

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**Avaliação de Modelos de Correção
Automática de Redações ENEM**
um estudo com redações oficiais

Laís Nuto Rossman

MONOGRAFIA FINAL
MAC 499 — TRABALHO DE
FORMATURA SUPERVISIONADO

Supervisor: Prof. Dr. Denis Deratani Mauá
Cossupervisor: Prof. Msc. Igor Cataneo Silveira

São Paulo
2025

*O conteúdo deste trabalho é publicado sob a licença CC BY 4.0
(Creative Commons Attribution 4.0 International License)*

Agradecimentos

Agradeço ao meu orientador, Denis, por todo o suporte, paciência e ajuda nesses últimos meses. Ele foi meu primeiro professor de uma disciplina de IA na USP, o que me motivou bastante a explorar mais o assunto e desenvolver este trabalho na área.

Ao meu coorientador Igor, por toda a ajuda, paciência e por ser sempre tão solícito. Foram muitas reuniões até tarde, inúmeros experimentos e várias vezes em que o desespero bateu ele estava disposto a ajudar e encontrar uma solução.

Aos meus amigos de faculdade, que fizeram esses quatro anos, apesar de muito puxados, serem também muito divertidos. Todos os domingos de estudo com pizza na minha casa, ajudas nos EPs, os bandejões diários, as conversas e risadas no LabX, as festas na USP nas sextas e todos os desesperos compartilhados. Essa experiência foi muito mais leve e memorável com vocês.

Aos meus amigos de Fortaleza, que sempre torceram por mim mesmo de longe e que, apesar do tempo e da distância, sempre se fizeram presentes e apoiam a minha jornada.

À família Nuto e à família Cavalcante, que sempre me apoiaram e torceram por cada passo que dei durante a graduação, mesmo de longe.

Ao vovô Nuto, que foi quem me ensinou Matemática de um jeito divertido e fez com que eu me apaixonasse por exatas desde pequena. Ele faleceu poucos meses antes de eu ser aprovada no vestibular, mas gostaria muito que tivesse visto até onde a Matemática me levou.

Ao meu irmão e também colega de apartamento, Pedro, por toda a parceria, por compensar as tarefas domésticas quando tudo apertava nas vésperas das entregas, e por fazer São Paulo parecer cada vez mais casa com ele por perto.

Aos meus pais, Rossman e Carla, que sempre me apoiaram em tudo. Meu pai, uma das minhas maiores referências como profissional e pesquisador, nunca mediou esforços para que eu tivesse acesso às melhores oportunidades e me ensinou a ir atrás dos meus objetivos desde muito nova. Minha mãe, um dos meus maiores exemplos de resiliência,

nunca deixou com que eu me sentisse sozinha, me ligando quase todos os dias nos últimos quatro anos, sendo meu porto seguro mesmo de longe e me ajudando a lembrar que todo o esforço durante a graduação valeria a pena.

Aos meus professores, pela paciência e dedicação ao longo dos anos. Entrei no curso sem saber programar e cheia de incertezas sobre a minha escolha, mas tive docentes excepcionais, que me ensinaram com cuidado e me fizeram me apaixonar por computação. Hoje vejo que não poderia ter escolhido área melhor para construir minha carreira.

Por fim, ao IME-USP, por todas as oportunidades e vivências proporcionadas, desde grupos de extensão e monitoria até intercâmbio e o contato com pessoas incríveis. Não trocaria essa experiência por nada, foram os melhores quatro anos da minha vida.

Resumo

Laís Nuto Rossman. **Avaliação de Modelos de Correção Automática de Redações**

ENEM: um estudo com redações oficiais. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2025.

Sistemas automatizados de correção de redações podem, por um lado, reduzir a carga de trabalho dos professores nessa tarefa e, por outro, permitir que os estudantes pratiquem com mais frequência devido a ciclos de *feedback* mais rápidos. No português brasileiro, há um interesse crescente em sistemas de correção automática voltados para o exame padronizado ENEM. No entanto, os conjuntos de dados atualmente disponíveis consistem apenas em redações produzidas como prática para o exame oficial.

A questão sobre se esses conjuntos de dados de simulados fornecem informações úteis para a avaliação de redações oficiais do ENEM ainda não havia sido investigada na literatura. Este trabalho preenche essa lacuna ao apresentar um novo conjunto de dados rotulados composto por 157 redações escritas no exame oficial do ENEM.

Foi observado que o conjunto de dados apresentado neste trabalho compartilha características semelhantes às de conjuntos de redações produzidas em simulados. Também foi demonstrado que, para conjuntos pequenos como o utilizado neste estudo, o uso de *Large Language Models* (LLMs) pré-treinados em redações de simulados melhora significativamente o desempenho de sistemas automáticos de correção aplicados a textos oficiais do ENEM, resultando em um ganho médio de 0,27 pontos na métrica *Quadratic Weighted Kappa* em comparação com o treinamento realizado exclusivamente com dados oficiais.

Palavras-chave: correção automática de redações. ENEM. BERT. fine-tuning. zero-shot.

Abstract

Laís Nuto Rossman. **Evaluating Automated Scoring Models on Official ENEM Essays: an analysis using a newly collected dataset.** Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2025.

Automated essay scoring systems can reduce teachers' workload and allow students to practice more frequently through faster feedback cycles. In Brazilian Portuguese, there is growing interest in automatic scoring systems for the national standardized exam ENEM. However, the only datasets available so far consist of essays written as practice for the official exam.

Whether these simulated-essay datasets provide useful information for scoring official ENEM essays remained an open question in the literature. This work addresses this gap by presenting a new annotated dataset composed of 157 essays written during the official ENEM exam.

It was observed that the dataset presented in this study shares linguistic characteristics with datasets composed of simulated essays. It was also demonstrated that, for small datasets such as the one used here, the use of *Large Language Models* (LLMs) pre-trained on simulated essays significantly improves the performance of automated scoring systems applied to official ENEM essays, resulting in an average gain of 0.27 points in the *Quadratic Weighted Kappa* metric compared to training performed exclusively on official data.

Keywords: automatic essay scoring. ENEM. BERT. fine-tuning. zero-shot.

Listas de abreviaturas

IME Instituto de Matemática e Estatística

USP Universidade de São Paulo

ENEM Exame Nacional do Ensino Médio

AES Correção Automática de Redações (*Automated Essay Scoring*)

OCR Reconhecimento Óptico de Caracteres (*Optical Character Recognition*)

LLM Modelo de Linguagem de Grande Porte (*Large Language Model*)

QWK Kappa Ponderado Quadrático (*Quadratic Weighted Kappa*)

PLN Processamento de Linguagem Natural

IA Inteligência Artificial

Lista de figuras

1.1	Arquitetura Transformer	6
3.1	Distribuição das redações coletadas por ano	17
3.2	Distribuição das notas por competência	18
3.3	Distribuição da nota final das redações	18
3.4	Distribuição do número de palavras por redação	19
3.5	Correlação entre tamanho do texto e nota final	19
5.1	mBERT – Matrizes de confusão para C1 sob (esq.) Fine-tuning Pré-treinado e (dir.) Fine-tuning Exclusivo.	31
5.2	BERTimbau – Matrizes de confusão para C2 sob (esq.) Fine-tuning Pré-treinado e (dir.) Fine-tuning Exclusivo.	32
A.1	Matrizes de confusão: mBERT – Zero-shot (C1–C5)	41
A.2	Matrizes de confusão: mBERT – Fine-tuning após pré-treinamento em simulados (C1–C5)	42
A.3	Matrizes de confusão: mBERT – Fine-tuning Exclusivo (C1–C5)	42
A.4	Matrizes de confusão: BERTuguês – Zero-shot (C1–C5)	42
A.5	Matrizes de confusão: BERTuguês – Fine-tuning após pré-treinamento em simulados (C1–C5)	43
A.6	Matrizes de confusão: BERTuguês – Fine-tuning Exclusivo (C1–C5)	43
A.7	Matrizes de confusão: BERTimbau – Zero-shot (C1–C5)	43
A.8	Matrizes de confusão: BERTimbau – Fine-tuning após pré-treinamento em simulados (C1–C5)	44

A.9	Matrizes de confusão: BERTimbau – Fine-tuning Exclusivo (C1–C5)	44
-----	---	----

Listas de tabelas

5.1	Resultados de QWK para todos os modelos e configurações de treinamento	28
5.2	À esquerda: variação de QWK ao comparar Fine-tuning vs. Zero-shot. À direita: variação entre Pretrained FT e Exclusive FT	28
5.3	Resultados do F1-ponderado para todos os modelos e configurações de treinamento	29
5.4	À esquerda: variação do F1-ponderado ao comparar Fine-tuning vs. Zero-shot. À direita: variação entre Pretrained FT e Exclusive FT	29
5.5	Significância estatística das diferenças entre experimentos, baseada em intervalos de confiança obtidos via bootstrap. Células verdes indicam diferença significativa (ICs não se sobrepõem).	29
5.6	Desempenho em QWK para mBERT, BERTuguês, BERTimbau e GPT-4o na configuração <i>zero-shot</i>	30
5.7	Variação de QWK entre GPT-4o e modelos BERT em modo <i>zero-shot</i>	30
6.1	QWK por competência para modelos Zero-shot sob diferentes protocolos de avaliação.	35
6.2	QWK por competência para modelos Pretrained Fine-tuned em diferentes protocolos de avaliação.	35
6.3	QWK por competência para modelos Exclusive Fine-tuned em diferentes protocolos de avaliação.	36

Listings

Sumário

Introdução	1
1 Referencial Teórico	5
1.1 Modelos baseados em Transformers	5
1.1.1 Arquitetura Transformer	5
1.1.2 Modelos <i>Encoder</i> : família BERT	7
1.1.3 Modelos <i>Decoder</i> : família GPT	7
1.2 Métricas de Avaliação	8
1.2.1 <i>Quadratic Weighted Kappa (QWK)</i>	8
1.2.2 F1 Ponderado	9
1.3 Testes estatísticos: <i>Bootstrap</i>	10
2 Trabalhos Relacionados	11
2.1 Conjuntos de Dados Existentes de AES	11
2.2 Modelos de Correção Automática	12
2.3 Métricas de Complexidade Textual e o NILC-Metrix	12
3 Conjunto de Dados do ENEM	15
3.1 O Exame Nacional do Ensino Médio (ENEM)	15
3.2 Critérios de Avaliação da Redação	16
3.3 Conjunto de Redações Oficiais do ENEM	16
3.4 Análise Exploratória do Conjunto de Dados	17
4 Metodologia	21
4.1 Análise de Complexidade Textual	21
4.2 Modelos Encoder e Estratégias de Treinamento	21
4.2.1 <i>Fine-tuning</i>	22
4.2.2 Grid Search	23
4.2.3 Validação Cruzada (<i>K-Fold Cross-Validation</i>)	23

4.2.4	Testes Estatísticos	24
4.3	Escalas de Notas do ENEM	24
4.3.1	Protocolos de Avaliação e Estratégias de Arredondamento	25
4.4	Avaliação com o Modelo GPT-4o	25
5	Resultados	27
5.1	Métricas Textuais	27
5.2	Modelos <i>Encoder</i>	28
5.2.1	Testes Estatísticos	29
5.3	Comparação com o Modelo GPT-4o	29
5.4	Discussão	30
6	Avaliação Complementar dos Efeitos de Arredondamento	35
6.1	Influência das Estratégias de Arredondamento no QWK	36
6.2	Conclusão da Análise Complementar	37
7	Conclusão	39
Apêndices		
A	Matrizes de Confusão por Modelo	41
A.1	mBERT	41
A.2	BERTuguês	42
A.3	BERTimbau	43
Referências		45

Introdução

A correção automática de redações (*Automated Essay Scoring* – AES) é um campo consolidado dentro do Processamento de Línguagem Natural (PLN), utilizada em vários exames padronizados e diversos contextos educacionais (ATTALI e BURSTEIN, 2006; BEIGMAN KLEBANOV e MADNANI, 2020). Além do inglês, o desenvolvimento de sistemas de AES se expandiu para várias outras línguas, abrangendo, por exemplo, o francês (LEMAIRE e DESSUS, 2001), o japonês (ISHIOKA e KAMEDA, 2006) e o chinês (SONG *et al.*, 2016). Apesar dos avanços recentes terem melhorado consideravelmente o desempenho desses sistemas, a literatura focada no português brasileiro ainda é restrita. São raros os modelos de avaliação automática que foram treinados e validados especificamente para o português, e a maioria dos estudos disponíveis utiliza, predominantemente, redações elaboradas em contextos simulados, por exemplo AMORIM e VELOSO, 2017; MARINHO, ANCHIÉTA *et al.*, 2021; SILVEIRA, BARBOSA e MAUÁ, 2024. Essa discrepância em relação ao cenário internacional ressalta a importância de pesquisas que levem em conta os dados oficiais gerados no âmbito do Exame Nacional do Ensino Médio (ENEM), que é a principal via de acesso ao ensino superior no país.

