COMPARATIVO DE MODELOS DE LINGUAGEM EM PORTUGUÊS DO BRASIL: UMA ANÁLISE DE GPT-2, GPORTUGUESE-2 E CABRITA-LORA-V0-1

Bruno Leonardo Santos Menezes[1]; Ricardo Gomes de Oliveira[2]; Raphael Souza de Oliveira[3]; Erick Giovani Sperandio Nascimento[4]

Resumo

Este artigo tem como objetivo realizar um comparativo entre três modelos de linguagem treinados a partir de corpora português do Brasil. Foram avaliados o *GPT-2*, *GPorTuguese-2* e *Cabrita-Lora-v0-1*, por intermédio das métricas *Sentence Textual Similarity* (STS) e *Recognizing Textual Entailment* (RTE). Atualmente existe uma carência de estudos que abordem este tema em modelos treinados em português. Sendo esta pesquisa uma contribuição nesta direção. Os resultados mostraram que *Cabrita-Lora-v0-1* obteve melhor desempenho em todas as métricas, existindo ainda oportunidades de melhorias e novos estudos em todos os modelos objeto desta pesquisa, pois os desempenhos podem ser afetados por configurações de hiper parâmetros. O comparativo elaborado por este estudo mostrou os pontos fortes e fracos de cada modelo e pode servir como base em novos direcionamentos que abordam este tema. Esta pesquisa pode representar uma contribuição significativa em Processamento de Linguagem Natural (NLP) em português.

Palavras-chave: Processamento de linguagem natural; modelos de linguagem; português do brasil; transformers.

Abstract

This paper presents a comparative analysis of three language models trained on Brazilian Portuguese corpora, namely GPT-2, GPorTuguese-2, and Cabrita-Lora-v0-1. These models were assessed using the Sentence Textual Similarity (STS) and Recognizing Textual Entailment (RTE) metrics. There is a notable scarcity of research focusing on this subject area for models trained in Portuguese, and this study aims to fill that gap. The results indicated that Cabrita-Lora-v0-1 outperformed the other models across all metrics. However, there remain opportunities for enhancement and additional research for all the models examined in this study, as performance can be influenced by hyperparameter configurations. The detailed comparison in this paper underscores the strengths and weaknesses of each model, providing valuable insights for future research in this domain. This work represents a meaningful contribution to Natural Language Processing (NLP) studies in Portuguese.

Keywords: Natural language processing; language models; brazilian portuguese; transformers.

[1] Doutor em Modelagem Computacional e Tecnologia Industrial pelo SENAI CIMATEC; professor da Fundação de Apoio à Escola Técnica do Estado do Rio de Janeiro. E-mail: bruno.menezes@faetecresende.com.

[2] Mestrando em Ciências da Computação pela Universidade Federal da Bahia-UFBA. E-mail: rgoliveira.ti@gmail.com.

[3]Doutorando em Modelagem Computacional e Tecnologia Industrial pelo SENAI CIMATEC. E-mail: raphael.oliveira@gmail.com.

[4] Doutor em Engenharia Ambiental pela Universidade Federal do Espírito Santo-UFES; professor da University of Surrey – UK e do SENAI CIMATEC. E-mail: erick.sperandio@surrey.ac.uk erick.sperandio@fieb.org.br.

1 Introdução

O processamento de linguagem natural (NLP) é um campo de estudo da inteligência artificial (IA) com diversas aplicações. Neste contexto, destaca-se os modelos de linguagem que são capazes de realizar tarefas como tradução, geração e classificação de textos por exemplo. Existe uma carência de estudos científicos que abordem modelos treinados em português do Brasil. Diante disso, este estudo tem como objetivo realizar um comparativo entre modelos de linguagem treinados com corpus em português do Brasil. Esta pesquisa visa contribuir com o aumento de estudos acadêmicos e científicos que tratam do tema NLP em nossa língua. Foram comparados três modelos: um modelo *GPT-2* (RADFORD, et. al., 2019) treinado a partir de corpora totalmente em português do Brasil, o *GPorTuguese-2* (*Portuguese GPT-2 small*), e o *Cabrita-Lora-v0-1*.

O *GPorTuguese-2* (https://huggingface.co/pierreguillou/gpt2-small-portuguese) é um modelo de linguagem, baseado no *GPT-2 small*. O modelo foi treinado com os dados da enciclopédia livre Wikipedia em português, no qual foram utilizadas técnicas de transferência de aprendizado e ajuste fino, em uma GPU NVIDIA V100 de 32GB e aproximadamente 1GB de dados de treinamento. *Cabrita-Lora-v0-1* (https://huggingface.co/22h/cabrita-lora-v0-1) é outro modelo de linguagem treinado também em português, ajuste fino realizado de um modelo LLaMA. Foi utilizado código disponível no Alpaca Lora, com uma A100 no Colab, baseado no LLaMA-7B e os dados de treinamento foram traduzidos para o português usando o ChatGPT.

