Uma Abordagem Eficiente e Sustentável para Classificação de Aderência Temática usando Sentence-BERT e LightGBM

Vinícius Santos Monteiro

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Carlos – SP – Brazil

vini.mon@usp.br

Abstract. *This study proposes a computationally efficient solution to the LeanDL-HPC 2025 challenge, which involves the multi-class classification of the adherence of academic productions to strategic themes. The methodology employed is based on a two-phase architecture: initially, semantic features are extracted from the texts using the pre-trained Sentence-BERT model; then, supervised classification takes place. Two classifiers with low computational cost were analyzed: Logistic Regression, used as a baseline model, and LightGBM, a gradient boosting model. The results indicate that the suggested architecture is efficient, with the LightGBM model obtaining a weighted F1-score of 0.54, representing an improvement of 5 points compared to the baseline. In addition, the carbon footprint of each stage was monitored using the CodeCarbon library, proving the viability of the solution in terms of low-cost computing and sustainability.*

Resumo. *Este estudo propõe uma solução eficiente em termos computacionais para o desafio LeanDL-HPC 2025, que envolve a classificação multi-classe da aderência de produções acadêmicas a temas estratégicos. A metodologia empregada fundamenta-se em uma arquitetura de duas fases: inicialmente, são extraídas características semânticas dos textos por meio do modelo pré-treinado Sentence-BERT; em seguida, ocorre a classificação supervisionada. Dois classificadores de baixo custo computacional foram analisados: Regressão Logística, utilizada como modelo de base (baseline), e LightGBM, um modelo de gradient boosting. Os resultados indicam que a arquitetura sugerida é eficiente, com o modelo LightGBM obtendo um F1-score ponderado de 0.54, representando uma melhoria de 5 pontos em comparação com o baseline. Além disso, a pegada de carbono de cada etapa foi acompanhada com o uso da biblioteca CodeCarbon, comprovando a viabilidade da solução em termos de computação de baixo custo e sustentabilidade.*

1. Introdução

A crescente demanda por alinhamento entre a produção científica e as necessidades estratégicas da sociedade traz desafios consideráveis na análise e categorização de grandes volumes de texto. O LeanDL-HPC Challenge 2025 apresenta um desafio prático nessa área: relacionar teses e dissertações a tópicos de importância estratégica, classificando o grau de aderência como Baixo, Médio ou Alto. Este relatório descreve uma metodologia criada para solucionar este problema, com um enfoque duplo na precisão e na eficiência

computacional. A ideia principal é que modelos de linguagem modernos podem ser usados de maneira ”enxuta” (lean), dissociando a fase de extração de características, que exige mais recursos computacionais, da formação de classificadores leves. Essa estratégia busca equilibrar a capacidade de representação semântica dos modelos de transformação (transformers) com as demandas de baixo consumo de recursos defendidas pela computação de alto desempenho (HPC) sustentável. O código-fonte e os experimentos completos estão disponíveis publicamente em um notebook do Google Colab[1].

2. Objetivos

Este estudo tem como objetivo criar e avaliar uma arquitetura de duas etapas (extração de características com Sentence-BERT e classificação com modelos leves) para enfrentar o desafio LeanDL-HPC 2025, concentrando-se no equilíbrio entre desempenho preditivo e pegada de carbono.

3. Abordagem Proposta

A solução foi implementada em um pipeline sequencial, compreendendo as etapas de pré-processamento de dados, extração de características e modelagem.

3.1. Pré-processamento e Preparação dos Dados

O conjunto de dados inicial, composto por informações textuais e categóricas de produções acadêmicas, passou por uma etapa de limpeza para tratamento de valores ausentes. Para cada par (produção, tema), foram criados dois documentos de texto consolidados:

- **Documento da Produção:** Concatenação do título, resumo (em português) e abstract (em inglês) da produção acadêmica.
- **Documento do Tema:** Concatenação do nome do tema estratégico e suas palavras-chave associadas.

3.2. Extração de Características com Sentence-BERT

A conversão de documentos de texto em vetores numéricos densos, também conhecidos como embeddings, é o núcleo de nossa metodologia. Para esta tarefa, escolheu-se o modelo `paraphrase-multilingual-MiniLM-L12-v2` da biblioteca Sentence-Transformers. Este modelo foi selecionado devido ao seu equilíbrio entre desempenho e eficiência, sendo mais leve do que os modelos BERT convencionais e tendo uma capacidade multilíngue nativa, apropriada para nosso corpus [Reimers and Gurevych 2019].

Cada ”Documento da Produção” e ”Documento do Tema” foi processado pelo modelo, gerando um embedding de 384 dimensões para cada. O vetor de características final para cada amostra de dados foi criado pela concatenação do embedding da produção e do embedding do tema, resultando em um vetor de 768 dimensões.

[1] `https://colab.research.google.com/drive/1o8BKSmGwF8HEuggTLdGtzNnVjbULDqB5?usp=sharing`

3.3. Modelos de Classificação

Com os vetores de características definidos, dois modelos de classificação foram treinados e avaliados:

1. **Regressão Logística (Baseline):** Um modelo linear selecionado devido à sua simplicidade, facilidade de interpretação e custo computacional de treinamento extremamente baixo.
2. **LightGBM:** Um modelo de Gradient Boosting Machine (GBM) fundamentado em árvores. Foi selecionado devido à sua habilidade de identificar relações não-lineares nos dados, desempenho superior e elevada eficiência de treinamento em relação a outros algoritmos de boosting [Ke et al. 2017].