Diante desse panorama, esta introdução apresenta, nas seções subsequentes, a justificativa para a realização do estudo, bem como a motivação que o fundamenta e os objetivos que o orientam, situando-o em seu contexto de relevância acadêmica e social.

Justificativa

A correção automática de redações foi proposta como uma forma de reduzir o esforço dos professores na tarefa trabalhosa de atribuir notas a textos escritos (PAGE, 1966). Ao tornar o processo de correção mais rápido, esses sistemas também permitem que estudantes pratiquem com maior frequência. Essa possibilidade é especialmente importante em contextos em que os candidatos precisam realizar um exame padronizado que exige a produção de um texto longo, geralmente uma redação.

No Brasil, o Exame Nacional do Ensino Médio (ENEM) representa a principal porta de entrada para o ensino superior. De acordo com dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), em 2025 o exame registrou mais de 4,8 milhões de inscrições confirmadas.¹

Dentre as etapas avaliativas, a prova de redação é amplamente reconhecida como

¹ <https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/enem/enem-2025-mais-de-4-8-milhoes-de-inscritos-confirmados>

um dos maiores desafios para os candidatos, pois exige não apenas domínio da norma padrão da língua portuguesa, mas também competências de interpretação, argumentação e proposição de soluções para problemas sociais.

Nesse contexto, o desenvolvimento de métodos mais eficientes e acessíveis para a correção de redações pode contribuir de forma significativa para a democratização do acesso ao ensino superior. O uso de modelos de Inteligência Artificial (IA) aplicados à avaliação de textos possibilita fornecer retornos rápidos e consistentes aos estudantes, ajudando na identificação de pontos de melhoria e na preparação para o exame. Assim, ferramentas automatizadas de correção podem desempenhar um papel relevante na redução de desigualdades educacionais e no apoio ao processo de aprendizagem.

Motivação

A maioria dos conjuntos de dados públicos para AES em português brasileiro é composta por redações enviadas para simulados aplicados por plataformas que imitam o ENEM (AMORIM e VELOSO, 2017; MARINHO, ANCHIÉTA *et al.*, 2021; SILVEIRA, BARBOSA e MAUÁ, 2024). Trabalhos como SILVEIRA, BARBOSA e MAUÁ, 2024 apontam que avaliadores experientes observam que os temas propostos por essas plataformas não reproduzem exatamente as características do exame oficial. Além disso, os estudantes que realizam simulados têm motivações diferentes dos candidatos do ENEM real, e as redações são escritas em condições também diferentes (e desconhecidas). Outro problema é o viés de seleção: não é público como esses sites coletam, corrigem e publicam as redações. Com isso, ainda não está estabelecida a validade ou a utilidade desses sistemas como ferramentas de preparação específica para o ENEM.

Essa lacuna torna-se ainda mais complexa devido à falta de conjuntos de dados representativos de redações oficiais. Embora os temas do exame sejam divulgados, apenas algumas redações com nota máxima são publicadas, impedindo que terceiros treinem modelos com dados realmente oficiais. Também não é possível verificar se modelos treinados em redações de simulados estão atribuindo notas semelhantes às do exame real.

Objetivo

Este trabalho busca preencher essa lacuna por meio da construção de um novo conjunto de redações oficiais do ENEM, acompanhadas de suas notas reais. A criação desse conjunto envolveu três etapas: elaboração de um formulário online para coleta dos dados, digitalização das redações manuscritas com apoio de modelos de linguagem, e uma etapa final de revisão manual das transcrições.

Com esse conjunto de dados, quatro questões de pesquisa foram investigadas:

1. Quão semelhantes são as redações de simulados em relação às redações oficiais do ENEM?
2. Modelos treinados apenas em simulados conseguem corrigir redações oficiais?

INTRODUÇÃO

3. Como modelos treinados exclusivamente no conjunto de redações oficiais se comportam?
4. O pré-treinamento em redações simuladas ajuda os modelos a se adaptar melhor às redações reais?

Para responder à primeira pergunta, foram extraídas métricas textuais utilizando a ferramenta NILC-Metrix (LEAL *et al.*, 2024) e ajustado um modelo de regressão linear para cada competência. Os resultados indicam que as características linguísticas mais relevantes do conjunto analisado são as mesmas encontradas em trabalhos anteriores com simulados (SILVEIRA, BARBOSA, COSTA *et al.*, 2025), sugerindo que os dois tipos de texto compartilham propriedades semelhantes.

Para responder às demais questões, foram utilizados três modelos disponibilizados por BARBOSA *et al.* (2025): BERTimbau-base (SOUZA *et al.*, 2020), BERTuguês (MAZZA ZAGO e AGNOLETTI DOS SANTOS PEDOTTI, 2024) e mBERT (DEVLIN *et al.*, 2019). Os modelos foram avaliados por meio das métricas *Quadratic Weighted Kappa* (QWK) e F1. Em seguida, foi realizado *fine-tuning* desses modelos no conjunto de dados deste estudo, e os resultados foram comparados ao cenário sem ajuste. De modo geral, observou-se que o desempenho melhora após o *fine-tuning*, embora o ganho varie entre competências. O maior aumento ocorreu no BERTimbau na Competência 5 (+0,316 em QWK), enquanto a maior queda foi observada no mBERT na Competência 1 (-0,04 em QWK).

Também foi conduzido um estudo adicional comparando o desempenho dos modelos pré-treinados e posteriormente ajustados com o desempenho de versões treinadas exclusivamente no conjunto de dados apresentado neste trabalho. Em média, os modelos pré-treinados apresentaram um ganho significativo de 0,27 em QWK em relação aos modelos sem pré-treinamento.

De forma resumida, as contribuições deste trabalho são:

- Construção e disponibilização de um conjunto de dados de redações oficiais do ENEM com suas respectivas notas;²
- Demonstração de que o conjunto de redações oficiais apresenta características semelhantes às de bases de simulados já existentes;
- Demonstração de que modelos pré-treinados apresentam boa capacidade de generalização para redações oficiais;
- Demonstração de que o pré-treinamento tem grande impacto no desempenho de modelos de correção automática.

O restante deste trabalho está organizado da seguinte forma: No Capítulo 1, é apresentado o referencial teórico sobre modelos *encoder* e *decoder* baseados em Transformers, métricas e testes estatísticos utilizados neste estudo. O Capítulo 2 discute os trabalhos relacionados. O Capítulo 3 descreve a criação do conjunto de dados de redações oficiais do ENEM. O Capítulo 4 detalha a metodologia empregada. Os resultados são apresentados e

² Disponibilizado em: <https://huggingface.co/datasets/laisnuto/self-collected-ENEM-dataset>

e discutidos no Capítulo 5, seguidos de análises complementares no Capítulo 6. Por fim, o Capítulo 7 apresenta as conclusões.

Capítulo 1

Referencial Teórico

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento deste trabalho. Aqui são apresentados conceitos centrais como Processamento de Linguagem Natural (PLN) e *Large Language Models* (LLMs), além de explicar a arquitetura *Transformer*, que é base dos modelos utilizados neste estudo. Por fim, são apresentadas as métricas empregadas para avaliar o desempenho dos modelos, com foco no *Quadratic Weighted Kappa* (QWK) e no F1 Ponderado, além de testes estatísticos como o *bootstrap* para avaliar se a mudança do desempenho foi significativa

1.1 Modelos baseados em Transformers

O Processamento de Linguagem Natural (PLN) é uma área da inteligência artificial que estuda como computadores podem compreender, analisar e gerar textos escritos em linguagem humana, incluindo tarefas como classificação, tradução automática, summarização e avaliação textual, incluindo a correção automática de redações, foco deste trabalho.

Nos últimos anos, grande parte dos avanços em PLN foi impulsionada por modelos baseados na arquitetura Transformer. Essa arquitetura permite capturar relações de longo alcance no texto e treinar modelos em grande escala, o que levou ao surgimento dos *Large Language Models* (LLMs). LLMs são modelos treinados em enormes quantidades de dados textuais e são capazes de executar uma ampla variedade de tarefas, como responder perguntas, resumir, traduzir, classificar ou gerar conteúdo.

Neste trabalho, os modelos empregados são baseados nessa arquitetura, mais especificamente da família BERT, que serão detalhados na subseção a seguir.

1.1.1 Arquitetura Transformer

A arquitetura Transformer, proposta por [VASWANI *et al.* \(2017\)](#), superou desafios importantes de modelos anteriores, como redes neurais recorrentes (RNNs) e convolucionais (CNNs). Esses modelos processam os dados de forma sequencial, o que dificulta o aprendizado de dependências longas e limita o paralelismo no treinamento.

O Transformer resolve esses problemas utilizando exclusivamente mecanismos de *self-attention*. Esse mecanismo permite que cada palavra em uma frase avalie diretamente sua relação com todas as outras, independentemente da distância no texto, capturando dependências longas de forma muito mais eficiente do que RNNs ou CNNs.

A arquitetura é composta por duas partes principais (Figura 1.1):

- **Encoder:** recebe o texto de entrada e transforma cada palavra em uma representação contextual. Cada camada combina *multi-head self-attention* com uma rede *feed-forward*. O mecanismo de *self-attention* permite que o modelo saiba quais palavras do texto são mais importantes para interpretar cada posição.
- **Decoder:** usa as representações criadas pelo encoder para gerar a saída final. Ele possui três partes: (1) um *masked self-attention*, que impede que uma palavra “veja” palavras futuras e garante que o texto seja gerado passo a passo; (2) um bloco de *encoder-decoder attention*, que conecta cada palavra gerada às informações do encoder; e (3) uma rede *feed-forward*.

Além disso, como o *Transformer* não possui recorrência, utiliza-se *positional encoding* para incorporar informação sobre a ordem das palavras na sequência, essencial para preservar significado.

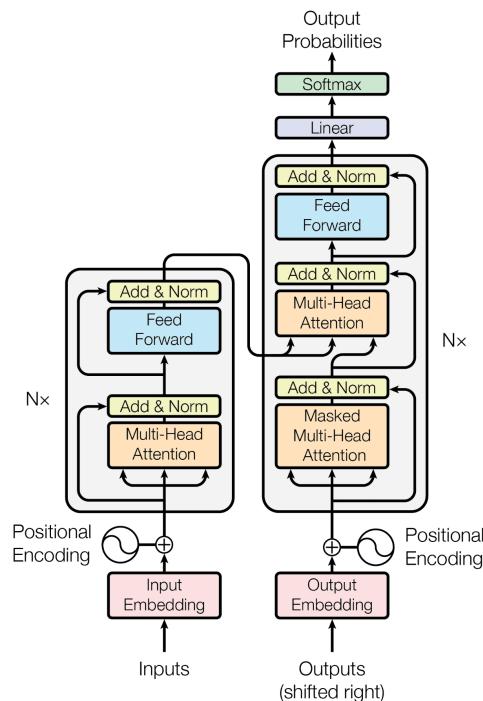


Figura 1.1: Arquitetura Transformer

Ao permitir que todas as palavras interajam entre si de forma paralela e eficiente, o *Transformer* tornou-se a base dos modelos modernos de Processamento de Linguagem Natural e, posteriormente, das LLMs, como o BERT, que usa apenas o *encoder*, e o GPT, que usa apenas o *decoder*.

1.1.2 Modelos *Encoder*: família BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) é um dos modelos mais influentes derivados do Transformer (DEVLIN *et al.*, 2019). Ele utiliza apenas a parte de *encoder* da arquitetura para gerar representações contextuais profundas. Por ser bidual, o BERT considera simultaneamente o contexto à esquerda e à direita de cada palavra, produzindo *embeddings* altamente informativos.

Essa capacidade de capturar o contexto amplo da frase tornou o modelo especialmente popular em tarefas de avaliação e análise textual, incluindo correção automática de redações (SILVEIRA, BARBOSA e MAUÁ, 2024; BARBOSA *et al.*, 2025). Por isso, neste trabalho são avaliadas variantes do BERT: modelos multilíngues, como o mBERT, e versões especializadas para o português, como o BERTuguês e o BERTimbau.

mBERT: O mBERT (*Multilingual BERT*) é a versão multilíngue do BERT, treinada em textos de 104 idiomas diferentes. Ele utiliza o mesmo vocabulário para todas as línguas, permitindo que o conhecimento adquirido em um idioma seja parcialmente transferido para outros.

BERTuguês: O BERTuguês é um modelo BERT treinado especificamente para o português brasileiro, mas o seu diferencial é que ele conta com algumas melhorias importantes no processo de tokenização: remoção de caracteres raros no português, inclusão de emojis e filtragem de textos de baixa qualidade. Como resultado, o modelo produz menos quebras de palavras e representa melhor o vocabulário da língua, apresentando desempenho superior ao BERTimbau-base em algumas tarefas.

BERTimbau: O BERTimbau é outro modelo BERT treinado especificamente para o português brasileiro. Ele foi disponibilizado em duas versões (*Base* e *Large*) e atingiu resultados muito bons em diversas tarefas de PLN em português, como reconhecimento de entidades nomeadas, similaridade textual e inferência textual.

