Avaliação de Modelos de Correção Automática de Redações ENEM
Um estudo com redações oficiais

Laís Nuto Rossman Orientador: Prof. Dr. Denis Deratani Mauá Coorientador: Prof. Msc. Igor Cataneo Silveira

Instituto de Matemática e Estatística — Universidade de São Paulo

INTRODUÇÃO

Apesar dos avanços em *Automated Essay Scoring* (AES) para o português, quase todos os modelos usam apenas redações de simulados. Este trabalho investiga quatro questões centrais:

- Os simulados são realmente parecidos com as redações oficiais?
- Modelos treinados só em simulados corrigem textos reais?
- O que ocorre ao treinar apenas no conjunto oficial (pequeno)?
- O pré-treino em simulados melhora a correção de textos oficiais?

METODOLOGIA

1. Coleta de dados

- Envio voluntário de redações oficiais via formulário online.
- Digitalização de textos manuscritos com apoio de LLMs.
- Revisão manual para garantir fidelidade total ao texto original.

2. Análise linguística

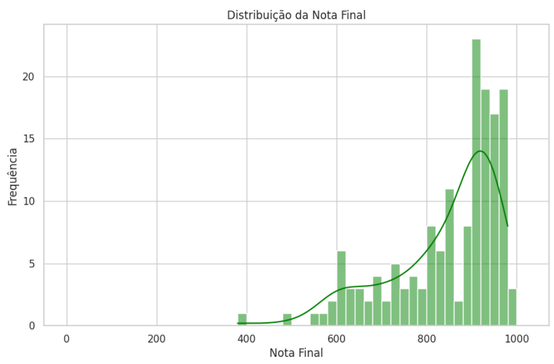
- Extração de 72 métricas textuais usando o NILC-Metrix.
- Modelos de regressão linear por competência (C1–C5).
- Identificação das métricas linguísticas mais influentes na nota.

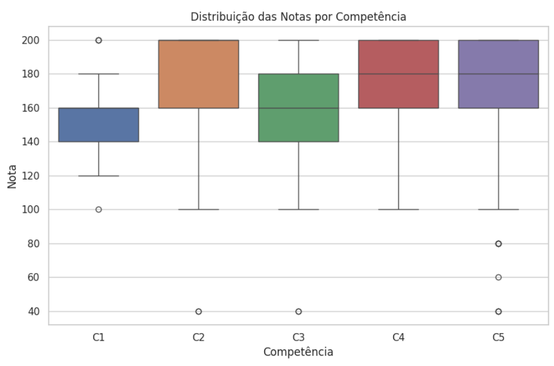
3. Experimentos com encoders

- Modelos avaliados: mBERT, BERTuguês e BERTimbau.
- Três cenários experimentais:
  - *Zero-shot*: uso direto sem ajuste.
  - *Pretrained Fine-tuning*: pré-treinado + ajustado no oficial.
  - *Exclusive Fine-tuning*: ajustado apenas nas redações oficiais.
- Divisão: 114 redações para treino, 43 para teste.
- Ajuste por *grid search* e validação cruzada.

RESULTADOS E DISCUSSÃO

Foram coletadas 157 redações oficiais do ENEM, incluindo o texto completo, o ano da prova e as notas oficiais em todas as competências.





Distribuição das notas: predominância de notas altas, reflexo do perfil dos participantes.

Análise textual: A regressão linear mostrou que as mesmas métricas linguísticas (substantivos, palavras de conteúdo, verbos, adjetivos e advérbios) dominam a predição das 5 competências. Esse padrão, também presente nos simulados, indica **forte semelhança linguística** e reforça que **simulados são eficazes para pré-treinamento**.