Sentence Textual Similarity (STS) e *Recognizing Textual Entailment* (RTE), foram utilizadas como métricas para avaliar os três modelos com o *dataset* da segunda Avaliação de Similaridade Semântica e Inferência Textual (ASSIN2, disponível em https://sites.google.com/view/assin2/). Sendo possível realizar uma avaliação quantitativa da capacidade dos modelos de entender e gerar texto em português do Brasil. Esta pesquisa limitou-se a um comparativo de três modelos de linguagem, existindo a possibilidade de existirem outros modelos para realização de análises. A variabilidade na qualidade dos dados e hiper parâmetros podem influenciar nos resultados. As tarefas STS e RTE que podem não expor todas as capacidades e limitações dos modelos.

2 Modelos em Português do Brasil

Apesar da carência de material acadêmico em relação ao tema, modelos de linguagem em português do Brasil, é um campo de estudo relevante na área de inteligência artificial. O estudo

realizado por Iman, Arabnia e Rasheed (2023), investigou a eficácia de técnicas de pré-processamento e arquiteturas no treinamento de modelos de linguagem. Examinaram a técnica de Aprendizado por Transferência (AT), detalhando alguns métodos. O ajuste fino, neste contexto, trata-se da técnica de transferência de aprendizagem mais comum, devido a sua simplicidade. Estudos em AT indicam algumas limitações, como por exemplo, modelos tendenciosos. Cada uma das técnicas possui aspectos positivos e negativos a serem considerados. Schneider, et al., (2021) utilizaram AT para treinar um modelo GPT-2 em português do Brasil. O modelo GPT-2 obteve melhor desempenho em relação ao *Elmo-Pt, embeddings* treinados em português, Word2Vec e LSTM. Ressaltando a efetividade da arquitetura Transformers. Baktash e Dawodi (2023) analisaram modelos mais atuais, o GPT-4 (modelo de linguagem de quarta geração na série GPT, desenvolvido pela OpenAI) por exemplo apresenta desafios e limitações: requisitos computacionais, requisitos de dados e preocupações éticas por exemplo. Em contrapartida, pode representar uma revolução na área de processamento de linguagem natural e a IA de forma geral, sua aplicação tem impacto significativo no mercado e sociedade.

Considerando o estudo de Souza, Nogueira, e Lotufo (2020) com foco nas métricas de avaliação, STS e RTE, foram adaptadas como métricas para realização de um comparativo entre os três modelos objeto deste estudo amplamente utilizadas na literatura. STS e RTE são tarefas relevantes no NLP que podem contribuir para a avaliação da eficiência e eficácia de modelos de linguagem. STS contribui para avaliar se o modelo é capaz de compreender a similaridade semântica entre duas frases, RTE tem como foco o reconhecimento se uma frase é uma inferência lógica de uma outra frase.

Embora exista uma carência de modelos de linguagem para português do Brasil, especificamente em relação a comparação de métricas de desempenho, este tudo representa uma contribuição nesta lacuna. Uma análise comparativa dos modelos *GPT-2, GPorTuguese-2* e *Cabrita-Lora-v0-1* poderá fornecer uma contribuição acadêmica valiosa na área, expondo capacidades e limitações dos modelos em questão e servindo como base para novos estudos. Espera-se variabilidade dos resultados, considerando que hiper parâmetros possam influenciar os resultados, RTS e RTE podem não ser suficientes para apresentar todas as capacidades e limitações dos modelos.

3 Metodologia

3.1 Dados

Foram consolidados diversos corpus, todos em português, para formação dos corpora de treinamento do modelo. Os conjuntos de dados foram:

LerNER-BR (318 mil palavras).

CETENFolha (24 milhões de palavras).

CETEMPúblico (180 milhões de palavras).

Português Now (1 milhão de palavras).

Projeto Corpus Brasileiro (1 bilhão de palavras).

Marc-Morpho (1 mil palavras).

Lácio-Ref (8 mil palavras).

Br-wac (2 bilhões de palavras).

PorPopular (600 mil palavras).

