# AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS DISCURSIVAS

GRAZIELE FAGUNDES MARTINS[1]; ULISSES BRISOLARA CORRÊA[2]

[1]*Universidade Federal de Pelotas – gfmartins@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – ulisses@inf.ufpel.edu.br*

## 1. INTRODUÇÃO

A avaliação de respostas discursivas desempenha um papel crucial nos sistemas educacionais, permitindo mensurar não apenas o conhecimento factual dos estudantes, mas também sua capacidade de argumentação, raciocínio e expressão escrita. Esse tipo de avaliação oferece informações mais ricas sobre a compreensão do aluno em comparação com questões de múltipla escolha, já que possibilita capturar nuances do raciocínio e identificar concepções alternativas.

No entanto, a correção manual dessas respostas é um processo que apresenta diversos desafios, como a subjetividade na atribuição de notas, a variabilidade entre avaliadores e o alto custo de tempo e esforço. Esses fatores tornam o processo menos escalável, especialmente em avaliações de larga escala, nas quais um grande número de respostas precisa ser corrigido em prazos curtos. Além disso, a correção desempenha um papel central em sistemas de ensino personalizado, pois é a partir dela que se torna possível identificar lacunas de conhecimento e adaptar o conteúdo às necessidades individuais dos estudantes.

Nesse cenário, com o avanço das técnicas de Processamento de Linguagem Natural (PLN) e do uso de modelos de linguagem pré-treinados, surge a possibilidade de automatizar parte desse processo, oferecendo avaliações mais rápidas, consistentes e com menor custo operacional. A tarefa de avaliação automática de respostas curtas (*Automatic Short Answer Grading* – ASAG) busca exatamente esse objetivo: classificar respostas abertas de estudantes com base em critérios objetivos de conteúdo, mesmo em contextos nos quais múltiplas formulações são semanticamente válidas (AHMED et al., 2022).

Inicialmente, as técnicas se apoiavam em métricas simples de similaridade textual, mas o surgimento de redes neurais, mecanismos de atenção e arquiteturas baseadas em *Transformers* ampliou consideravelmente a capacidade de capturar relações semânticas complexas. Atualmente, abordagens modernas exploram modelos de linguagem de grande escala (*Large Language Models* – LLMs), que não apenas avaliam respostas com maior precisão, mas também conseguem fornecer *feedback* construtivo e identificar equívocos conceituais sutis, antes considerados exclusivos da análise humana (MAITY; DEROY, 2024).

Neste trabalho, propõe-se a aplicação de técnicas de Inteligência Artificial, especificamente o ajuste fino (*fine-tuning*) supervisionado do modelo BERTimbau (SOUZA et al., 2020), para automatizar a avaliação de respostas discursivas em português. O estudo utiliza o conjunto de dados PT ASAG 2018 (GALHARDI et al., 2018), que reúne perguntas de ciências, respostas de referência e respostas de alunos com notas atribuídas por avaliadores humanos.

Além de investigar a eficácia do ajuste fino supervisionado, este trabalho avalia estratégias para lidar com o desbalanceamento das classes, como o uso de *undersampling* e a ponderação da função de perda, bem como o impacto da inclusão da pergunta no *input* do modelo sobre o seu desempenho. Para isso, são

utilizadas métricas amplamente empregadas na área, como acurácia, precisão, *recall* e *F1-score* (RAINIO et al., 2024), bem como a métrica *Quadratic Weighted Kappa* (DOEWES et al., 2023), a fim de mensurar a qualidade das predições em relação ao julgamento humano.

O objetivo final é analisar a viabilidade de empregar o BERTimbau como ferramenta de apoio a professores e instituições de ensino, reduzindo o tempo de correção e aumentando a consistência das avaliações, sem perder de vista a importância de manter a qualidade pedagógica do processo avaliativo. A escolha pelo BERTimbau se deve ao fato de ser um modelo consolidado e amplamente utilizado em pesquisas em língua portuguesa, com documentação acessível e custo computacional relativamente menor em comparação a arquiteturas mais recentes.

## 2. METODOLOGIA

O *dataset* PT ASAG 2018 foi desenvolvido com a participação de 245 estudantes do ensino fundamental II (8º e 9º anos), 13 professores e 12 graduandos em Biologia, totalizando 3.675 respostas para 15 questões de ciências com temas como corpo humano, respiração e circulação (GALHARDI et al., 2018). Cada questão contém entre 2 e 4 respostas de referência e uma lista de palavras-chave elaboradas por professores. As respostas foram avaliadas segundo uma escala de quatro níveis: 0 (totalmente incorreta), 1 (parcialmente correta com erros relevantes), 2 (majoritariamente correta com pequenas falhas) e 3 (correta e completa).

O pré-processamento consistiu na tokenização via *tokenizer* do modelo BERTimbau, com dois formatos de entrada:

- **Abordagem 1**: pergunta + resposta de referência [SEP] resposta do aluno;
- **Abordagem 2**: resposta de referência [SEP] resposta do aluno.

Para lidar com o desbalanceamento do *dataset*, aplicaram-se duas estratégias:

- ***Undersampling*:** redução da classe majoritária até igualar a frequência da menor classe, visando reduzir viés para a classe mais frequente;
- **Ponderação da função de perda:** atribuição de pesos inversamente proporcionais à frequência das classes, penalizando mais erros em classes minoritárias.

