# AVALIAÇÃO AUTOMÁTICA DE REDAÇÕES UTILIZANDO TRANSFORMERS

Vitor da Rocha Machado (PIC/UEM), Valéria Delisandra Feltrim (Orientadora). E-mail: vdfeltrim@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Maringá, PR.

**Ciência da Computação, Metodologia e Técnicas da Computação**

**Palavras-chave:** modelos de linguagem; processamento de linguagem natural; aprendizagem de máquina.

## RESUMO

Este projeto teve como objetivo o desenvolvimento de modelos preditores para a avaliação automática de redações no formato dissertativo-argumentativo, conforme o padrão do Exame Nacional do Ensino Médio (ENEM). Para treinamento e testes dos modelos, utilizou-se um subconjunto de um córpus público, composto por redações coletadas do *site* Educação UOL, todas avaliadas por avaliadores humanos de acordo com as diretrizes das cinco competências do exame. Foram construídos preditores individuais para cada competência, baseados na arquitetura *Transformer*, mais especificamente no modelo BERTimbau, utilizando a biblioteca *Simple Transformers* para o processo de *fine-tuning*. Em uma análise comparativa com abordagens tradicionais que utilizam *features* manuais, os modelos baseados em *Transformers* demonstraram superioridade consistente em todas as competências. Os resultados, embora limitados pelo tamanho reduzido e desbalanceamento do córpus, indicam que a abordagem é promissora e mais eficaz para a avaliação automática de redações no contexto avaliado.

## INTRODUÇÃO

A correção de redações é fundamental no processo educacional, sobretudo em exames de larga escala como o ENEM, que avalia cinco competências de 0 a 200 pontos. Entretanto, o processo demanda tempo, recursos humanos e está sujeito a variações entre avaliadores (SILVEIRA et al., 2024). Nesse contexto, a Avaliação Automática de Redações (*Automatic Essay Scoring* – AES) se apresenta como alternativa para reduzir custos e manter critérios consistentes (MARINHO et al., 2022). Embora já consolidada em outros idiomas, no português brasileiro ainda é limitada pela escassez de córpus adequados e publicamente disponíveis (SILVEIRA et al., 2024; MARINHO et al., 2022). As primeiras abordagens foram majoritariamente baseadas no uso de características manuais extraídas de textos, como métricas de complexidade lexical, coesão e uso da norma culta (AMORIM; VELOSO, 2017). Posteriormente, métodos baseados em redes neurais profundas passaram a gerar *embeddings* contextualizados (representações vetoriais densas), chegando em modelos *Transformers* (VASWANI et al., 2017), como o BERT

(DEVLIN et al., 2019), que produzem representações contextuais mais eficazes (MARINHO et al., 2022). Diante disso, este trabalho investigou o uso de modelos BERT pré-treinados para o português, aplicados ao córpus de Silveira et al. (2024), para a predição de notas de redações do ENEM, comparando-os a modelos tradicionais de aprendizado de máquina.

## MATERIAIS E MÉTODOS

*Córpus*

Para os experimentos, foi utilizada uma versão filtrada do subconjunto "*Source A with Graders*" do córpus de Silveira et al. (2024), composta por redações feitas entre os anos de 2015 e 2020 e extraídas do *site* Educação UOL, a maioria com três avaliações: uma realizada por avaliadores do *site* e duas realizadas por novos avaliadores, totalizando 1.164 redações. Para cada redação do subconjunto, havia o comando, texto de apoio, título, texto e ano da redação, notas (por competência e total), comentários gerais ou específicos (opcional) e a fonte. Os dados foram disponibilizados em formato *Parquet* por meio do *site Hugging Face* e seguiram a mesma divisão disponibilizada na plataforma: 65% para treinamento, 17% para validação e 18% para teste. Tal partição preservou o equilíbrio entre os temas das redações, garantindo que textos referentes a um mesmo comando permanecessem no mesmo subconjunto, evitando vazamento de informação entre as divisões.

*Treinamento*

Os experimentos foram conduzidos por meio da criação de preditores baseados no modelo BERTimbau-base (*neuralmind/bert-base-portuguese-cased*), utilizando a biblioteca *Simple Transformers* para *fine-tuning*. Cada competência foi tratada como uma tarefa independente, isto é, cada uma teve uma parametrização diferente para o ajuste fino. Para cada redação, a nota de cada competência (0 a 200 pontos, com saltos de 40) foi mapeada para classes de 0 a 5 para que a tarefa de classificação fosse aplicada, tanto por meio de um modelo classificador quanto por meio de um regressor com saídas arredondadas para os valores de classes possíveis.
Em linhas gerais, o treinamento foi executado a partir de seis épocas, variando-se para cada competência a taxa de aprendizado, perda de pesos, tamanho de lote e outros parâmetros. Além disso, houve um *early stopping* monitorado pela função de perda em validação, que ocorreu durante os treinos. Por fim, o modelo foi testado na partição de teste e seu desempenho foi medido em função de métricas, como *Quadratic Weighted Kappa* (QWK) e *Root Mean Squared Error* (RMSE), além de métricas de classificação, como acurácia, F1-Score e matriz de confusão. Tais informações foram salvas em registros e o código desenvolvido está disponível no repositório do projeto da plataforma *GitHub* https://github.com/hito-boo/pic-aes.