3.2 Pré-processamento

O conjunto de dados foram convertidos para o formato UTF-8 usando o comando iconv no terminal Linux do supercomputador SENAI CIMATEC. Em seguida, utilizamos o vim para remover todas as informações desnecessárias de alguns dos conjuntos de dados, deixando apenas o texto com significado semântico. Finalmente, os conjuntos de dados foram concatenados utilizando o comando cat.

3.3 Desenvolvimento do Tokenizador

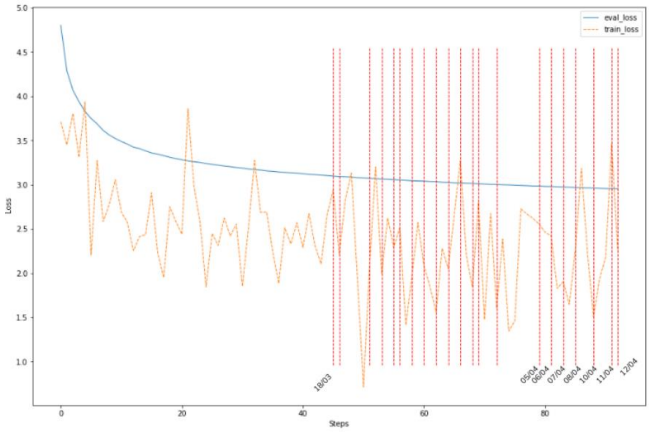
Os estudos realizados por Mikolov (2013) e Sennrich, et al., (2015) foram adaptados para realizar o desenvolvimento de um tokenizador para o novo corpora completo gerado na etapa de pré-processamento.

3.4 Treinamento

Para o treinamento, escolhemos a biblioteca *Simple Transformers* (https://simpletransformers.ai/), uma ferramenta de processamento de linguagem natural projetada para simplificar o uso de modelos *Transformer*. Esta biblioteca é baseada no trabalho da *Hugging Face* (https://huggingface.co/) e sua biblioteca *Transformers*. Utilizamos o modelo Megatron 345m da Nvidia

(https://catalog.ngc.nvidia.com/orgs/nvidia/models/megatron_bert_345m) como base para o treinamento, sendo realizado utilizando 4 GPUs A100 e 32 CPUs (figura 1).

Figura 1 - Treinamento.



Fonte: Dados da pesquisa.

O treinamento foi realizado em uma única época, com um tamanho de batch igual a 1. Acompanhamos a perda de avaliação e a perda de treinamento a cada 10.000 etapas globais. O treinamento foi interrompido na etapa 95.000 de um total de 4.302.246 etapas. A figura 1 ilustra a perda de avaliação (eixo y) e a perda de treinamento (eixo y) durante cada etapa de treinamento (eixo x). Este acompanhamento permite monitorar o progresso e a eficácia do treinamento ao longo do tempo.

4 Resultados

A tabela abaixo apresenta os melhores resultados dos desempenhos dos três modelos objeto deste estudo na tarefa STS. Foram avaliadas a correlação de *Pearson*, correlação de *Spearman* e o Erro Quadrático Médio (MSE). *Pearson* tem como foco avaliar a correlação linear entre as previsões e valores reais. Spearman avalia a correlação de *rank*, ou seja, se os modelos estão capturando a semelhança semântica entre as sentenças. MSE mede a diferença média entre previsões e valores reais.

Tabela 1 - Avaliação na tarefa STS.

Data	Modelo	Otimizador	batch	epoch	lr	eps	Pearson (DevSet/ TestSet)	Spearman (DevSet/T estSet)	MSE (DevSet/ TestSet)
4/4/2023	GPT-2	AdamW	16	7	1e-4	1e-10	0.933/0.7 01	0.884/0.67 6	0.0075/0. 0506
4/4/2023	GPortugu ese-2	AdamW	16	7	1e-4	1e-10	0.941/0.7 26	0.883/0.68 9	0.0072/0. 0471

4/4/2023	Cabrita-Lora-v0-1	AdamW	8	7	1e-4	1e-10	0.956/0.761	0.891/0.713	0.0037/0.0313
14/4/2023	GPT-2	AdamW	16	7	1e-6	1e-10	0.933/0.689	0.891/0.674	0.0074/0.0518
14/4/2023	GPortuguese-2	AdamW	16	7	1e-6	1e-10	0.935/0.731	0.874/0.687	0.0074/0.0475
14/4/2023	Cabrita-Lora-v0-1	AdamW	8	7	1e-6	1e-10	0.960/0.771	0.911/0.705	0.0039/0.0323

Fonte: Dados da pesquisa.

