Análise de Eficiência em NLP: Impacto do *Transfer Learning* na Classificação de Sentimentos com DistilBERT

Ronaldo Uchôa Oliveira[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Campus Maracanaú – Maracanaú – CE – Brasil

ronaldo.uchoa.oliveira60@aluno.ifce.edu.br

Abstract. *This work investigates the effectiveness of Transfer Learning in scenarios of data scarcity versus moderate data availability. We compared the pre-trained DistilBERT architecture against a randomly initialized version using the IMDB dataset. Results demonstrate that pre-training offers an accuracy advantage of 24.67% in low-data scenarios (86% vs 61.33%) and maintains 5.2% advantage with increased dataset size (91.9% vs 86.7%), besides converging significantly faster. The findings validate that pre-trained models are not just performance optimizers, but technical viability requirements in resource-constrained scenarios.*

Resumo. *Este trabalho investiga a eficácia do Transfer Learning em cenários de escassez de dados versus cenários de dados moderados. Comparou-se a arquitetura DistilBERT pré-treinada contra uma inicializada aleatoriamente (from scratch) utilizando o dataset IMDB. Os resultados demonstram que o pré-treinamento oferece uma vantagem de acurácia de 24,67% em cenários de poucos dados (86% vs 61,33%) e mantém vantagem de 5,2% com aumento de dataset (91,9% vs 86,7%), além de convergir significativamente mais rápido. Os achados validam que modelos pré-treinados não são apenas otimizadores de performance, mas requisitos de viabilidade técnica em cenários com restrição de recursos.*

1. Introdução

O treinamento de Redes Neurais Profundas (*Deep Learning*) tradicionalmente exige datasets massivos e alto poder computacional, o que é inviável para muitos projetos reais. No contexto de Processamento de Linguagem Natural (NLP), modelos como BERT [Devlin et al. 2019] revolucionaram o campo ao permitir a reutilização de conhecimento linguístico prévio através do *Transfer Learning*.

A hipótese central deste trabalho é que a utilização de modelos pré-treinados permite alcançar alta performance mesmo com poucos dados e hardware limitado, "transferindo" conhecimento semântico prévio adquirido em grandes corpora. Esta abordagem contrasta com o treinamento *from scratch*, onde o modelo precisa aprender simultaneamente a estrutura da linguagem e a tarefa específica.

O objetivo deste estudo é comparar quantitativamente o desempenho do modelo DistilBERT [Sanh et al. 2019] em dois estados: (A) Com pesos pré-treinados na língua inglesa e (B) Com pesos inicializados aleatoriamente (Tabula Rasa). A análise é conduzida em dois cenários experimentais que simulam diferentes níveis de disponibilidade de dados.

2. Trabalhos Relacionados

O conceito de *Transfer Learning* em NLP foi popularizado por Peters et al. [Peters et al. 2018] com ELMo, seguido pelo trabalho seminal de Devlin et al. [Devlin et al. 2019] que introduziu o BERT. Estudos subsequentes [Howard and Ruder 2018] demonstraram que o pré-treinamento é especialmente eficaz em cenários de baixo recurso.

Sanh et al. [Sanh et al. 2019] propuseram o DistilBERT, uma versão destilada que mantém 97% da performance do BERT com 40% menos parâmetros, tornando-o ideal para ambientes com recursos limitados. Este trabalho estende a análise empírica desses benefícios em diferentes regimes de dados.

3. Metodologia

3.1. Arquitetura do Modelo

Utilizou-se o DistilBERT (`distilbert-base-uncased`), uma versão destilada do BERT escolhida por ser 40% menor e 60% mais rápida que o BERT original, mantendo 97% da performance [Sanh et al. 2019]. A arquitetura possui 66 milhões de parâmetros e utiliza 6 camadas transformer com 768 dimensões ocultas.

3.2. Ambiente Experimental

Todos os experimentos foram executados no Google Colab com GPU Tesla T4 (16GB VRAM). Frameworks utilizados: PyTorch 2.0, Hugging Face Transformers 4.30 e Scikit-Learn 1.3. O tokenizador empregado foi o `DistilBertTokenizer` com estratégia de padding e truncamento em 512 tokens.

