PERGUNTAS E RESPOSTAS COM *TRANSFORMERS*: UM COMPARATIVO DOS MODELOS ALBERTINA E BERTIMBAU

JÚLIA DA ROCHA JUNQUEIRA[1]; LARISSA ASTROGILDO DE FREITAS[2]; ULISSES BRISOLARA CORRÊA[3]

[1]*Universidade Federal de Pelotas – julia.rjunqueira@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – ulisses@inf.ufpel.edu.br*

1. INTRODUÇÃO

O processamento de perguntas e respostas é uma faceta essencial do campo de Processamento de Linguagem Natural (PLN), com amplas aplicações em assistentes virtuais, motores de busca, sistemas de recomendação e entre outros. Nas últimas décadas, o PLN tem testemunhado avanços significativos, e um dos marcos mais notáveis nesse progresso tem sido a utilização das redes neurais conhecidas como *Transformers*. Impulsionados pelo Aprendizado Profundo, esses modelos têm demonstrado grandes capacidades em tarefas de compreensão de linguagem natural (EKMAN, 2021).

A tarefa de Perguntas e Respostas (do inglês, *Question Answering* - QA), combina diversos campos de pesquisa de PLN. Os métodos empregados têm como objetivo principal solucionar e sugerir respostas que estejam em sintonia com as perguntas formuladas na linguagem natural escolhida, mutuamente. O QA pode ser composto por três módulos: classificação da pergunta, recuperação de informações e extração de respostas (ALLAM et al., 2012). A classificação da pergunta desempenha o papel de identificar o tipo de resposta esperada com base na pergunta apresentada. Por exemplo, ao formular a pergunta "Quando começou a pandemia do COVID?", espera-se que retorne uma data como resposta. A etapa de recuperação de informações, por sua vez, tem a função de fornecer resultados de pesquisa relevantes com base na pergunta submetida e em seu tipo. Já a extração de respostas visa recuperar e apresentar a resposta de acordo com as expectativas geradas pela pergunta original.

À medida que a popularidade e necessidade de sistemas de QA cresce, se demonstra necessário a avaliação de modelos e tecnologias que se destacam nesse cenário, tornando-se uma prioridade fundamental para acompanhar a evolução da demanda. É essencial garantir que esses sistemas ofereçam resultados confiáveis, precisos e culturalmente sensíveis (KRÜGEL et al., 2023). Deste modo, este artigo traz uma avaliação do modelo de linguagem para português brasileiro Albertina PT-BR (RODRIGUES, 2023), demonstrando sua eficácia na tarefa de QA, em comparação com o modelo BERTimbau (SOUZA et al., 2020).

Este artigo está organizado como segue: A Seção 2 informa acerca da metodologia e descreve as etapas realizadas para conduzir os experimentos, incluindo informações sobre conjuntos de dados e *fine-tuning*. A Seção 3 apresenta os resultados, a discussão acerca dos mesmos e fornece uma análise crítica da eficácia dos modelos. A Seção 4 discorre sobre as considerações finais e trabalhos futuros.

2. METODOLOGIA

Nesta seção, descreve-se em detalhes os procedimentos e abordagens utilizados para avaliar a eficácia do modelo Albertina PT-BR para a tarefa de QA em língua portuguesa brasileira. Primeiramente, foi realizado o *fine-tuning* do modelo Albertina para a tarefa. Em seguida, foi utilizado o *dataset* SQUAD v1.1 (RAJPURKAR et al., 2016) para o teste do modelo. Por fim, avaliamos as métricas em cima do modelo (Figura 1).

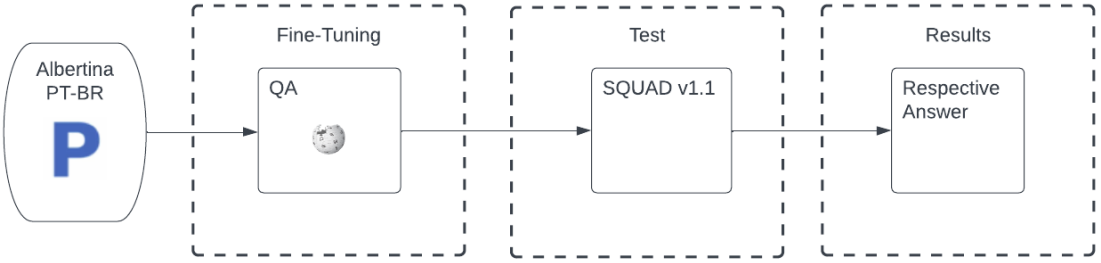

Figura 1: Metodologia deste teste.

Para os experimentos, foi utilizado o *dataset* SQUAD v1.1 (RAJPURKAR et al., 2016), onde o conteúdo deste foi criado traduzindo automaticamente o *dataset* SQUAD original (RAJPURKAR et al., 2016), utilizando a Google Cloud API. O *dataset* contém o total de 98169 sentenças, sendo dividida em 87599 sentenças para treino e 10570 para teste. Cada sentença é composta por um título, um contexto, uma questão e uma resposta. Essas informações são retiradas de artigos da Wikipédia, onde a resposta de cada pergunta é um segmento retirado do contexto correspondente.

Foi utilizado o Trainer API para realizar a etapa de treinamentos e *fine-tuning*, como também sua biblioteca de avaliação. Para o experimento, utilizou-se 3 épocas, com o *learning rate* equivalente a $1 * 10^{-5}$ e *batch size* de 16 e 8, para a versão *base* e *large* do modelo, respectivamente (os demais hiperparâmetros são apresentados na Tabela 1) .