A tabela 1 indica que o modelo *Cabrita-Lora-v0-1* apresenta melhor desempenho em relação ao *GPT-2* e o *GPortuguese-2*. *Cabrita-Lora-v0-1* apresenta maior correlação de *Pearson* e *Spearman*, indicando que o modelo é capaz de capturar a similaridade semântica entre as sentenças. O MSE mais baixo indica que as previsões do *Cabrita-Lora-v0-1* são próximas dos valores reais. *GPT-2* e *GPortuguese-2* apresentam desempenho próximo. Empiricamente, de acordo com as métricas alcançadas, a capacidade de generalização e interpretação semântica das sentenças são aproximadas. Na tarefa RTE, na tarefa RTE, é avaliado a capacidade dos modelos em determinar se uma sentença implica logicamente em outra. O treinamento ocorreu com diferentes hiper parâmetros, foram avaliadas a precisão, recall e pontuação F1 para cada classe (0 e 1), bem como a média macro e ponderada das mesmas.

Tabela 2 - Avaliação na tarefa RTE.

Modelo	**pad_type**	**padding with pad**	**bos_eos**	**epochs**	**lr**	**eps**	**weight_decay_rate**	**batch**	**F1**
GPT-2	after	N	S	50	1.00E-05	1.00E-07	0,01	20	0.83
GPT-2	after	N	S	32	1.00E-05	1.00E-07	0,01	20	0.81
Gportuguese-2	after	N	S	50	0,00001	0,0000001	0,01	20	0.84
Gportuguese-2	after	N	S	32	0,00001	0,0000001	0,01	20	0.83
Cabrita-Lora-v0-1	after	N	S	32	0,00001	0,0000001	0,01	12	0.90
Cabrita-Lora-v0-1	after	S	N	50	0,00001	0,0000001	0,01	12	0.89

Fonte: Dados da pesquisa.

Considerando que os resultados variam de acordo com cada hiper parâmetro utilizado, *Cabrita-Lora-v0-1* apresenta melhor pontuação F1, seguido pelo *Gportuguese-2* e *GPT-2*, que apresentam valores com maior proximidade. Com base nos dados de Souza, et al., (2020), a tabela 3 indica, na tarefa STS (Pearson), Cabrita-Lora-v0-1 apresenta melhor desempenho em relação ao GPT-2 e GPortuguese-2. BERTimbau-Large apresentou o melhor resultado, considerando todos os modelos.

Tabela 3 - Comparação de STS (Correlação de Pearson).

Modelos	**Pearson**
BERTimbau-Large	0.852
BERTimbau-Base	0.836
mBERT	0.809
SOTA Anterior	0.830
GPT-2	0.701
GPortuguese-2	0.726
Cabrita-Lora-v0-1	0.761

Fonte: Dados da pesquisa.

A tabela 4 destaca os resultados na tarefa RTE (F1-score). Cabrita-Lora-v0-1 e BERTimbau-Large obtiveram os melhores desempenhos.

Tabela 4 - Comparação de RTE (F1-score).

Modelo	**F1-score**
BERTimbau-Large	90.0%
BERTimbau-Base	89.2%
mBERT	86.8%
SOTA Anterior	88.3%
GPT-2	83.0%
GPortuguese-2	84.0%
Cabrita-Lora-v0-1	90.0%

Fonte: Dados da pesquisa.

BERTimbau-Large apresentou o melhor desempenho na tarefa STS (Pearson). Cabrita-Lora-v0-1 em segundo lugar em termos de desempenho. GPT-2 e GPortuguese-2 tiveram desempenhos inferiores. O desempenho do GPT-2 foi o mais próximo do BERTimbau-Large,

enquanto o GPortuguese-2 apresentou o mais baixo resultado entre os modelos avaliados. A proximidade dos resultados indica uma competição acirrada entre esses modelos de linguagem.

5 Discussão

Este estudo comparou a capacidade de três modelos treinados em português, no que tange compreender e gerar textos, por intermédio das métricas STS e RTE. Trata-se de uma contribuição efetiva para a área de NLP em português do Brasil. *Cabrita-Lora-v0-1* teve o melhor desempenho em relação aos demais, com maiores correlação de *Pearson* e *Spearman*, e menor MSE. O mesmo ocorreu em RTE, com melhor F1. Esses resultados podem indicar que o modelo *Cabrita-Lora-v0-1* é mais eficiente e eficaz na em considerar as semelhanças semânticas entre as sentenças e se uma determinada sentença implica logicamente em outra.