3.3. Dataset

O dataset IMDB Reviews [Maas et al. 2011] foi escolhido por ser um benchmark consolidado em classificação de sentimentos binária (Positivo/Negativo). O dataset original contém 50.000 reviews balanceados.

3.4. Configuração Experimental

O experimento foi conduzido em duas fases distintas:

1. **Fase Piloto (Baixo Recurso):** Treinamento com apenas 1.000 amostras de treino e 200 de teste para simular escassez extrema de dados rotulados.
2. **Fase de Validação (Médio Recurso):** Treinamento com 10.000 amostras de treino e 2.000 de teste para validar a consistência estatística e analisar a curva de aprendizado.

Para cada fase, treinaram-se dois modelos:

- **Modelo Pré-treinado:** DistilBERT com pesos inicializados a partir do checkpoint `distilbert-base-uncased`, pré-treinado em corpus da língua inglesa.
- **Modelo *From Scratch*:** DistilBERT com mesma arquitetura, porém com pesos inicializados aleatoriamente seguindo distribuição normal Xavier [Glorot and Bengio 2010].

3.5. Hiperparâmetros

Os hiperparâmetros foram mantidos idênticos para ambos os modelos em todas as fases:

- Taxa de aprendizado: 2×10^{-5}
- Otimizador: AdamW com $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Épocas: 4
- Batch size: 16
- Weight decay: 0.01

3.6. Métricas de Avaliação

A métrica principal foi a acurácia no conjunto de teste. Adicionalmente, foram analisadas as matrizes de confusão e curvas de aprendizado (loss vs. épocas) para avaliar a convergência dos modelos.

4. Resultados e Discussão

4.1. Cenário 1: Escassez de Dados (1.000 Amostras)

No cenário de baixo recurso, observou-se uma diferença drástica entre as duas abordagens. O modelo sem pré-treinamento apresentou dificuldade severa em generalizar, atingindo apenas 61,33% de acurácia – apenas 11,33 pontos percentuais acima do acaso (50%).

Em contraste, o modelo pré-treinado atingiu 86% de acurácia, demonstrando que o conhecimento linguístico prévio é crucial quando não há dados suficientes para aprender a sintaxe e semântica do zero. Esta diferença de 24,67 pontos percentuais representa um ganho relativo de 40,2% em performance.

A Figura 1 apresenta a evolução da acurácia durante o treinamento, evidenciando que o modelo *from scratch* apresenta convergência limitada mesmo após 3 épocas. A matriz de confusão na Figura 2 revela que o pré-treinado mantém precisão superior.

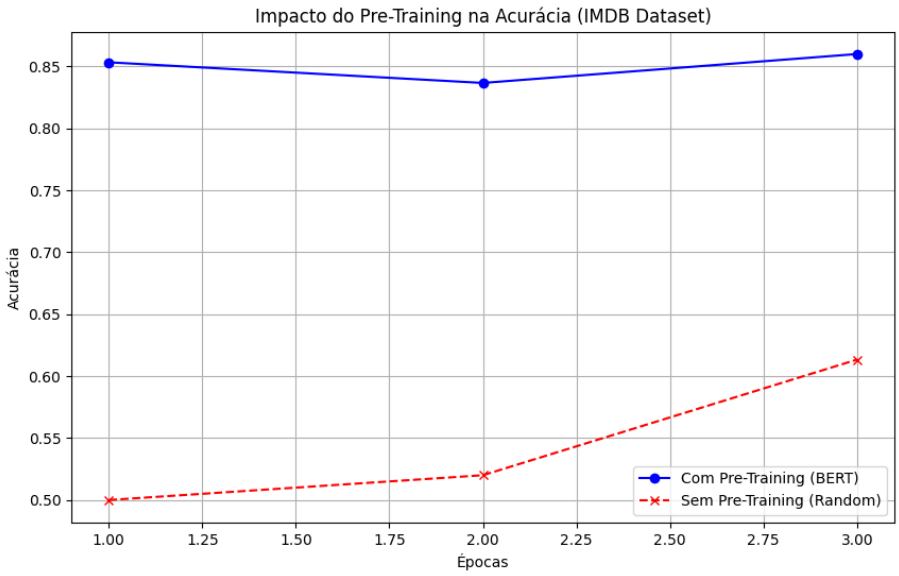

Figura 1. Evolução da acurácia durante treinamento – Cenário 1 (1.000 amostras). O modelo pré-treinado (azul) demonstra convergência estável atingindo 86%, enquanto o modelo *from scratch* (vermelho) apresenta performance limitada com 61,33% final.