1.1.3 Modelos *Decoder*: família GPT

Os modelos GPT (*Generative Pre-trained Transformer*) fazem parte da classe das LLMs e diferem dos modelos BERT por utilizarem exclusivamente a parte de *decoder* da arquitetura Transformer. Enquanto os *encoders*, como o BERT, produzem representações contextuais do texto de entrada, os modelos GPT são autoregressivos: eles geram o texto passo a passo, prevendo cada próximo token com base no que veio antes, o que os torna muito úteis para tarefas de geração de texto.

Esses modelos são pré-treinados em grandes quantidades de dados, aprendendo padrões linguísticos gerais antes de serem aplicados a tarefas específicas. A OpenAI foi a primeira empresa a lançar esses modelos em larga escala, começando com o GPT-1 em RADFORD *et al.*, 2018.

GPT-4o: O GPT-4o (OPENAI *et al.*, 2024) é uma versão recente multimodal da família GPT, sendo capaz de processar e combinar diferentes modalidades, desde texto, áudio,

imagem e vídeo, em um único sistema.

1.2 Métricas de Avaliação

A avaliação dos modelos requer métricas capazes de refletir tanto a concordância com os avaliadores humanos quanto a distribuição real das classes no conjunto de dados. Com esse objetivo, foi utilizado o *Quadratic Weighted Kappa* (QWK) e o F1 Ponderado, cujas definições são apresentadas nas próximas subseções.

1.2.1 *Quadratic Weighted Kappa* (QWK)

O Kappa Quadrático Ponderado (QWK) é uma métrica bastante usada para comparar notas atribuídas (DOEWES *et al.*, 2023), já que mede o nível de concordância entre dois avaliadores que classificam itens em uma escala ordinal, levando em conta a gravidade das discordâncias. Seu valor varia de -1 (discordância completa) até 1 (concordância perfeita), sendo que 0 indica um nível de concordância equivalente ao acaso.

A fórmula geral do QWK é dada por:

$$\kappa = 1 - \frac{\sum_{i,j} W_{ij} O_{ij}}{\sum_{i,j} W_{ij} E_{ij}}$$

em que:

- i e j representam categorias de notas atribuídas pelo primeiro e pelo segundo avaliador, variando de 0 a $N - 1$, onde N é o número total de categorias possíveis.
- O_{ij} é a frequência observada, ou seja, a proporção de vezes que um item recebeu nota i pelo primeiro avaliador e nota j pelo segundo (equivalente a uma matriz de confusão).
- E_{ij} é a frequência esperada, calculada a partir do produto externo das distribuições marginais das notas dos avaliadores, normalizada para ter o mesmo total de O . Esse termo representa a concordância esperada ao acaso.
- W_{ij} é o peso de discordância entre as categorias i e j , definido como:

$$W_{ij} = \frac{(i - j)^2}{(N - 1)^2}$$

Dessa forma, quanto maior a diferença entre as notas, maior é o peso da penalização aplicada.

Em resumo, o QWK mede o quanto dois avaliadores concordam levando em consideração não apenas se discordam, mas também o quanto discordam.

1.2.2 F1 Ponderado

O F1 Ponderado é uma métrica bastante utilizada em tarefas de classificação multiclasse, bastante útil quando as classes possuem distribuições muito diferentes. Ele combina o F1-score individual de cada classe com o número de exemplos reais daquela classe no conjunto de teste. Assim, classes mais frequentes têm maior peso no resultado final.

Para cada classe i , o F1-score é calculado a partir da precisão e da revocação:

$$\text{Precisão}_i = \frac{\text{VP}_i}{\text{VP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{VP}_i}{\text{VP}_i + \text{FN}_i},$$

em que:

- VP (verdadeiros positivos): número de exemplos cuja classe real foi corretamente prevista pelo modelo;
- FP (falsos positivos): número de exemplos cuja classe real não pertence a uma classe, mas o modelo os classificou como pertencentes a ela;
- FN (falsos negativos): número de exemplos cuja classe real pertence a uma classe, mas o modelo os classificou como outra.

Com esses valores, o F1-score individual de cada classe é definido como:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}},$$

em que:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \quad \text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

Cada classe possui ainda um *support*, definido como:

$$\text{support}_i = \text{número de exemplos reais da classe } i.$$

O F1 Ponderado combina os F1-scores de todas as classes em uma média ponderada:

$$\text{F1 Ponderado} = \frac{\sum_i (\text{support}_i \times F1_i)}{\sum_i \text{support}_i}.$$

Assim, o F1 Ponderado leva em conta a frequência real de cada classe no conjunto de teste, evitando que classes pouco frequente distorçam muito a métrica e permite uma avaliação mais fiel em cenários de desbalanceamento.

1.3 Testes estatísticos: *Bootstrap*

Testes estatísticos são ferramentas usadas para avaliar se uma hipótese sobre uma população é compatível com os dados observados. Quando a distribuição dos dados é desconhecida ou a amostra é pequena, técnicas de reamostragem oferecem uma alternativa prática para estimar a variabilidade de uma medida. Entre elas, está o *bootstrap*, usado neste trabalho para verificar a estabilidade das métricas e estimar seu intervalo de confiança.

O *bootstrap* é um método estatístico que estima a incerteza de uma métrica gerando várias novas amostras a partir da amostra original, por meio de reamostragem com reposição. Em cada reamostragem, é calculado novamente a estatística de interesse, o que permite construir uma distribuição empírica dessa estatística. Como a distribuição da população verdadeira não é conhecida, a própria amostra é usada como aproximação.

O procedimento pode ser resumido em três passos principais:

1. Criar uma nova amostra, do mesmo tamanho da original, por meio de reamostragem com reposição.
2. Calcular, nessa nova amostra, a estatística de interesse (no caso deste trabalho o QWK).
3. Repetir os passos anteriores várias vezes para construir uma distribuição empírica dessa estatística e, a partir dela, estimar intervalos de confiança e variabilidade.

Esse método é bastante útil quando o número de observações é pequeno ou quando não é possível assumir uma distribuição teórica específica. Por esses motivos, o *bootstrap* foi adotado neste trabalho para avaliar a estabilidade das métricas e quantificar a incerteza associada a elas.

Capítulo 2

Trabalhos Relacionados

Este capítulo apresenta os principais estudos, bases de dados e ferramentas relacionados ao tema deste trabalho. A seguir, são discutidos alguns conjuntos de dados já existentes sobre redações do ENEM e outros exames, depois são revisados trabalhos envolvendo modelos de correção automática e, por fim, é descrito o uso do NILC-Metrix em pesquisas anteriores e seu papel neste estudo.

2.1 Conjuntos de Dados Existentes de AES

Conjuntos de dados anteriores relacionados a redações do ENEM foram criados a partir da extração de textos e notas de plataformas online ([AMORIM e VELOSO, 2017](#); [MARINHO, ANCHIÉTA et al., 2021](#); [SILVEIRA, BARBOSA e MAUÁ, 2024](#)). Esses sites publicam mensalmente um tema, os estudantes escrevem uma redação sobre esse tema e a enviam para a plataforma, que então corrige o texto e disponibiliza o resultado final online. Essas redações são posteriormente coletadas e utilizadas com a nota atribuída pela plataforma ([AMORIM e VELOSO, 2017](#); [MARINHO, ANCHIÉTA et al., 2021](#)) ou reavaliadas por especialistas ([SILVEIRA, BARBOSA e MAUÁ, 2024](#)).

Essa abordagem de coleta em plataformas terceiras mostrou um importante *trade-off* na área de AES: o ganho em volume de dados em detrimento da especificidade e alinhamento com o exame oficial. Por exemplo, trabalhos como o de [AMORIM e VELOSO, 2017](#) utilizam um grande volume de dados, estimado em torno de 1.840 redações, enquanto o [MARINHO, ANCHIÉTA et al., 2021](#) conta com 4.570 redações. Segundo [SILVEIRA, BARBOSA e MAUÁ, 2024](#), especialistas observaram que os temas propostos nessas plataformas não se alinham adequadamente às características do exame oficial. Embora esses conjuntos de dados representem contribuições importantes, o objetivo principal ao desenvolver sistemas de correção automática não é avaliar redações de simulados, mas sim redações oficiais. Este trabalho apresenta um conjunto de dados formado por redações produzidas no exame oficial, juntamente com suas notas oficiais.

Outros conjuntos de dados para AES disponíveis em português brasileiro incluem o *Narrative Dataset* ([MELLO et al., 2024](#); [OLIVEIRA et al., 2025](#)) e o Diplomatix ([CAVALCANTI et al., 2025](#)). O primeiro é composto por textos narrativos escritos por estudantes do 5º ao

9º ano do Ensino Fundamental e não está associado a um exame padronizado. O segundo é formado por redações escritas para o exame oficial de diplomacia. Esse conjunto, assim como o conjunto de dados dessa pesquisa, apresenta um viés para notas altas. No caso do exame oficial de diplomacia, os respondentes foram aprovados, já no caso deste trabalho, a maioria foi aprovada no vestibular.

Esses exemplos mostram que conjuntos muito grandes, embora úteis para treinar modelos em larga escala, nem sempre refletem as condições e os critérios do ENEM, enquanto os conjuntos mais específicos enfrentam a dificuldade de reunir um volume significativo de dados oficiais. Para endereçar essa lacuna de dados específicos do exame, este trabalho reúne 157 redações oficiais do ENEM, garantindo especificidade e alinhamento com o processo real de correção, apesar do menor volume.

2.2 Modelos de Correção Automática

Além dos conjuntos de dados, diversos modelos de correção automática foram testados para avaliar redações de simulados do ENEM em [BARBOSA et al. \(2025\)](#). Esses modelos variam desde abordagens baseadas em características linguísticas até grandes modelos de linguagem com capacidade de raciocínio. Os autores concluíram que, embora nenhum modelo seja o melhor em todas as competências, modelos baseados em arquiteturas *Encoder* apresentam um bom equilíbrio entre tamanho, desempenho e custo computacional. Com base nessas conclusões, foram utilizados neste trabalho os modelos disponibilizados pelos autores para conduzir os experimentos. As versões *large* desses modelos não foram utilizadas, pois o estudo de referência indicou que os modelos *base* já atingem desempenho competitivo com custo computacional significativamente menor.

Além disso, o GPT-4o também foi avaliado em modo *zero-shot*. A motivação segue dos resultados desse mesmo artigo de referência, que mostrou que esse modelo supera outras alternativas de grande porte, como Sabiá e DeepSeek, quando usados sem ajuste. Dessa forma, o GPT-4o funciona como uma linha de base para comparação direta com os demais experimentos *zero-shot*.

2.3 Métricas de Complexidade Textual e o NILC-Metrix

O NILC-Metrix ([LEAL et al., 2024](#)) é uma ferramenta para calcular a complexidade textual em português. Ele extrai diferentes características, que abrangem diversas dimensões da escrita:

- contagens básicas, como número de palavras, sentenças e parágrafos;
- complexidade sintática, que inclui medidas de estruturas gramaticais e distância de dependência;
- coesão referencial, que avalia repetições de referentes e uso de pronomes;
- coesão semântica, baseada em similaridade entre sentenças e parágrafos via LSA;

- conectivos, que contabilizam a presença de conectores lógicos, causais e temporais;
- frequência lexical, que mede a quantidade de vezes que uma palavra aparece em um determinado corpus ou texto

Essas características foram empregadas em trabalhos anteriores para verificar a diferença entre Diplomatix e Essay-br (CAVALCANTI *et al.*, 2025). Neste trabalho, o NILC-Metrix foi utilizado para comparar o conjunto de dados apresentado com um conjunto de dados de testes simulados.

SILVEIRA, BARBOSA, COSTA *et al.* (2025) empregaram regressão linear baseada em características do NILC-Metrix. Os autores observaram que as características mais importantes para cada competência geralmente eram as mesmas. De acordo com os resultados reportados no artigo, para as competências C1, C2 e C3 as quatro métricas mais influentes apareceram sempre na mesma ordem: *adverbs*, *adjective_ratio*, *noun_ratio* e *verbs*. Já na competência C4, além dessas medidas, surgiram também métricas agregadas como *function_words*, *content_density* e o uso de conectivos causais negativos (*cau_neg_conn_ratio*). Para a competência C5, as métricas mais importantes voltaram a seguir um padrão muito parecido com o das três primeiras, mas nessa ordem: *adjective_ratio*, *verbs*, *adverbs* e *noun_ratio*.

O presente trabalho reproduz esse experimento com o objetivo de investigar se, em seu conjunto de dados, as características mais preditivas de cada competência coincidem com aquelas identificadas no estudo original.

Capítulo 3

Conjunto de Dados do ENEM

Neste capítulo, é apresentada uma visão geral sobre o ENEM e explicado como a redação é avaliada oficialmente. Em seguida, é descrito como o novo conjunto de redações deste trabalho foi coletado, digitalizado e revisado, além das principais características que compõem o *dataset* final utilizado nas análises.