Experimentos com encoders

Tabela 1: QWK em todos os experimentos

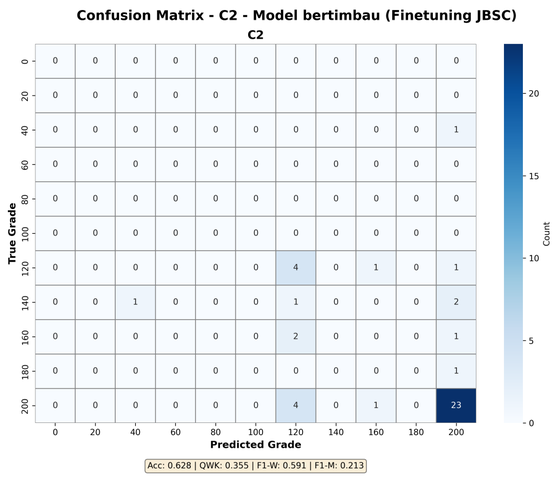
	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Mock Exams	.520	.220	.350	.500	.000	.318	.620	.330	.290	.540	.360	.428	.600	.360	.350	.550	.630	.498
Zero-shot	.501	.349	.447	.626	.030	.391	.382	.375	.290	.619	.271	.387	.378	.253	.387	.628	.069	.343
Pretrained FT	.456	.464	.519	.727	.000	.433	.467	.449	.286	.685	.489	.475	.393	.355	.453	.658	.385	.449
Exclusive FT	.000	.348	.000	.272	.359	.196	.073	.313	.150	.315	.020	.174	.000	-.023	.000	.452	.346	.155

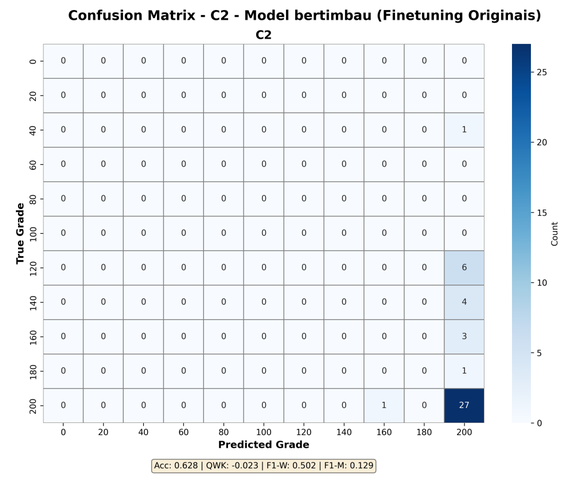
Tabela 2: F1 ponderado em todos os experimentos

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	avg	C1	C2	C3	C4	C5	avg	C1	C2	C3	C4	C5	avg
Zero-shot	.373	.571	.143	.334	.002	.285	.296	.565	.221	.369	.272	.345	.302	.517	.193	.396	.082	.298
Pretrained FT	.375	.630	.209	.354	.180	.350	.372	.625	.195	.376	.373	.388	.360	.591	.183	.391	.368	.379
Exclusive FT	.271	.608	.088	.260	.386	.323	.289	.595	.161	.280	.187	.302	.271	.502	.088	.308	.382	.310

Zero-shot: modelos pré-treinados em simulados já apresentam **bom desempenho inicial**, indicando **transferência efetiva** para redações oficiais.

Pretrained fine-tuning: o ajuste com poucas redações oficiais produz **os melhores resultados e maior estabilidade**, superando os demais cenários em QWK e F1.





Exclusive fine-tuning: o modelo teve a previsão concentrada em poucas classes, resultando em um **QWK muito baixo**; mesmo o F1, menos penalizado por previsões homogêneas, permanece **inferior ao cenário pré-treinado**.

CONCLUSÃO

- Redações simuladas e oficiais apresentam **alta semelhança linguística**
- Simulados são **muito mais abundantes** e de **coleta simples**
- Pré-treinamento em simulados **melhora consistentemente** o desempenho dos modelos.
- Ajustar o modelo apenas com um pequeno conjunto oficial garante **adaptação ao padrão real de correção**.

A estratégia mais eficaz para AES no ENEM é pré-treinar em grandes conjuntos de simulados e fazer fine-tuning final com poucas redações oficiais.

REFERÊNCIAS

1. Silveira, I. C.; Barbosa, A.; Mauá, D. D. *A New Benchmark for Automatic Essay Scoring in Portuguese*. PROPOR, 2024.
2. Leal, S. E.; Duran, M. S.; Scarton, C.; Aluísio, S. M. *NILC-Metrix: Assessing the Complexity of Written Language in Brazilian Portuguese*. Language Resources and Evaluation, 2024.
3. Barbosa, A.; Silveira, I. C.; Mauá, D. D. *An Empirical Analysis of Large Language Models for Automated Cross-Prompt Essay Trait Scoring in Brazilian Portuguese*. Journal of the Brazilian Computer Society (JBCS), 2025.
4. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL, 2019.