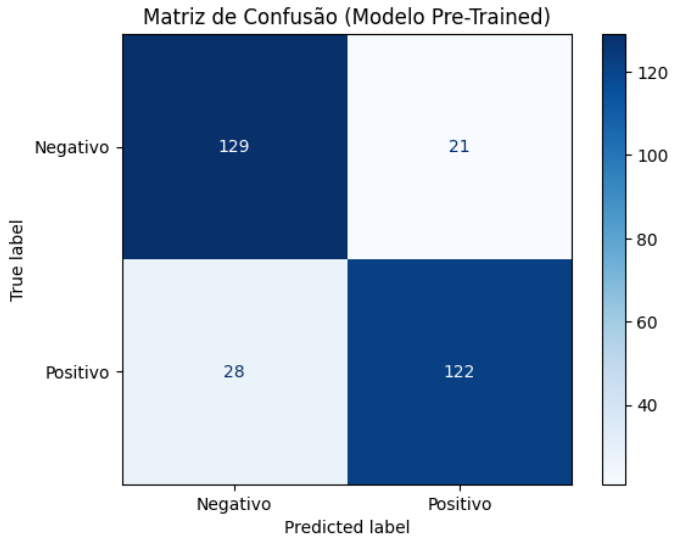

Figura 2. Matriz de confusão – Cenário 1 (300 amostras de teste). O modelo pré-treinado apresenta 251 acertos contra 49 erros, demonstrando balanceamento adequado.

4.2. Cenário 2: Aumento de Escala (10.000 Amostras)

Ao aumentar o conjunto de treinamento em 10x, o modelo sem pré-treinamento demonstrou capacidade substancial de aprendizado, melhorando sua performance para 86,7% de acurácia – um ganho impressionante de 25,37 pontos percentuais. Este resultado prova que o modelo é capaz de aprender representações linguísticas complexas quando há volume suficiente de dados rotulados.

O modelo pré-treinado atingiu 91,9% de acurácia, mantendo vantagem de 5,2 pontos percentuais. Observa-se que enquanto o modelo *from scratch* teve ganho absoluto de 25,37% ao aumentar os dados, o modelo pré-treinado ganhou apenas 5,9%, indicando que já estava próximo de sua capacidade máxima com poucos dados.

A análise das curvas de acurácia (Figura 3) revela um padrão interessante: o modelo pré-treinado atinge 91,3% já na primeira época. Esta velocidade de convergência tem implicações práticas diretas em economia de tempo de GPU e energia.

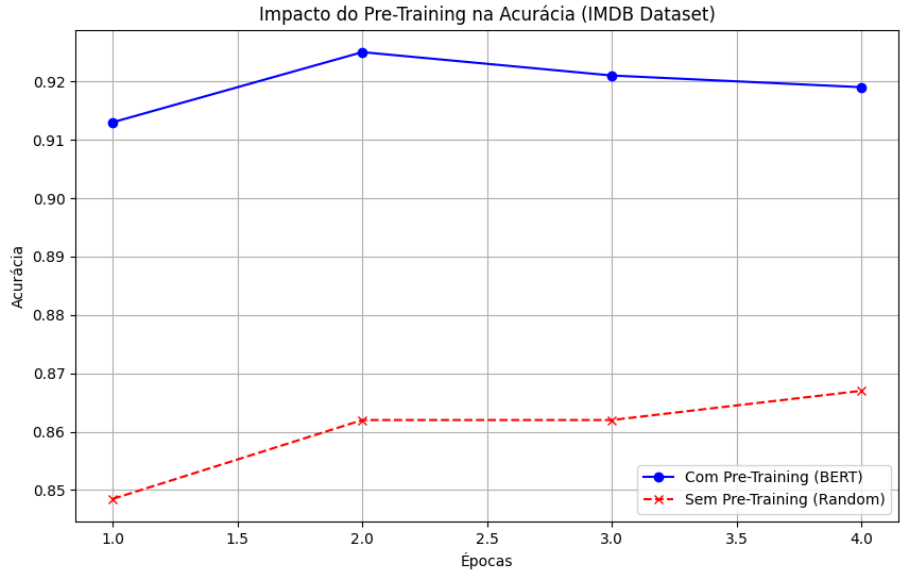

Figura 3. Evolução da acurácia durante treinamento – Cenário 2 (10.000 amostras). O modelo pré-treinado atinge 91,3% já na primeira época e estabiliza em 91,9%.