3.1 O Exame Nacional do Ensino Médio (ENEM)

O ENEM é uma avaliação aplicada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) com o objetivo de mensurar o desempenho acadêmico dos estudantes ao final da educação básica. Realizado por milhões de participantes em todo o território nacional, o ENEM é amplamente utilizado como forma de ingresso em instituições de ensino superior públicas e privadas, além de servir como critério de seleção para programas governamentais como o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento Estudantil (FIES).

A prova é composta por quatro áreas de conhecimento:

- Linguagens, Códigos e suas tecnologias;
- Ciências Humanas e suas tecnologias;
- Ciências da Natureza e suas tecnologias;
- Matemática e suas tecnologias.

Cada área contém 45 questões objetivas, totalizando 180 questões, além de uma prova de redação com caráter dissertativo-argumentativo. As notas das áreas objetivas são calculadas com base na Teoria de Resposta ao Item (TRI), que considera o grau de dificuldade de cada questão e o padrão de acertos do participante.

A redação é avaliada separadamente, com pontuação máxima de 1000 pontos, distribuída igualmente entre cinco competências (200 pontos cada). A nota final do ENEM é obtida pela média aritmética das pontuações das quatro áreas e da redação. Dessa forma, a redação exerce um peso significativo na composição da nota final.

3.2 Critérios de Avaliação da Redação

De acordo com a *Cartilha do Participante – Redação no ENEM 2024*,¹ a avaliação da redação é realizada com base em cinco competências, descritas a seguir:

- **Competência I** – domínio da modalidade escrita formal da Língua Portuguesa, incluindo ortografia, gramática, concordância e pontuação.
- **Competência II** – compreensão adequada da proposta de redação e desenvolvimento do tema dentro dos limites do texto dissertativo-argumentativo.
- **Competência III** – seleção, organização e interpretação de informações e argumentos em defesa de um ponto de vista.
- **Competência IV** – domínio dos mecanismos linguísticos necessários para garantir coesão e coerência entre frases e parágrafos.
- **Competência V** – elaboração de uma proposta de intervenção relacionada ao tema, respeitando os direitos humanos.

Cada competência recebe notas de 0, 40, 80, 120, 160 ou 200 pontos, atribuídas por dois corretores independentes. Em caso de divergências significativas, a redação é encaminhada a um terceiro avaliador.

3.3 Conjunto de Redações Oficiais do ENEM

Um dos principais desafios na construção de sistemas automáticos de correção de redações para o ENEM é a ausência de redações oficiais disponíveis publicamente. Para contornar essa limitação, foi criado um novo conjunto de dados composto exclusivamente por redações escritas no exame oficial.

As redações foram coletadas por meio de um formulário online distribuído em comunidades universitárias, cursinhos preparatórios e redes pessoais. Os participantes enviaram voluntariamente cópias digitais das redações oficiais acompanhadas das respectivas folhas de correção.

Ao todo, foram coletadas 173 redações de participantes reais, abrangendo oito edições do ENEM (2016, 2018, 2019, 2020, 2021, 2022, 2023 e 2024). Os textos manuscritos foram extraídos e transcritos utilizando Modelos de Linguagem de Grande Porte (LLMs) combinados com técnicas de engenharia de *prompts* para automatizar parcialmente o processo de digitalização e normalização textual. Cada transcrição passou por revisão manual para garantir fidelidade ao texto original.

Após a limpeza dos dados e verificação da consistência entre textos e notas, redações incompletas ou ilegíveis foram excluídas, resultando em um conjunto final de 157 amostras completas. Cada registro contém o texto digitalizado, o ano do exame e as notas oficiais atribuídas às cinco competências.

¹ <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>

3.4 Análise Exploratória do Conjunto de Dados

Para complementar a descrição do conjunto de redações coletadas, foi realizada uma análise exploratória para entender melhor as características do dataset. Os gráficos a seguir mostram algumas propriedades importantes do conjunto.

A distribuição das redações por ano na Figura 3.1 mostra que foi coletado textos de oito edições diferentes do ENEM, entre 2016 e 2024. Com essa distribuição de anos, foi possível separar treinamento e teste usando anos diferentes. Isso evita o viés de treinar e testar redações do mesmo tema, impedindo que o modelo aprenda apenas características específicas daquele tema em vez de generalizar de forma adequada.

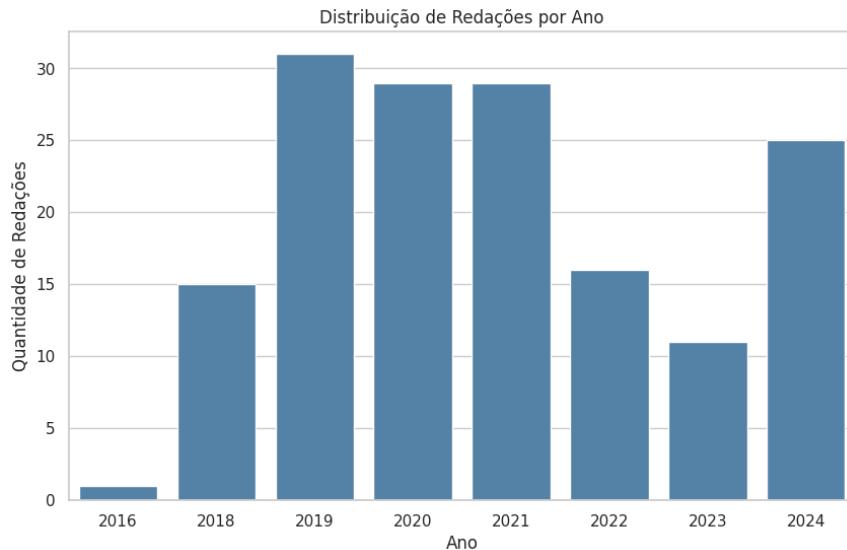


Figura 3.1: Distribuição das redações coletadas por ano

Como o formulário de coleta de dados foi amplamente compartilhado na comunidade universitária, muitos dos participantes que enviaram suas redações obtiveram pontuações relativamente altas no ENEM. Consequentemente, a distribuição das pontuações difere da distribuição geral das pontuações do ENEM, apresentando uma maior concentração de redações com pontuações acima da média nacional, tanto nas competências individuais (Figura 3.2) quanto na nota total (Figura 3.3).

Ao mesmo tempo, a distribuição das pontuações também difere significativamente da distribuição das provas simuladas (SILVEIRA, BARBOSA e MAUÁ, 2024), o que atenua alguns dos vieses introduzidos pela estratégia de coleta de dados.

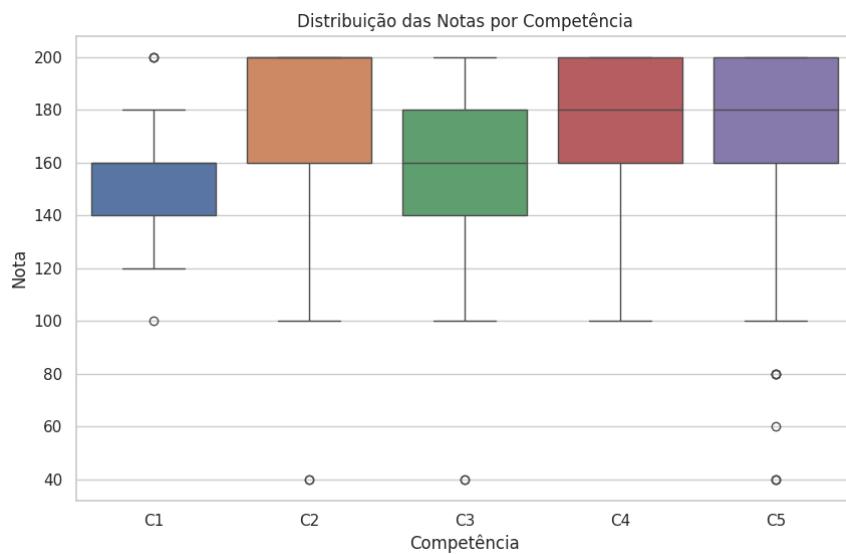


Figura 3.2: Distribuição das notas por competência

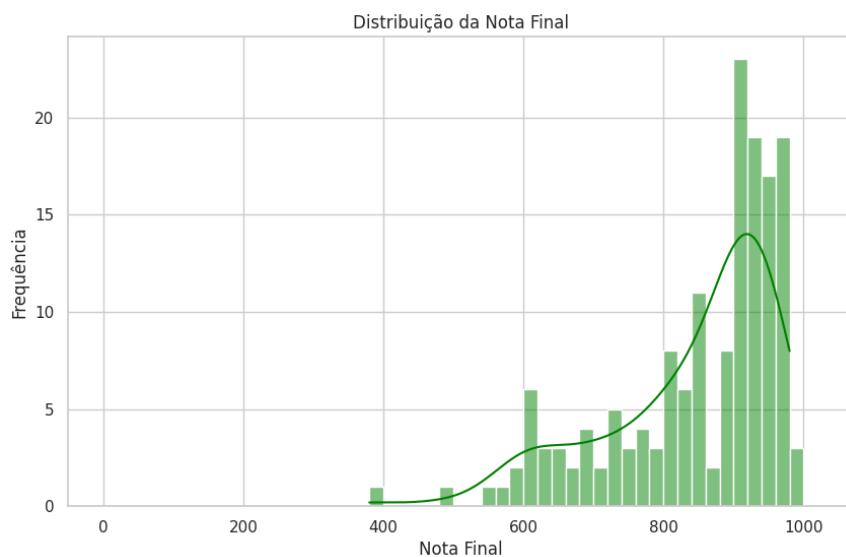


Figura 3.3: Distribuição da nota final das redações

Por fim, ao observar a Figura 3.4, é possível verificar que a distribuição do número de palavras segue aproximadamente uma curva normal, concentrada no intervalo esperado do ENEM, que exige pelo menos 7 linhas e permite até 30.

A análise da relação entre tamanho do texto e nota final (Figura 3.5), evidencia uma correlação positiva entre as variáveis: textos mais extensos tendem a apresentar notas mais altas. No conjunto de dados avaliado, o coeficiente de correlação linear de Pearson entre as duas variáveis foi de aproximadamente 0,65, valor que indica uma associação de magnitude moderada a forte.

Mesmo sabendo que há textos curtos com notas elevadas e textos extensos com notas medianas, essa correlação reforça a ideia de que o tamanho do texto funciona como um

3.4 | ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS

índicio de maior desenvolvimento argumentativo, melhor organização e maior aderência às competências avaliadas.

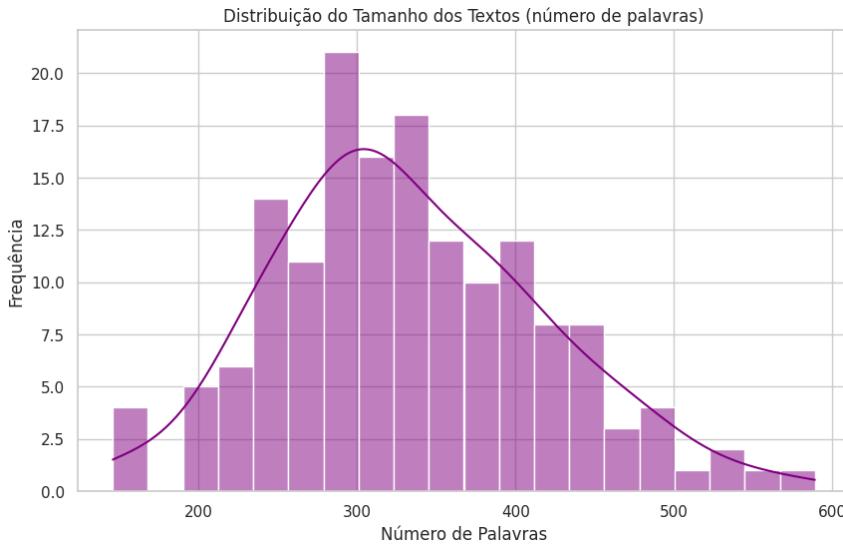


Figura 3.4: Distribuição do número de palavras por redação

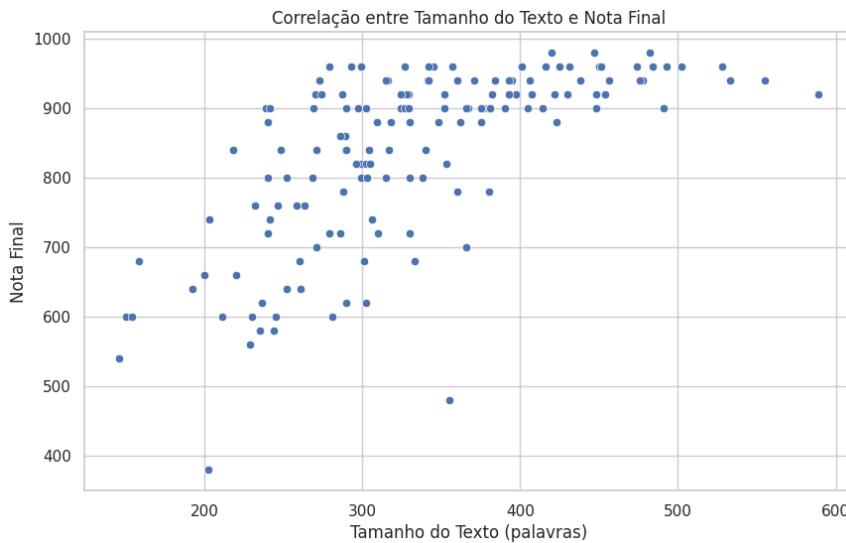


Figura 3.5: Correlação entre tamanho do texto e nota final

Esse comportamento também é importante para modelos automáticos de correção. Embora o tamanho do texto não deva ser usado como único critério, ele pode atuar como um sinal auxiliar na predição, já que, estatisticamente, redações mais extensas têm maior probabilidade de apresentar argumentação mais elaborada. Além disso, mesmo que o ENEM não penalize diretamente textos mais curtos ou mais longos, o atendimento aos critérios de coerência, coesão e argumentação geralmente demanda certo nível de desenvolvimento, o que tende a favorecer textos mais extensos.