## RESULTADOS E DISCUSSÃO

No total, foram treinados 90 modelos de classificação e 38 modelos de regressão. A Tabela 1 apresenta as principais métricas obtidas pela parametrização dos modelos de classificação e regressão que obtiveram os melhores resultados. Os resultados completos, contendo todas as métricas utilizadas e todos os modelos gerados, estão disponíveis no repositório do projeto.

**Tabela 1 - Melhores resultados de classificação e regressão por competência**

| Preditor | QWK | RMSE | Accuracy | F1-Score |
|---|---|---|---|---|
| C1 - Classif. | 0,38 | 43,40 | 0,54 | 0,25 |
| C1 - Regres. | 0,45 | 42,39 | 0,49 | 0,23 |
| C2 - Classif. | 0,22 | 57,24 | 0,37 | 0,18 |
| C2 - Regres. | 0,28 | 53,95 | 0,36 | 0,19 |
| C3 - Classif. | 0,26 | 53,58 | 0,36 | 0,21 |
| C3 - Regres. | 0,42 | 47,41 | 0,35 | 0,23 |
| C4 - Classif. | 0,36 | 45,88 | 0,50 | 0,26 |
| C4 - Regres. | 0,43 | 43,01 | 0,49 | 0,26 |
| C5 - Classif. | 0,30 | 61,75 | 0,26 | 0,16 |
| C5 - Regres. | 0,34 | 51,95 | 0,24 | 0,15 |

Os modelos baseados no BERTimbau apresentaram desempenho moderado na predição das notas por competência, o qual foi superior na previsão das classes intermediárias. Esses achados estão em linha com a literatura, que aponta maior dificuldade em prever valores extremos (0 e 200 pontos) devido ao desbalanceamento das classes (SILVEIRA et al., 2024). Também foi possível observar um desempenho superior dos regressores em relação aos classificadores.
Ao comparar esses resultados com os obtidos por modelos baseados em abordagens tradicionais, como o que utiliza *features* TF-IDF e métricas de complexidade textual (CHALEGRE, 2025), observou-se uma vantagem dos modelos *Transformers*, que alcançaram métricas de QWK consistentemente superiores nas cinco competências. Já em relação ao trabalho de Silveira et al. (2024), os resultados foram 13% abaixo dos relatados pelos autores. No entanto, destaca-se que os melhores resultados de Silveira et al. (2024) foram obtidos por modelos que contaram com um pré-treinamento adicional.
Em síntese, os resultados indicam que o uso de *Transformers* é uma abordagem promissora para a avaliação automática de redações no contexto do ENEM, embora o tamanho reduzido do córpus e o desequilíbrio entre classes ainda limitem o desempenho dos modelos e se apresentam como um desafio.

## CONCLUSÕES

Este trabalho avaliou a viabilidade da utilização de Transformers para a avaliação automática de redações do ENEM. Observou-se desempenho consistente, ainda que moderado, sobretudo dos modelos regressores, com melhora em relação a abordagens baseadas em *features* manuais em todas as cinco competências.

Contudo, o desempenho foi limitado pelo tamanho reduzido do córpus de treinamento e pelo forte desbalanceamento das classes, dificultando a predição de notas extremas. Como trabalhos futuros, sugere-se ampliar o conjunto de redações e aplicar técnicas de mitigação do desbalanceamento, reforçando o potencial dos Transformers como ferramenta robusta para a automação e otimização do processo avaliativo no contexto educacional brasileiro.

## REFERÊNCIAS

AMORIM, E.; VELOSO, A. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. In: 15th CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: STUDENT RESEARCH WORKSHOP, 1., 2017, Valencia, Spain. **Proceedings** [...] Association for Computational Linguistics, 2017.

CHALEGRE, P. C. **Avaliação automática de redações por meio de aprendizado de máquina**. 2025. Relatório em andamento (Programa de Iniciação Científica) - Universidade Estadual de Maringá, Maringá, 2025 (**a publicar**).

DEVLIN, J. CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, Volume 1 (Long and Short Papers), 2019, **Proceedings** [...] Association for Computational Linguistics, 2019, p. 4171–4186.

MARINHO, J.; ANCHIÊTA, R. MOURA, R. Essay-br: a Brazilian corpus to automatic essay scoring task. **Journal of Information and Data Management**, p. 65–76, 2022.

SILVEIRA, I. C.; BARBOSA, A.; MAUÁ, D. D. A New Benchmark for Automatic Essay Scoring in Portuguese. In: 16th INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR), 1., 2024, Galiza, **Proceedings** [...] Association for Computational Linguistics, 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: 31st International Conference on Neural Information Processing Systems (NIPS'17), 2017, **Proceedings** […] Curran Associates Inc., Red Hook, NY, USA, 2017, p.6000–6010.