Cabrita-Lora-v0-1 pode ser uma escolha valiosa para tarefas de NLP em português do Brasil. A realização de um comparativo entre modelos de linguagem, especificamente em português, permite identificar os métodos e modelos mais eficazes para treinamento e tarefas específicas na área de NLP. Contribuindo para a otimização e desenvolvimento de novos modelos. Apesar de *Cabrita-Lora-v0-1* superar os outros dois modelos, existem lacunas para otimizações. As métricas alcançadas podem ser otimizadas. Podendo ser um indicativo de oportunidades de futuras pesquisas e melhorias nos modelos objetos deste estudo.

Existem potencialidades no GPT-2 e GPortuguese-2, cada um possui aspectos próprios e positivos que podem ser adequados para tarefas específicas. O Cabrita-Lora-v0-1 pode ser uma escolha para determinadas tarefas, ainda existem oportunidades de otimizações envolvendo todos os modelos que foram objetivo deste estudo. As métricas alcançadas são indicadores oportunidades para futuras pesquisas e melhoria contínua dos mesmos.

6 Conclusão

Este estudo concentrou-se no comparativo do desempenho dos modelos *GPT-2*, *GPortuguese-2* e *Cabrita-Lora-v0-1* nas tarefas STS e RTE. *Cabrita-Lora-v0-1* apresentou melhores resultados. Indicativo que o modelo é eficiente e eficaz para tarefas objetos desta pesquisa. Ressaltando a capacidade desses modelos em lidar com os aspectos inerentes ao português do Brasil. Embora os experimentos tenham sido realizados com rigor científico, o desempenho pode variar de acordo com hiper parâmetros, tamanho e variedade do conjunto de dados e da tarefa específica.

Existem outros fatores que podem exercer influência no desempenho do modelo, por exemplo a estrutura computacional existente. Requerendo uma investigação mais aprofundada

em relação ao tema. Este artigo oferece uma contribuição efetiva para a área de NLP em português do Brasil, atualmente carente em relação à literatura existente. A crescente demanda por soluções de NLP em português do Brasil, ressalta a importância deste estudo, a identificação de modelos pode contribuir para soluções práticas, contribuindo em diversas áreas como tradução, análise de sentimentos e assistentes virtuais.

Podem ser realizadas pesquisas futuras, com foco na avaliação de outros modelos de linguagem treinados em português do Brasil, com diferentes conjuntos de dados e a utilização de maiores quantidades de métricas. Este artigo pode representar uma contribuição efetiva na literatura de NLP em português do Brasil, fornecendo resultados substanciais e direcionamento para futuros estudos.

Referências

BAKTASH, J.A.; DAWODI, M. Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing. **arXiv preprint arXiv:2305.03195**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2305.03195. Acesso em: 08 jul. 2024.

IMAN, M.; ARABNIA, H.R.; RASHEED, K. A Review of Deep Transfer Learning and Recent Advancements. **Technologies**, 2023, 11, 40. Disponível em: https://doi.org/10.3390/technologies11020040 .Acesso em: 08 jul. 2024.

MIKOLOV, T. *et al*. Efficient estimation of word representations in vector space. **arXiv preprint arXiv: 1301.3781**, 2013. Disponível em: https://doi.org/10.48550/arXiv.1301.3781. Acesso em: 08 jul. 2024.

RADFORD, A. *et al*. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019. Disponível em: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf . Acesso em: 08 jul. 2024.

SCHNEIDER, E. T. R. *et al*. A GPT-2 Language Model for Biomedical Texts in Portuguese. *In*: 21 IEEE 34TH INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS (CBMS). Aviero (Portugal): IEEE, 2021. pp. 474-479. Disponível em: https://doi.org/10.1109/CBMS52027.2021.00056. Acesso em: 08 jul. 2024.

SENNRICH, R., HADDOW, B., BIRCH, A. Neural Machine Translation of Rare Words with Subword Units. **arXiv preprint arXiv:1508.07909**, 2015. Disponível em: https://doi.org/10.48550/arXiv.1508.07909. Acesso em: 08 jul. 2024.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: CERRI, R.; PRATI, R.C. (eds). Intelligent Systems. BRACIS 2020. **Lecture Notes in Computer Science**, vol 12319. Springer, Cham. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-61377-8_28. Acesso em: 08 jul. 2024.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese Named Entity Recognition using BERT-CRF. **arXiv preprint arXiv:1909.10649**. 2010. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-61377-8_28. Acesso em: 08 jul. 2024.

VASWANI, A. *et al*. *In*: 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS. Long Beach (California): NIPS, 2017. p. 5998-6008. Disponível em: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Acesso em: 08 jul. 2024.