Hiperparâmetros	Base	Large
Attention Heads	12	16
Batch Size (*)	16	8
Epochs	3	3
Hidden Size	768	1536
Hidden Layers	12	24
Learning Rate	1e-5	1e-5
Loss Function	CrossEntropy	CrossEntropy
Optimizer	AdamW	AdamW
Parameters	100 M	900 M

Tabela 1: Hiperparâmetros utilizados para a avaliação.

É fundamental ressaltar que as escolhas dos hiperparâmetros foram determinadas por restrições específicas e podem não corresponder à configuração ideal para o problema em questão, já que o processo de treinamento do modelo é ajustado para encontrar um equilíbrio entre as limitações computacionais e a necessidade de alcançar um desempenho aceitável.

Para a avaliação do modelo para a tarefa de QA, foram utilizadas as métricas de *exact match* (EM) e *f-measure* (BROWNLEE, 2016). A métrica de *exact match* mede a proporção de respostas geradas pelo sistema que são idênticas às respostas de referência, já o *f-measure* considera a taxa de precisão e o *recall* de um modelo, permitindo uma avaliação equilibrada do desempenho, especialmente em situações em que os dados podem estar desequilibrados.

3. RESULTADOS E DISCUSSÃO

Nas Tabelas 2 e 3, dispõe-se os resultados adquiridos nos experimentos. Utiliza-se como comparação os resultados adquiridos em um estudo anterior, onde foi investigado o modelo BERTimbau sobre algumas tarefas (DA ROCHA JUNQUEIRA, 2023).

	Task	Dataset	F-Measure	EM%
B	QA	SQUAD v1-PT	0.56	43.29
A	QA	SQUAD v1-PT	**0.57**	**45.12**

Tabela 2: Resultados obtidos utilizando os modelos Albertina PT-BR (A) e BERTimbau (B) *base*.

	Task	Dataset	F-Measure	EM%
B	QA	SQUAD v1-PT	0.62	47.15
A	QA	SQUAD v1-PT	0.32	**47.30**

Tabela 3: Resultados obtidos utilizando os modelos Albertina PT-BR (A) e BERTimbau (B) *large*.

Ao observarmos os resultados retornados, obteve-se um valor de 45,12% de EM no Albertina PT-BR, em comparação com 43,29% no BERTimbau, ambos nas versões *base*. Já nas versões *large*, 47,30% de EM no Albertina PT-BR, em comparação com 47,15% no BERTimbau. Visualiza-se uma maior vantagem do modelo Albertina sobre a tarefa de QA, que pode ser relacionado ao fato do modelo Albertina PT-BR (900 milhões) possuir mais parâmetros que o modelo BERTimbau (335 milhões) (Tabela 1).

No entanto, ao compararmos com o trabalho de (GUILLOU, 2021), o resultado obtido é consideravelmente inferior ao dele, de 70,49% de EM. Isso se deve à escolha dos hiperparâmetros de configuração, já que optamos especificamente por reduzir nossos hiperparâmetros para que os experimentos pudessem ser executados sem atingir limitações computacionais. Além disso, a métrica EM utilizada como avaliação tende a ser bem sensível, já que uma pequena mudança na escrita ou pontuação do texto afetaria a avaliação de correto ou não.

4. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

A análise do desempenho do modelo Albertina PT-BR na tarefa de perguntas e respostas revelou resultados interessantes, visto que apresenta um desempenho superior em comparação com o modelo utilizado anteriormente na mesma tarefa, demonstrando uma tendência favorável ao uso do modelo Albertina PT-BR. É importante ressaltar que este resultado pode não refletir para outras tarefas, já que a arquitetura do modelo, o *fine-tunning*, os hiperparâmetros e a qualidade dos dados de treinamento são fatores que podem influenciar significativamente os resultados.

Os resultados fornecem uma base para trabalhos futuros e refletem sobre a necessidade contínua de otimização de modelos e seleção criteriosa de parâmetros para alcançar o melhor desempenho possível em diferentes contextos de aplicação. Para os trabalhos futuros, seria interessante analisar outros escopos possíveis para utilização do modelo, além de integrar outros *datasets* para analisar os resultados sobre outros tipos de dados. Ademais, explorar diferentes hiperparâmetros e estratégias de *fine-tunning* para determinar se alguma estratégia em específico se adapta melhor ao problema em questão. Agradecemos à CNPq, à FAPERGS e à NVIDIA Corporation pelo financiamento parcial deste trabalho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALLAM, A. M. N.; HAGGAG, M. H. The question answering systems: A survey. **International Journal of Research and Reviews in Information Sciences (IJRRIS)**, v. 2, n. 3, 2012.

BROWNLEE, J. **Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end**. Machine Learning Mastery, 2016.

DA ROCHA JUNQUEIRA, J. et al. BERTimbau in Action: An Investigation of its Abilities in Sentiment Analysis, Aspect Extraction, Hate Speech Detection, and Irony Detection. In: **The International FLAIRS Conference Proceedings**. 2023.

EKMAN, M. **Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow**. Reino Unido: Addison-Wesley Professional, 2021.

GUILLOU, P. (2021). Portuguese bert base cased qa (question answering), finetuned on squad v1.1 v1. 1. **Sl: sn**, 2021.

KRÜGEL, S.; OSTERMAIER, A.; UHL, M.. The moral authority of ChatGPT. **arXiv preprint arXiv:2301.07098**, 2023.

RAJPURKAR, P. et al. Squad: 100,000+ questions for machine comprehension of text. **arXiv preprint arXiv:1606.05250**, 2016.

RODRIGUES, J. et al. Advancing Neural Encoding of Portuguese with Transformer Albertina PT. **arXiv preprint arXiv:2305.06721**, 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9**. Springer International Publishing, 2020. p. 403-417.