos sobre IA:

- Biblioteca Digital de Teses e Dissertações da USP (www.teses.usp.br)
- Repositório Institucional da UFMG (repositorio.ufmg.br)
- Repositório Digital da Unicamp (repositorio.unicamp.br)
- Repositório Digital da UNIFESP (repositorio.unifesp.br)
- Biblioteca Digital da UFPE (repositorio.ufpe.br)
- Repositório Virtual da UFF (**https://app.uff.br/riuff/**)
- Lume - Repositório Digital da UFRGS (lume.ufrgs.br)
- Repositório Institucional da FGV (bibliotecadigital.fgv.br)
- Biblioteca Digital de Monografias da UFABC (biblioteca.ufabc.edu.br)

Para acompanhar os avanços mais recentes na pesquisa brasileira, recomenda-se também os anais das conferências BRACIS, SBIA, ENIAC e workshops especializados organizados pela Sociedade Brasileira de Computação (SBC).