Capítulo 4

Metodologia

A metodologia adotada neste trabalho foi estruturada para responder a quatro questões centrais: (i) quão semelhantes são as redações de simulados em relação às redações oficiais do ENEM, (ii) se modelos treinados apenas em dados de simulados conseguem corrigir redações oficiais, (iii) como se comportam modelos treinados exclusivamente no novo conjunto de dados oficial, (iv) se o pré-treinamento em redações simuladas ajuda o modelo a se adaptar melhor às redações reais do ENEM.

Para responder a essas questões, foi realizada uma análise de complexidade textual comparando redações simuladas e oficiais, seguida por experimentos de treinamento e avaliação utilizando diferentes modelos *encoder* da família BERT, além de uma avaliação complementar com o modelo GPT-4o.

4.1 Análise de Complexidade Textual

O objetivo desta análise foi investigar se as características linguísticas que mais influenciam a pontuação de redações do ENEM em conjuntos não oficiais também são relevantes para o conjunto oficial coletado neste trabalho. Para isso, foram extraídas 72 métricas textuais por meio do NILC-Metrix para as 157 redações oficiais do ENEM incluídas no dataset. Essas métricas abrangem aspectos sintáticos, morfológicos e de coesão, tais como: contagem de palavras e sentenças, proporção de classes gramaticais e uso de conectivos discursivos.

Em seguida, foram treinados modelos independentes de regressão linear para cada uma das cinco competências (C1–C5), utilizando 80% das redações para treino e 20% para teste. Todas as variáveis foram padronizadas usando normalização *z-score*, o que permite interpretar diretamente os coeficientes dos modelos como a variação esperada na nota da competência correspondente para um incremento de um desvio padrão no valor da métrica.

4.2 Modelos Encoder e Estratégias de Treinamento

Foram selecionados três modelos base da família BERT:

- **mBERT**, disponibilizado por **DEVLIN et al.**, 2019
- **BERTuguês**, disponibilizado por **MAZZA ZAGO e AGNOLETTI DOS SANTOS PEDOTTI**, 2024
- **BERTimbau**, disponibilizado por **SOUZA et al.**, 2020

Cada um deles foi avaliado em três experimentos distintos:

- **Zero-shot** — desempenho dos modelos pré-treinados em redações não oficiais, disponibilizado por **BARBOSA et al.**, 2025.
- **Pretrained Fine-tuning** — modelos previamente treinados em redações simuladas (JBSC) e posteriormente ajustados no conjunto oficial.¹
- **Exclusive Fine-tuning** — modelos treinados exclusivamente com as 114 redações oficiais.²

Para cada competência (C1–C5), todas essas variantes foram treinadas e avaliadas de forma independente.

Para avaliar o desempenho dos modelos, foram utilizadas duas métricas amplamente empregadas na literatura de correção automática de redações: (1) *Quadratic Weighted Kappa* (QWK), métrica padrão em AES (**DOEWES et al.**, 2023; **FONSECA et al.**, 2018; **MARINHO, ANCHIETÀ et al.**, 2022; **MARINHO, CORDEIRO et al.**, 2022), variando de -1 (discordância total) a 1 (acordo perfeito); e (2) F1 ponderado, que favorece modelos capazes de identificar corretamente as classes mais frequentes (**MELLO et al.**, 2024; **BARBOSA et al.**, 2025).

4.2.1 Fine-tuning

O *fine-tuning* consiste em ajustar um modelo pré-treinado para uma tarefa específica utilizando um conjunto menor e especializado de dados. Em vez de treinar o modelo do zero, o conhecimento linguístico adquirido durante o pré-treinamento é aproveitado e o modelo para capturar as características particulares da tarefa específica é refinado, que neste caso é a predição das notas das competências da redação do ENEM.

Para realizar o *fine-tuning*, o conjunto de dados foi dividido considerando os anos das redações, de modo a evitar que textos do mesmo ano aparecessem simultaneamente em treino e teste. As redações dos anos 2016, 2018, 2022 e 2023 (43 textos) formaram

¹ Coleções no Hugging Face:

- mBERT: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-jbsc-mbert>
- BERTuguês: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-jbsc-bertugues>
- BERTimbau: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-jbsc-bertimbau>

² Coleções no Hugging Face:

- mBERT: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-mbert>
- BERTuguês: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-bertugues>
- BERTimbau: <https://huggingface.co/collections/laisnuto/finetuning-with-official-enem-essays-bertimbau>

o conjunto de teste, enquanto os anos 2019, 2020, 2021 e 2024 (114 textos) formaram o conjunto de treinamento.

Todos os modelos seguiram o mesmo *pipeline*, que incluiu a seleção de hiperparâmetros por meio de *grid search* e a utilização de validação cruzada baseada em anos. Esses procedimentos são detalhados nas subseções seguintes.

4.2.2 Grid Search

Hiperparâmetros como taxa de aprendizado, número de épocas e tamanho do *batch* podem influenciar significativamente o desempenho de um modelo durante o treinamento. Por essa razão, neste trabalho a técnica de *grid search* foi utilizada no processo de *fine-tuning*.

O grid search é um método bastante comum em aprendizado de máquina para ajustar hiperparâmetros. A técnica consiste em testar diversas combinações possíveis desses parâmetros e escolher aquela que obtém o melhor desempenho segundo uma métrica definida.

No presente trabalho, o grid search foi aplicado utilizando o QWK como métrica de otimização. As seguintes combinações foram avaliadas:

- taxa de aprendizado: 10^{-5} ;
- tamanhos de lote: 16 e 32;
- número de épocas: 8, 12 e 16.

Para cada combinação, o modelo era treinado e posteriormente avaliado, e a escolha final foi feita com base no maior valor médio de QWK obtido durante a validação.

4.2.3 Validação Cruzada (*K-Fold Cross-Validation*)

A validação cruzada é uma técnica usada para avaliar o quanto um modelo é capaz de generalizar para dados que ele não “viu” durante o treinamento. Em vez de realizar apenas uma divisão entre treino e teste, a validação cruzada utiliza várias divisões do conjunto de dados, produzindo uma estimativa mais confiável de desempenho, muito útil quando o conjunto é pequeno.

No método *k-fold*, o conjunto de dados é dividido em *k* partes e a cada iteração, o modelo é treinado em *k-1* partes e validado na parte restante. Esse processo se repete *k* vezes, alternando qual parte é usada para validação. Assim, cada instância do conjunto é utilizada exatamente uma vez como dado de validação.

Essa abordagem permite verificar se o modelo está aprendendo padrões reais ou apenas decorando os dados de treino (*overfitting*), além de fornecer uma medida mais estável de desempenho, já que a pontuação final corresponde à média das *k* iterações.

Como o conjunto de dados é pequeno e a variação entre anos pode afetar o conteúdo das redações, foi adotada uma validação cruzada baseada nos anos do exame. O treinamento foi dividido em 4 *folds*, cada um correspondendo a uma combinação diferente de anos usados para validação, garantindo que nenhuma redação do mesmo ano aparecesse

simultaneamente em treino e validação. Essa estratégia permite que o modelo seja avaliado em contextos variados, reduzindo o risco de *overfitting* a temas específicos de um ano.

O desempenho final de cada configuração foi definido como a média dos valores de QWK obtidos nos quatro folds, refletindo melhor a performance esperada em dados não vistos.

4.2.4 Testes Estatísticos

O método de *bootstrap* aplicado à métrica *Quadratic Weighted Kappa* (QWK) foi utilizado para quantificar o intervalo de confiança associado ao desempenho dos modelos.

Em cada configuração (modelo × experimento × competência), foram realizadas $B = 3000$ reamostragens com reposição a partir do conjunto de teste. Em cada amostra, o QWK foi recalculado, produzindo uma distribuição empírica de valores. O processo pode ser descrito da seguinte forma:

```

1: function BOOTSTRAPQWK( $y_{\text{true}}$ ,  $y_{\text{pred}}$ ,  $B$ )
2:    $D \leftarrow []$ 
3:    $N \leftarrow \text{tamanho}(y_{\text{true}})$ 
4:   for  $b \leftarrow 1$  até  $B$  do
5:      $I \leftarrow \text{sorteio\_com\_reposicao}(N, N)$ 
6:      $q \leftarrow \text{QWK}(y_{\text{true}}[I], y_{\text{pred}}[I])$ 
7:      $D.append(q)$ 
8:   end for
9:   return média( $D$ ), percentil( $D$ , 2,5%), percentil( $D$ , 97,5%)
10: end function
```

Para comparar dois experimentos, foi analisado se os intervalos de confiança se sobrepõem. Quando os intervalos não se sobrepõem, a diferença é classificada como estatisticamente significativa. Esse critério foi aplicado às comparações entre *Zero-shot* e *Pretrained Fine-tuning*, e entre *Exclusive Fine-tuning* e *Pretrained Fine-tuning*.

4.3 Escalas de Notas do ENEM

Cada redação oficial do ENEM recebe duas notas independentes por competência, variando entre 0 e 200 em incrementos de 40. A nota final de cada competência é a média dessas duas avaliações, o que pode gerar valores intermediários (por exemplo, 20, 60, 100, 140, 180). Assim, surge um descompasso estrutural: os modelos produzem valores apenas múltiplos de 40, enquanto os rótulos oficiais incluem intervalos de 20 pontos.

Para tornar os rótulos compatíveis com o espaço de saída dos modelos durante o treinamento, todas as notas finais que não fossem múltiplas de 40 foram arredondadas para o múltiplo inferior mais próximo antes do *fine-tuning*. No entanto, durante a avaliação, a comparação foi feita sempre entre valores reais na escala de 20 pontos e previsões na escala de 40 pontos, sem qualquer transformação adicional.

Outro fator importante é que a média de duas avaliações humanas não permite inferir o grau real de concordância entre os corretores. Por exemplo, um valor final de 80 pode

significar total concordância (80 e 80) ou discordância elevada (40 e 120). Assim, mesmo quando previsões e médias coincidem, isso não necessariamente significa que o modelo imitou corretamente o padrão dos avaliadores.

4.3.1 Protocolos de Avaliação e Estratégias de Arredondamento

A métrica QWK é sensível a como as classes são distribuídas e comparadas. Diferentes formas de converter as notas de 20 para 40 pontos resultam em valores de QWK distintos (DOEWES *et al.*, 2023). Por isso, diferentes protocolos de avaliação foram testados, variando as regras de arredondamento dos rótulos reais.

Embora esses protocolos produzam flutuações nas métricas, todos levaram às mesmas conclusões gerais: modelos pré-treinados e posteriormente ajustados superam as demais variantes. Assim, a análise principal deste trabalho considera apenas o protocolo *no changes*, que compara diretamente rótulos reais na escala de 20 pontos com previsões na escala de 40 pontos, sem arredondamento durante a avaliação.

Uma análise detalhada sobre o impacto de cada protocolo e das diferentes regras de arredondamento é apresentada no Capítulo 6.

4.4 Avaliação com o Modelo GPT-4o

Além dos modelos *encoder* da família BERT, um experimento complementar utilizando o modelo GPT-4o foi feito, com o objetivo de comparar o desempenho dos modelos *decoder* de grande porte. A configuração adotada seguiu o padrão metodológico apresentado em BARBOSA *et al.* (2025).

O experimento com GPT-4o envolveu três decisões principais, conforme descrito no artigo de referência:

- escolha do modelo,
- uso ou não de um *prompt* explícito,
- seleção da diretriz de avaliação (*guideline*) para orientar a atribuição das notas.