Em ambos os modelos, usou-se o parâmetro de balanceamento de classes para reduzir o impacto da distribuição desequilibrada dos rótulos (Baixo, Médio, Alto) no conjunto de dados.

3.4. Cálculo da Pegada de Carbono

A biblioteca CodeCarbon foi utilizada para monitorar a pegada de carbono. As emissões de CO_2eq (em gramas) foram medidas individualmente nas duas fases mais onerosas do processo: a criação de todos os embeddings e o treinamento de cada um dos classificadores [Lacoste et al. 2021].

4. Resultados

Os modelos foram treinados em 75% do conjunto de dados e avaliados nos 25% restantes. A Tabela 1 apresenta os resultados de performance, enquanto a Tabela 2 detalha a pegada de carbono.

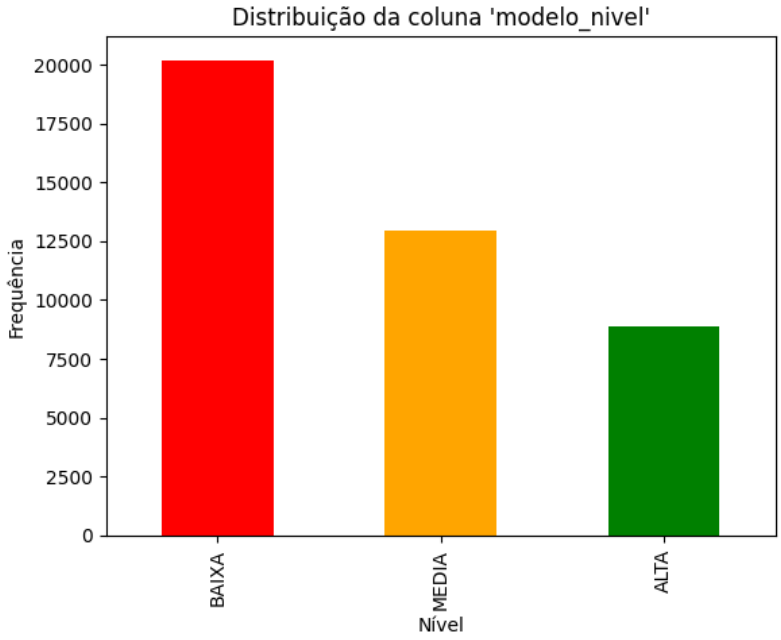

Figure 1. Distribuição das classes de aderência (Baixa, Média, Alta) no conjunto de dados.

Table 1. Resultados de Performance dos Modelos de Classificação

	F1-Score			
Modelo	**ALTA**	**MEDIA**	**BAIXA**	**Ponderado**
Regressão Logística	0.48	0.38	0.56	0.49
LightGBM	0.55	0.39	0.62	0.54

Table 2. Pegada de Carbono Estimada das Etapas do Pipeline

Etapa do Processo	**Emissões (gCO_2eq)**
Geração de Embeddings (Custo único)	2.67840
Treinamento - Regressão Logística	0.33014
Treinamento - LightGBM	0.45240

O modelo LightGBM apresentou uma vantagem evidente em todas as métricas, obtendo um F1-score ponderado de 0,54, o que representa um crescimento de 5 pontos percentuais em comparação com o baseline da Regressão Logística. O ganho mais significativo foi na classificação da classe ”ALTA”, que teve um crescimento de 7 pontos no F1-score, sinalizando uma habilidade melhorada para identificar corretamente as produções de maior aderência. Para ambos os modelos, a classe ”MEDIA” continuou sendo a mais difícil.

Em termos de eficiência, os dados apresentados na Tabela 2 indicam que o maior custo da abordagem está na criação dos embeddings. O custo de treinamento de ambos os classificadores é consideravelmente reduzido, sendo que o LightGBM exibe um consumo de recursos um pouco maior do que a Regressão Logística. No entanto, esse aumento é amplamente compensado pelo ganho significativo em desempenho.

5. Conclusão

Este estudo apresentou com êxito uma metodologia eficaz e econômica para a classificação de aderência temática. A abordagem de extração de características usando Sentence-BERT e classificação com LightGBM demonstrou ser robusta, superando de forma significativa o modelo de baseline, com um custo computacional de treinamento apenas marginalmente maior.

A análise da pegada de carbono confirma a eficácia da arquitetura ”lean”, na qual o custo computacional mais alto é incorrido apenas uma vez durante a etapa de vetorização, possibilitando experimentações rápidas e econômicas com diversos modelos de classificação. Conclui-se que a solução apresentada atende aos critérios do desafio LeanDL-HPC, fornecendo um modelo com excelente desempenho preditivo e em conformidade com as práticas de computação sustentável.

Para trabalhos futuros, recomenda-se a experimentação com modelos de embedding mais recentes e a aplicação de técnicas de fine-tuning do modelo de linguagem para tarefas específicas, visando um aumento ainda maior na performance, sempre observando o impacto no custo computacional, além da parte ambiental focada nesse projeto que é a pegada de carbono.

6. References

References

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems 30*.

Lacoste, A., Schmidt, A., Dandres, T., and Anthony, J. (2021). Codecarbon: A machine learning emissions calculator. *arXiv preprint arXiv:2110.11432*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.