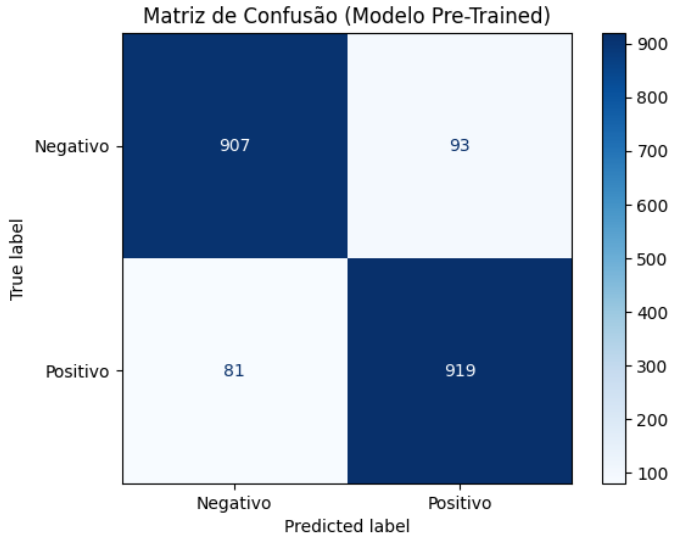

Figura 4. Matriz de confusão – Cenário 2 (2.000 amostras de teste). O modelo pré-treinado apresenta 1.826 acertos contra 174 erros, resultando em acurácia de 91,3% no teste.

4.3. Análise Comparativa

A Tabela 1 sumariza os resultados quantitativos de ambos os cenários. Observa-se que:

1. O ganho absoluto do pré-treinamento é drasticamente maior em cenários de baixo recurso (24,67% vs. 5,2%).
2. O modelo *from scratch* melhora impressionantes 25,37 pontos percentuais com 10x mais dados.

3. O modelo pré-treinado melhora apenas 5,9 pontos, indicando proximidade do limite de performance da arquitetura.

Tabela 1. Comparação de acurácia entre modelos nos dois cenários experimentais

Cenário	Pré-treinado	From Scratch
1.000 amostras	86,0%	61,33%
10.000 amostras	91,9%	86,7%
Ganho absoluto	**+5,9%**	**+25,37%**

4.4. Implicações Práticas

Do ponto de vista de engenharia de software e viabilidade econômica, os resultados indicam que o *Transfer Learning* com DistilBERT oferece três vantagens principais:

1. **Redução de custos de anotação:** Alcançar 86% com apenas 1.000 amostras rotuladas versus necessitar de 10.000+ amostras para resultados comparáveis.
2. **Economia de tempo de treinamento:** Convergência em 1 época versus 3+ épocas, economizando custos de GPU em nuvem.
3. **Democratização do NLP:** Viabiliza aplicações de alta qualidade sem necessidade de infraestrutura massiva.

5. Conclusão

Este trabalho validou empiricamente a hipótese de que o *Transfer Learning* com modelos pré-treinados não é apenas um otimizador de performance, mas um requisito de viabilidade técnica em cenários com restrição de dados. O experimento demonstrou que:

- Em cenários de baixo recurso (1.000 amostras), o pré-treinamento é essencial, oferecendo ganho de 40,2% relativo (86% vs 61,33%).
- Mesmo com 10x mais dados, o pré-treinamento mantém vantagem de 6% relativo (91,9% vs 86,7%).
- O modelo *from scratch* demonstra forte dependência de volume de dados.
- O tempo de convergência é significativamente reduzido com pré-treinamento.

Conclui-se que o pré-treinamento atua como um "catalisador", permitindo que a rede neural foque na tarefa específica em vez de gastar recursos aprendendo a estrutura fundamental da linguagem. Como trabalho futuro, propõe-se investigar o impacto em domínios específicos e técnicas de *few-shot learning*.

Referências

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.