O GPT-4o foi escolhido por ter se destacado em estudos anteriores de avaliação textual. Assim como em BARBOSA *et al.* (2025), o *extended prompt* não foi usado, pois ele melhora apenas as competências C2 e C3 e não é diretamente comparável ao cenário adotado para os modelos *encoder*. Dessa forma, o modelo recebeu apenas o texto da redação, sem informações sobre tema ou textos de apoio. A diretriz utilizada foi a *Student guideline*, que apresentou maior estabilidade nos experimentos anteriores e gerou respostas mais consistentes com padrões humanos de correção.

Nessa configuração, o GPT-4o atua em regime totalmente *zero-shot*. Cada redação foi enviada ao modelo de forma independente, e a resposta retornada consistiu em cinco valores correspondentes às competências C1–C5. As saídas foram então mapeadas para as categorias oficiais do ENEM {0, 40, 80, 120, 160, 200} e avaliadas com a métrica QWK.

Capítulo 5

Resultados

Este capítulo apresenta os principais resultados obtidos nos experimentos conduzidos, organizados em duas partes: (i) uma análise das métricas textuais derivadas do uso do NILC-Metrix e (ii) uma comparação detalhada do desempenho dos modelos BERT-base sob diferentes configurações de treinamento. Ao final, discute-se o que esses resultados significam considerando as características do conjunto de dados e das métricas utilizadas.

5.1 Métricas Textuais

Para entender melhor a relação entre características linguísticas com as notas atribuídas pelos corretores humanos em cada competência do ENEM, um modelo de regressão linear foi treinado utilizando apenas as redações oficiais do conjunto coletado. Em todas as competências, observou-se um padrão consistente: as cinco métricas com maior impacto absoluto nos coeficientes foram sempre as mesmas:

- noun_ratio – proporção de substantivos;
- content_words – proporção de palavras de conteúdo;
- verbs – proporção de verbos;
- adjective_ratio – proporção de adjetivos;
- adverbs – proporção de advérbios.

Esse comportamento reproduz de forma muito parecida os achados de SILVEIRA, BARBOSA, COSTA *et al.*, 2025, em que as mesmas métricas do NILC-Metrix figuram como preditores principais em um conjunto de dados formado apenas por redações simuladas. Essa semelhança sugere que as estruturas linguísticas e os padrões estilísticos do conjunto de redações oficiais deste trabalho são bastante próximos dos encontrados em textos simulados, reforçando a utilidade desses para pré-treinamento de modelos.

Isso indica que as propriedades lexicais capturadas pelo NILC-Metrix refletem aspectos estruturais da escrita que se mantêm estáveis mesmo em contextos distintos de coleta

5.2 Modelos *Encoder*

Os resultados de desempenho dos modelos são apresentados nas Tabelas 5.1 e 5.3. A Tabela 5.1 apresenta os valores de *Quadratic Weighted Kappa* (QWK) e a Tabela 5.3 mostra os resultados em termos de F1 ponderado.

A primeira linha de cada tabela, *Mock Exams*, corresponde ao desempenho original dos modelos pré-treinados no trabalho de referência BARBOSA *et al.*, 2025, calculados utilizando as notas individuais de dois corretores humanos e computando a média entre os dois valores resultantes. A linha *Zero-shot* mostra o desempenho desses mesmos modelos quando aplicados diretamente às redações oficiais sem qualquer ajuste. Já a linha *Pretrained FT* apresenta os resultados obtidos após realizar fine-tuning sobre as redações oficiais. Por fim, a linha *Exclusive FT* corresponde aos modelos que passaram apenas pelo *fine-tuning* no novo conjunto de dados, sem qualquer pré-treinamento em redações simuladas.

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
<i>Mock Exams</i>	.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 5.1: Resultados de QWK para todos os modelos e configurações de treinamento

Para aprofundar a interpretação, foram analisadas as diferenças entre as condições experimentais. A Tabela 5.2 resume os deltas entre: (i) *Zero-shot* e *Pretrained FT*, indicando o ganho proporcionado pelo *fine-tuning*, e (ii) *Pretrained FT* e *Exclusive FT*, evidenciando o impacto do pré-treinamento em redações simuladas.

	mBERT	BERTuguês	BERTimbau	avg		mBERT	BERTuguês	BERTimbau	avg
$\Delta C1$	-.045	.085	.015	.018	$\Delta C1$.456	.394	.393	.414
$\Delta C2$.115	.074	.102	.097	$\Delta C2$.116	.136	.332	.194
$\Delta C3$.072	-.004	.066	.044	$\Delta C3$.519	.136	.453	.369
$\Delta C4$.101	.066	.030	.065	$\Delta C4$.455	.370	.206	.343
$\Delta C5$	-.030	.218	.316	.168	$\Delta C5$	-.359	.469	.039	.049
avg.	.042	.087	.105	.078	avg.	.237	.301	.284	.274

Tabela 5.2: À esquerda: variação de QWK ao comparar Fine-tuning vs. Zero-shot. À direita: variação entre Pretrained FT e Exclusive FT

Embora o QWK seja a métrica principal para avaliar a concordância com os corretores humanos, ele não captura certas nuances do comportamento dos modelos, principalmente em conjuntos de dados desbalanceados como o utilizado neste trabalho. Por isso, o F1 ponderado também foi analisado, que dá maior peso às classes mais frequentes e oferece uma visão complementar. A Tabela 5.3 apresenta esses resultados para todos os modelos e experimentos.

5.3 | COMPARAÇÃO COM O MODELO GPT-4O

	mBERT					BERTuguês					BERTimbau							
	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean
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

Tabela 5.3: Resultados do F1-ponderado para todos os modelos e configurações de treinamento

Os valores de F1 ponderado apresentados na Tabela 5.3 mostram que o *Pretrained FT* geralmente obtém resultados superiores ao *Zero-shot* e que o *Exclusive FT*, padrão parecido com o observado na análise do QWK. Para facilitar a visualização dessas diferenças, a Tabela 5.4 apresenta os deltas entre as condições avaliadas

	mBERT	BERTuguês	BERTimbau	avg		mBERT	BERTuguês	BERTimbau	avg
$\Delta C1$.002	.076	.058	.045	$\Delta C1$.104	.083	.089	.092
$\Delta C2$.059	.060	.074	.064	$\Delta C2$.022	.030	.089	.047
$\Delta C3$.066	-.026	-.010	.010	$\Delta C3$.121	.034	.095	.083
$\Delta C4$.020	.007	-.005	.007	$\Delta C4$.094	.096	.083	.091
$\Delta C5$.178	.101	.286	.188	$\Delta C5$	-.206	.186	-.014	-.011
avg.	.065	.044	.081	.063	avg.	.027	.086	.068	.060

Tabela 5.4: À esquerda: variação do F1-ponderado ao comparar Fine-tuning vs. Zero-shot. À direita: variação entre Pretrained FT e Exclusive FT

5.2.1 Testes Estatísticos

A Tabela 5.5 resume, para cada modelo e competência, se a melhoria observada no experimento de *Pretrained Fine-tuning* usando como métrica o QWK é estatisticamente significativa, mesmo considerando o tamanho reduzido do conjunto de teste.

	mBERT					BERTuguês					BERTimbau				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Zero-shot vs. Pretrained FT	Não	Não	Não	Não	Sim	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não
Exclusive FT vs. Pretrained FT	Sim	Não	Sim	Sim	Sim	Não	Não	Não	Não	Sim	Sim	Sim	Sim	Não	Não

Tabela 5.5: Significância estatística das diferenças entre experimentos, baseada em intervalos de confiança obtidos via bootstrap. Células verdes indicam diferença significativa (ICs não se sobrepõem).

5.3 Comparação com o Modelo GPT-4o

Para complementar a análise dos modelos *encoder*, o desempenho do modelo GPT-4o em regime *Zero-shot* também foi avaliado, utilizando apenas o texto da redação e a *Student guideline*. O objetivo foi comparar sua capacidade de predição com os modelos BERT sem qualquer tipo de ajuste.

A Tabela 5.6 resume os valores de QWK para os cenários de interesse: os três modelos BERT em configuração *Zero-shot* e o GPT-4o.

Modelo	C1	C2	C3	C4	C5	avg.
mBERT Zero-shot	.501	.349	.447	.626	.030	.391
BERTuguês Zero-shot	.382	.375	.290	.619	.271	.387
BERTimbau Zero-shot	.378	.253	.387	.628	.069	.343
GPT-4o	.327	.362	.365	.330	.268	.330

Tabela 5.6: Desempenho em QWK para mBERT, BERTuguês, BERTimbau e GPT-4o na configuração zero-shot.

A Tabela 5.7 resume as diferenças dos valores de QWK entre o GPT-4o e os os três modelos BERT no experimento *zero-shot*. Dessa forma, os valores positivos indicam vantagem do GPT-4o.

	mBERT	BERTuguês	BERTimbau	avg
$\Delta C1$	-.174	-.055	-.051	-.093
$\Delta C2$.013	-.013	.109	.036
$\Delta C3$	-.082	.075	-.022	-.010
$\Delta C4$	-.296	-.289	-.298	-.294
$\Delta C5$.238	-.003	.199	.145
avg.	-.060	-.057	-.012	-.043

Tabela 5.7: Variação de QWK entre GPT-4o e modelos BERT em modo zero-shot

5.4 Discussão

Nesta seção, são discutidos os resultados obtidos, as implicações do uso de redações simuladas no treinamento de sistemas de avaliação automática, bem como algumas limitações da metodologia utilizada.

Os resultados da análise textual mostram que as redações simuladas apresentam características linguísticas muito semelhantes às redações oficiais do ENEM. Apesar de as redações oficiais serem produzidas em um contexto específico de prova, com tempo limitado e condições socioambientais particulares, as redações simuladas conseguem refletir, pelo menos do ponto de vista textual, padrões de estrutura e de estilo bastante próximos.

Ao analisar o desempenho dos modelos *encoder*, é possível ver que no primeiro experimento (*Zero-shot*) os modelos treinados em redações simuladas alcançam um desempenho razoável nas redações oficiais do ENEM. Isso sugere que as características linguísticas dos dois domínios são suficientemente próximas para permitir transferência de aprendizado, mesmo sem qualquer ajuste adicional.

Além disso, a comparação entre *Zero-shot* e *Pretrained FT* mostra que o fine-tuning gera ganhos consistentes: em 12 dos 15 casos (cinco competências para três modelos), o valor de QWK aumenta, e o mesmo ocorre para o F1 ponderado. Esses resultados apresentados nas tabelas 5.2 e 5.4 indicam que, mesmo com um conjunto reduzido de redações oficiais, os modelos conseguem se adaptar ao padrão de correção do ENEM e melhorar suas previsões.

O contraste mais expressivo ocorre entre *Pretrained FT* e *Exclusive FT*. Em 14 de 15 comparações, o modelo pré-treinado em redações simuladas e depois ajustado supera aquele que passou apenas pelo *fine-tuning* no conjunto reduzido de redações oficiais, muitas vezes com margens substanciais. Isso evidencia que os modelos não conseguem aprender padrões complexos de pontuação do ENEM apenas a partir de um conjunto reduzido, o pré-treinamento em redações não oficiais fornece informações essenciais sobre estrutura e estilo que são indispensáveis para um bom desempenho.

Esses padrões também aparecem nos testes estatísticos. Embora os resultados do *Pretrained FT* geralmente sejam melhores do que os do *Zero-shot* (Tabela 5.2), o bootstrap mostra que essa diferença não é significativa. Isso, significa que os modelos pré-treinados em redações simuladas já chegam muito próximos do desempenho obtido após o ajuste com dados oficiais. Como o conjunto de teste é pequeno, fica difícil identificar melhorias pequenas ou moderadas, mas o fato de não ter diferença significativa reforça que o treinamento com simulados produz modelos fortes desde o início. Já na comparação entre *Pretrained FT* e *Exclusive FT*, as diferenças são significativas em várias competências, mostrando que o pré-treinamento com redações simuladas realmente melhora o modelo e que esses ganhos não são apenas variações ao acaso. Em resumo, a análise estatística reforça que o pré-treinamento ajuda a obter bons resultados no contexto do ENEM.

Uma análise mais detalhada das matrizes de confusão ajuda a compreender como esses comportamentos se manifestam. Elas revelam não apenas o nível agregado das métricas, mas a forma como cada modelo distribui suas previsões entre as classes. Essa análise revelou limitações importantes relacionadas ao uso da métrica QWK.