A divisão dos dados utilizou *holdout* estratificado (80% treino, 10% validação e 10% teste). O modelo BERTimbau Base foi treinado com a biblioteca *Transformers* e otimizado via *Optuna*, buscando maximizar a métrica *Quadratic Weighted Kappa*. Foram testados valores de *learning rate* entre $10^{-6}$ e $5\times10^{-5}$, *weight decay* entre 0 e 0,2, 2 ou 3 épocas e *batch sizes* de 8 ou 16.

## 3. RESULTADOS E DISCUSSÃO

As métricas utilizadas para avaliar as diferentes abordagens foram:

- **Acurácia (*Accuracy*):** proporção de respostas corretamente classificadas;
- **Precisão (*Precision*):** proporção de acertos entre as respostas previstas para uma classe;

- **Revocação (*Recall*):** proporção de respostas corretas de uma classe que foram previstas corretamente;
- **Medida-F (*F1-score*):** média harmônica entre precisão e *recall*;
- ***Quadratic Weighted Kappa*:** concordância ponderada entre predições e notas reais, penalizando mais erros com maior diferença.

A Tabela 1 apresenta o desempenho obtido por quatro combinações: ponderação da função de perda (*Weight*) e *undersampling* (*Under*), com e sem inclusão da pergunta (*WQuestion* / *WOQuestion*).

**Tabela 1 – Desempenho Dos Modelos No Conjunto De Teste**

| **Métrica** | ***Weight WOQuestion*** | ***Weight WQuestion*** | ***Under WOQuestion*** | ***Under WQuestion*** |
|---|---|---|---|---|
| Acurácia | 0,6677 | **0,6717** | 0,6058 | 0,6332 |
| Precisão | 0,6146 | 0,6230 | 0,6262 | **0,6553** |
| Revocação | 0,6294 | 0,6141 | 0,6109 | **0,6396** |
| Medida-F | 0,6197 | 0,6160 | 0,6174 | **0,6454** |
| *Kappa* (linear) | 0,6590 | **0,6643** | 0,5843 | 0,6131 |
| *Kappa* (quad.) | 0,7696 | **0,7784** | 0,7022 | 0,7241 |
| *BK Score* (méd.) | 0,7143 | **0,7214** | 0,6432 | 0,6686 |

O melhor resultado foi obtido pelo modelo **Weight WQuestion**, com acurácia de 67,17% e QWK de 77,84%. Ele apresentou desempenho equilibrado entre classes e maior sensibilidade para as categorias menos representadas.

A análise da matriz de confusão indicou alta precisão (92,70%) e *recall* (82,33%) para a classe 0, com 368 acertos de 447 exemplos. As classes intermediárias (1 e 2) tiveram desempenho mais modesto, com confusão frequente entre si, evidenciando a dificuldade do modelo em capturar nuances semânticas sutis. A classe 3 apresentou *F1-score* de 58,82%, com precisão de 62,50% e *recall* de 55,56%.

Esses resultados confirmam que o contexto adicional fornecido pela pergunta melhora a interpretação do modelo e que a ponderação da função de perda ajuda a equilibrar o desempenho entre classes, mitigando o viés para a classe majoritária. No entanto, o domínio restrito do conjunto de dados e o predomínio de respostas com nota 0 podem impactar a generalização para outros contextos.

## 4. CONCLUSÕES

O estudo demonstrou que é viável aplicar ajuste fino supervisionado do BERTimbau para avaliação automática de respostas discursivas em português. A

inclusão da pergunta no *input* e a ponderação da função de perda se mostraram estratégias eficazes para melhorar a performance e o equilíbrio entre classes.

Apesar dos bons resultados, ainda há limitações, especialmente na diferenciação de respostas parcialmente corretas. Trabalhos futuros devem explorar modelos maiores, ampliar a diversidade temática do *dataset* e integrar mecanismos de *feedback* textual para apoiar decisões do modelo.

O sistema proposto tem potencial para apoiar professores em avaliações de grande escala, reduzindo o tempo de correção e aumentando a consistência das notas atribuídas.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

AHMED, A.; JOORABCHI, A.; HAYES, M. J. On deep learning approaches to automated assessment: Strategies for short answer grading. **CSEDU**, v.2, p.85-94, 2022.

GALHARDI, L.; BARBOSA, C.; SOUZA, R.; BRANCHER, J. Portuguese automatic short answer grading. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v.29, n.1, p.1373, 2018.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: CERRI, R.; PRATI, R. C. (Eds.) **Intelligent Systems**. Cham: Springer, 2020. p.403-417.

MAITY, S.; DEROY, A. **The future of learning in the age of generative AI: automated question generation and assessment with large language models.** arXiv, 2024. Acessado em 20 ago. 2025. Online. Disponível em: https://arxiv.org/abs/2410.09576.

RAINIO, O.; TEUHO, J.; KLÉN, R. Evaluation metrics and statistical tests for machine learning. ***Scientific Reports***, v.14, n.1, p.6086, 2024.

DOEWES, A.; KURDHI, N.; SAXENA, A. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In: **INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING**, 16., Bangalore, 2023. Proceedings...: International Educational Data Mining Society, 2023. p.103-113.