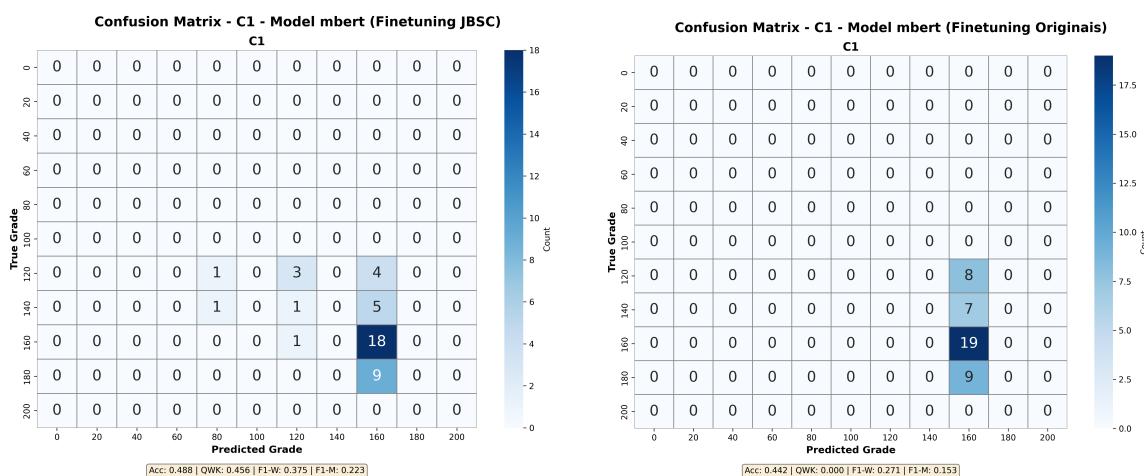


Figura 5.1: *mBERT* – Matrizes de confusão para C1 sob (esq.) Fine-tuning Pré-treinado e (dir.) Fine-tuning Exclusivo.

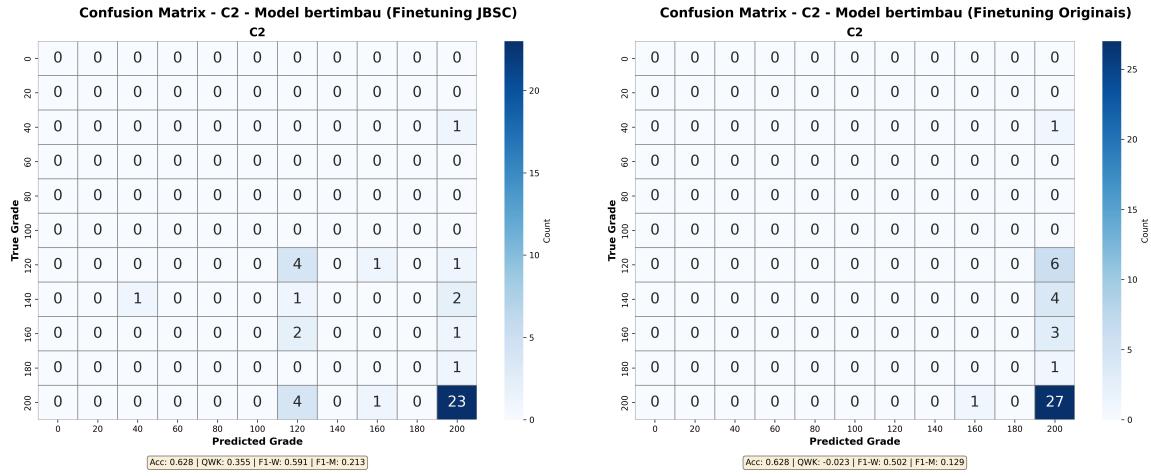


Figura 5.2: BERTimbau – Matrizes de confusão para C2 sob (esq.) Fine-tuning Pré-treinado e (dir.) Fine-tuning Exclusivo.

Como mostrado nas Figuras 5.1 e 5.2, em cenários de distribuição altamente concentrada em poucas classes, caso comum do conjunto deste trabalho, composto majoritariamente por redações de alto desempenho, o QWK tende a penalizar modelos que fazem previsões homogêneas. Mesmo quando o modelo identifica corretamente a classe majoritária (como prever sempre 160 ou 200 pontos), o QWK pode assumir valores próximos de zero ou até negativos. Isso acontece não porque o modelo falha totalmente, mas porque a métrica penaliza fortemente previsões pouco diversas, mesmo quando são compatíveis com o padrão real dos dados.

Esse padrão se repete em diversas competências e modelos. Para fornecer uma visão completa, todas as matrizes de confusão, cobrindo as cinco competências, os três modelos (mBERT, BERTuguês e BERTimbau) e os três cenários de avaliação (*Zero-shot*, *Pretrained* *Fine-tuning* e *Exclusive Fine-tuning*), estão reunidas no Apêndice A.

Por esse motivo, o F1 ponderado também foi analisado. Embora QWK e F1 não sejam diretamente comparáveis, o F1 ponderado mostra desempenhos relativamente melhores para os modelos *Exclusive FT*, pois ele recompensa previsões corretas da classe majoritária em datasets desbalanceados. Ainda assim, os modelos *Pretrained FT* mantêm desempenho superior, sugerindo que capturaram não apenas o padrão dominante, mas também nuances importantes das competências do ENEM.

Além dos modelos *encoder*, o GPT-4o *zero-shot* também foi avaliado. Mesmo sendo um modelo muito maior, com estimativa de centenas de bilhões de parâmetros, em contraste com aproximadamente 110 milhões nos modelos BERT, seus resultados não superaram os encoders. Em várias competências, o desempenho do GPT-4o foi semelhante ou inferior ao dos modelos *zero-shot*, e ficou consistentemente abaixo dos modelos pré-treinados e ajustados. Como o GPT-4o é mais caro de executar e exige mais recursos computacionais, enquanto os modelos BERT são gratuitos e podem ser usados localmente, o uso do GPT-4o para correção automática não apresenta bom custo-benefício neste contexto. Assim, para a tarefa de avaliação de redações do ENEM, modelos menores e bem treinados oferecem uma relação desempenho-custo mais favorável do que um LLM de grande porte.

Portanto, todos esses achados demonstram que redações não oficiais são um recurso valioso para pré-treinamento. Como redações oficiais são difíceis de obter, o uso de simulados permite treinar modelos em larga escala e, posteriormente, adaptá-los com um pequeno conjunto de textos reais, estratégia que se mostrou claramente vantajosa nos experimentos deste trabalho.

Capítulo 6

Avaliação Complementar dos Efeitos de Arredondamento

Embora o Capítulo 5 apresente os resultados principais utilizando um único protocolo de avaliação sem arredondamento, diferentes estratégias de arredondamento entre as escalas de 20 e 40 pontos podem influenciar diretamente o valor do QWK. Como esse comportamento é relevante para interpretar corretamente o desempenho dos modelos, este capítulo apresenta uma análise complementar. O objetivo é mostrar que, apesar de pequenas variações numéricas, as conclusões gerais permanecem as mesmas: o *Pretrained Fine-tuning* supera consistentemente as demais configurações.

As tabelas 6.1, 6.2, 6.3 resumem o desempenho dos modelos nos 3 experimentos usando os diferentes protocolos avaliados.

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean
<i>Mock Exams</i>	.520	.220	.350	.500	.000	.318	.620	.330	.290	.540	.360	.428	.600	.360	.350	.550	.630	.498
no changes	.501	.349	.447	.626	.030	.391	.382	.375	.290	.619	.271	.387	.378	.253	.387	.628	.069	.343
duplicate	.438	.346	.420	.587	.030	.364	.355	.370	.270	.579	.261	.367	.343	.250	.359	.587	.065	.321
floor	.479	.367	.397	.632	.034	.382	.340	.371	.290	.591	.306	.380	.305	.262	.309	.611	.114	.320
ceiling	.405	.324	.437	.543	.027	.347	.369	.369	.254	.568	.215	.355	.385	.236	.410	.562	.022	.323
amplitude	.096	.043	.050	.089	.007	.057	.042	.006	.036	.051	.091	.045	.080	.026	.101	.066	.092	.073

Tabela 6.1: QWK por competência para modelos *Zero-shot* sob diferentes protocolos de avaliação.

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean
<i>Mock Exams</i>	.520	.220	.350	.500	.000	.318	.620	.330	.290	.540	.360	.428	.600	.360	.350	.550	.630	.498
no changes	.456	.464	.519	.727	.000	.433	.467	.449	.286	.685	.489	.475	.393	.355	.453	.658	.385	.449
duplicate	.386	.456	.475	.687	.000	.401	.384	.442	.263	.642	.471	.440	.331	.350	.411	.618	.371	.416
floor	.448	.473	.458	.686	.000	.413	.501	.456	.284	.625	.491	.471	.410	.369	.402	.639	.411	.446
ceiling	.343	.438	.488	.688	.000	.392	.308	.426	.246	.660	.450	.418	.279	.330	.420	.596	.329	.391
amplitude	.113	.035	.061	.041	.000	.050	.193	.030	.040	.060	.041	.073	.131	.039	.051	.062	.082	.073

Tabela 6.2: QWK por competência para modelos *Pretrained Fine-tuned* em diferentes protocolos de avaliação.

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean	C1	C2	C3	C4	C5	Mean
<i>Mock Exams</i>	.520	.220	.350	.500	.000	.318	.620	.330	.290	.540	.360	.428	.600	.360	.350	.550	.630	.498
no changes	.000	.348	.000	.272	.359	.196	.073	.313	.150	.315	.020	.174	.000	-.023	.000	.452	.346	.155
duplicate	.000	.342	.000	.254	.343	.188	.057	.306	.129	.288	.019	.160	.000	-.022	.000	.418	.333	.146
floor	.000	.344	.000	.251	.366	.192	-.006	.303	.130	.272	.023	.144	.000	-.022	.000	.365	.351	.139
ceiling	.000	.339	.000	.258	.318	.183	.106	.310	.128	.306	.014	.173	.000	-.023	.000	.490	.312	.156
amplitude	.000	.009	.000	.021	.048	.016	.112	.010	.021	.043	.009	.039	.000	.001	.000	.125	.039	.033

Tabela 6.3: QWK por competência para modelos *Exclusive Fine-tuned* em diferentes protocolos de avaliação.

6.1 Influência das Estratégias de Arredondamento no QWK

Nesta análise complementar, é investigado como diferentes estratégias de arredondamento afetam o valor da métrica QWK. Como os rótulos oficiais são atribuídos em uma escala de 20 pontos e os modelos geram saídas em múltiplos de 40, torna-se necessário alinhar essas escalas durante a avaliação.

Foram consideradas quatro estratégias:

- **no changes:** mantém os rótulos reais na escala original de 20 pontos. Nesse caso, o modelo é penalizado por não ter acesso à mesma granularidade, e o desempenho tende a ser subestimado. Esta foi a estratégia adotada nos resultados principais do trabalho.
- **duplicate bounds:** cada rótulo real é expandido em duas versões, uma arredondada para cima e outra para baixo para o múltiplo mais próximo de 40. Essa estratégia tenta preservar a distribuição original, mas gera casos em que o modelo acerta um dos rótulos e inevitavelmente erra o outro.
- **floor:** arredonda todo rótulo para o múltiplo de 40 imediatamente inferior. Essa abordagem agrupa valores e tende a comprimir a distribuição.
- **ceiling:** arredonda todo rótulo para o múltiplo de 40 imediatamente superior, também resultando em maior concentração em poucas classes.

As Tabelas 6.1, 6.3 e 6.2 apresentam os resultados de QWK por competência para cada família de modelos nos três cenários avaliados: *Zero-shot*, *Exclusive Fine-tuning* e *Pretrained Fine-tuning*. Em cada tabela, também é reportada a amplitude, definida como a diferença entre o maior e o menor valor observado em cada competência, refletindo a sensibilidade de cada modelo às diferentes estratégias de avaliação.

Observa-se que, em praticamente todas as competências e experimentos, os maiores valores são obtidos ou pela estratégia *no changes*, ou pela estratégia *floor*. O fato de *no changes* produzir QWK elevados pode estar relacionado à maior dispersão dos rótulos nessa escala, já que o QWK penaliza fortemente avaliações em que a distribuição é muito concentrada. O bom desempenho ocasional da estratégia *floor* é menos intuitivo e possivelmente decorrente de uma coincidência entre a distribuição final dos rótulos arredondados e a forma como cada modelo distribui suas previsões.

Outro ponto importante é que a escolha da estratégia afeta cada modelo de maneira distinta. Mesmo dentro de uma mesma competência, a amplitude varia entre os três modelos, mostrando que a sensibilidade ao protocolo de arredondamento não é uniforme. A amplitude máxima observada (0.131 para o BERTimbau em C1) é suficientemente alta para que a escolha do protocolo influencie a interpretação do desempenho do modelo.

Também é possível notar que a ordenação relativa entre modelos muda dependendo da estratégia. Por exemplo no *Pretrained Fine-tuning*, em C1, o BERTuguês supera o mBERT em duas estratégias e fica atrás dele em outras duas. Isso reforça que comparações diretas entre modelos podem ser enviesadas dependendo do protocolo adotado.

6.2 Conclusão da Análise Complementar

De forma geral, esta análise mostra que diferentes protocolos de arredondamento podem alterar o valor absoluto do QWK e até modificar a ordem relativa dos modelos. Ainda assim, para as questões centrais deste trabalho, as conclusões permanecem consistentes: modelos *Pretrained Fine-tuned* superam suas variantes *Zero-shot* e, principalmente, as versões *Exclusive FT*, independentemente da estratégia usada.

Assim, embora seja importante reconhecer o impacto do protocolo de avaliação, os resultados globais sustentam a robustez das conclusões apresentadas no Capítulo 5.

Capítulo 7

Conclusão

Este trabalho investigou a utilidade de redações de simulados para o treinamento de modelos de AES aplicados ao ENEM, preenchendo uma lacuna importante da literatura: a ausência de análises baseadas em redações oficiais do exame. A partir de um novo conjunto de dados com 157 redações reais e notas oficiais, foram avaliadas as seguintes questões: (i) se simulados realmente se assemelham às redações do ENEM, (ii) se modelos treinados apenas nesses dados conseguem generalizar para textos oficiais, (iii) se modelos treinados exclusivamente em um conjunto reduzido de redações reais são suficientes para a tarefa e, por fim, (iv) se o pré-treinamento em simulados melhora a capacidade dos modelos de se adaptarem às redações oficiais.

Para responder à primeira questão, foram comparadas as características linguísticas do conjunto utilizado com aquelas de conjuntos de simulados anteriores, empregando o NILC-Metrix. Os resultados mostraram que as métricas textuais mais relevantes foram praticamente as mesmas em ambos os domínios, sugerindo que redações simuladas podem se aproximar da estrutura e do estilo característicos da escrita oficial do ENEM. Essa análise reforça que simulados constituem um recurso útil para o pré-treinamento de modelos de AES.

Quanto à segunda questão, foi observado que os experimentos *Zero-shot* nos três modelos baseados em *encoders* (mBERT, BERTuguês e BERTimbau) já generalizam razoavelmente bem para as redações oficiais do ENEM sem qualquer ajuste adicional. Mesmo quando comparados a modelos muito maiores e mais caros, como o GPT-4o, seu desempenho permanece competitivo. Isso indica que parte do conhecimento aprendido a partir de simulados é transferido para textos reais, dando suporte ao uso de simulados como base de treinamento inicial.

Em seguida, foi investigado como se comportam os modelos que passaram apenas pelo *fine-tuning* no conjunto oficial de 157 redações. Esses modelos, correspondentes ao experimento de *Exclusive FT*, apresentaram desempenho inferior em praticamente todas as competências quando comparados aos modelos pré-treinados. Esse resultado era esperado, dado o tamanho reduzido do conjunto de dados oficial, e confirma que o volume limitado de textos reais dificulta a aprendizagem de padrões complexos de avaliação de redações.

Por fim, foi avaliado se o pré-treinamento em redações simuladas auxilia os modelos a se adaptarem melhor às redações reais. A comparação entre o *fine-tuning* pré-treinado em redações simuladas (*Pretrained FT*) e apenas o *fine-tuning* (*Exclusive FT*) evidenciou a diferença mais expressiva de todo o estudo. Os modelos que passaram por pré-treinamento apresentaram desempenho substancialmente superior, com ganho médio de cerca de 0,27 no QWK. Além disso, análises estatísticas baseadas em *bootstrap* indicaram que essa melhoria é significativa, mesmo com tamanho reduzido do conjunto oficial, o que reforça que o pré-treinamento não apenas ajuda, mas é essencial para obter bons resultados no contexto do ENEM.

Dessa forma, o conjunto de achados deste trabalho sugere uma abordagem prática para a construção de sistemas de avaliação automática de redações (AES) para o ENEM: utilizar grandes conjuntos de redações simuladas para o pré-treinamento e, em seguida, adaptar os modelos utilizando um conjunto menor de redações oficiais.

Apêndice A

Matrizes de Confusão por Modelo

Este apêndice reúne todas as matrizes de confusão geradas nos experimentos para os três modelos avaliados: mBERT, BERTuguês e BERTimbau. Para cada modelo, apresentamos as matrizes nos três cenários:

1. *Zero-shot* – inferência direta sem ajuste;
 2. *Pretrained Fine-tuning (JBSC)* – modelos pré-treinados em simulados e ajustados com dados oficiais.
 3. *Exclusive Fine-tuning* – fine-tuning com as redações oficiais, sem nenhum pré-treinamento;

Cada imagem contém cinco matrizes mostrando os resultados das previsões nas 5 competências do ENEM

A.1 mBERT

Zero-shot

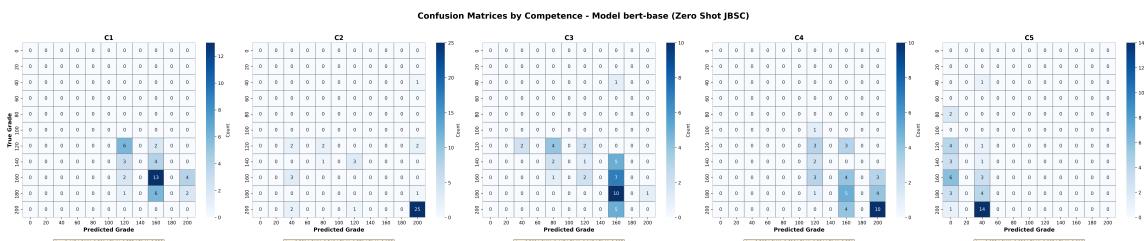


Figura A.1: Matrizes de confusão: *mBERT* – Zero-shot (C1–C5)

Pretrained Fine-tuning (JBSC)

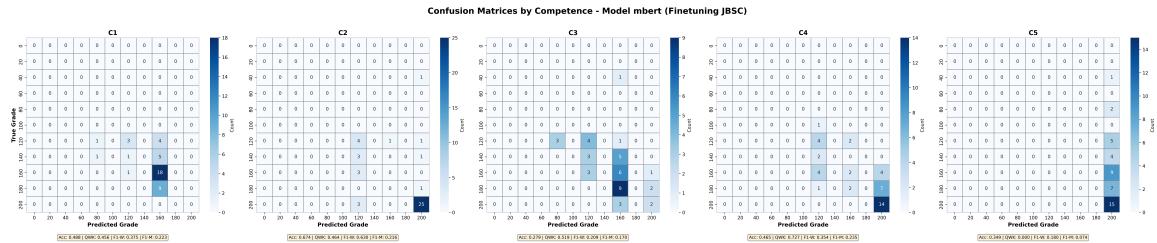


Figura A.2: Matrizes de confusão: mBERT – Fine-tuning após pré-treinamento em simulados (C1–C5)

Exclusive Fine-tuning

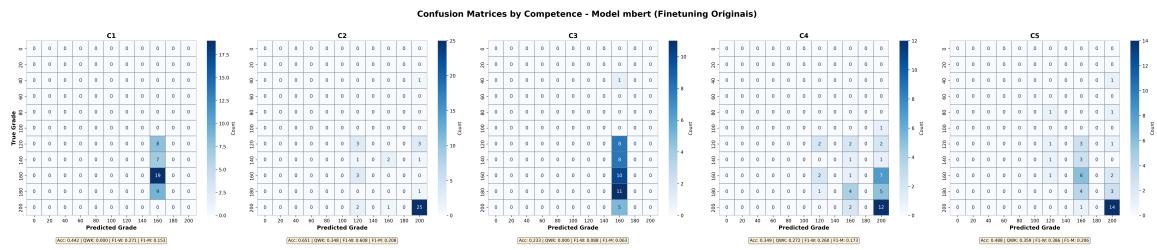


Figura A.3: Matrizes de confusão: mBERT – Fine-tuning Exclusivo (C1–C5)

A.2 BERTuguês

Zero-shot

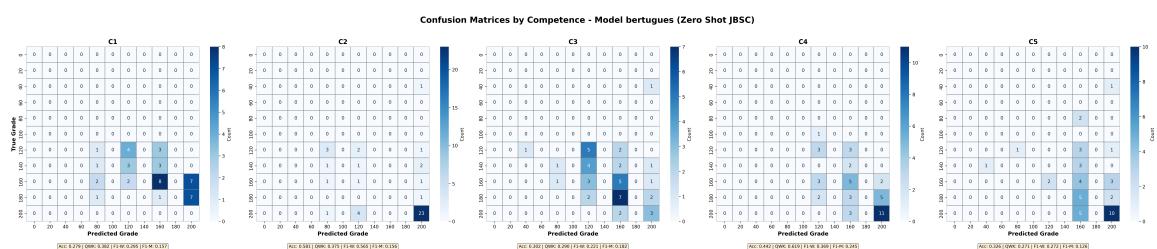


Figura A.4: Matrizes de confusão: BERTuguês – Zero-shot (C1–C5)

A.3 | BERTIMBAU

Pretrained Fine-tuning (JBSC)

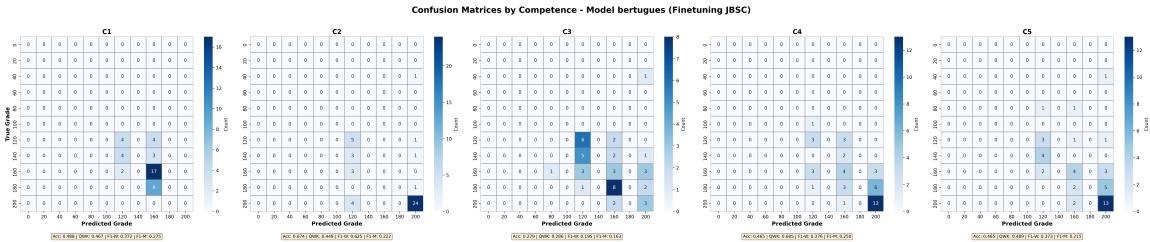


Figura A.5: Matrizes de confusão: BERTuguês – Fine-tuning após pré-treinamento em simulados (C1–C5)

Exclusive Fine-tuning

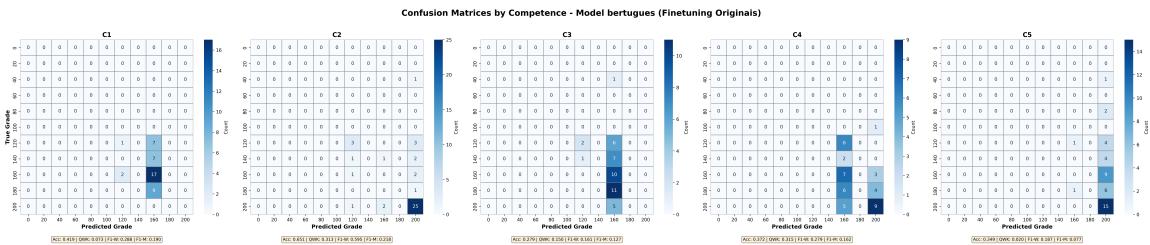


Figura A.6: Matrizes de confusão: BERTuguês – Fine-tuning Exclusivo (C1–C5)

A.3 BERTimbau

Zero-shot

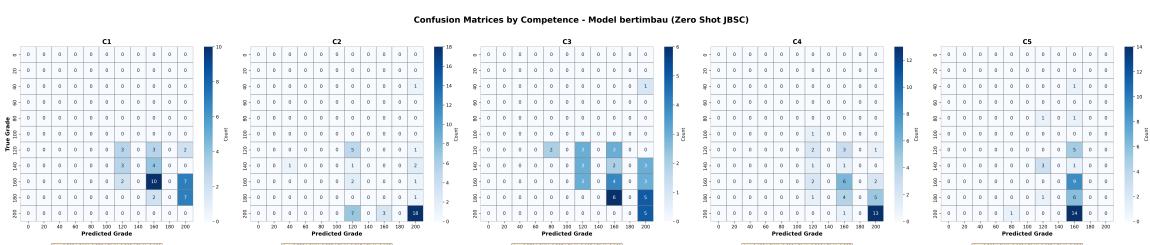


Figura A.7: Matrizes de confusão: BERTimbau – Zero-shot (C1–C5)

Pretrained Fine-tuning (JBSC)

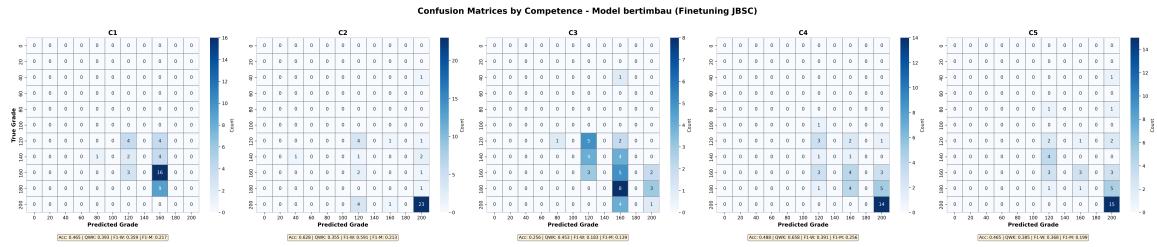


Figura A.8: Matrizes de confusão: BERTimbau – Fine-tuning após pré-treinamento em simulados (C1–C5)

Exclusive Fine-tuning

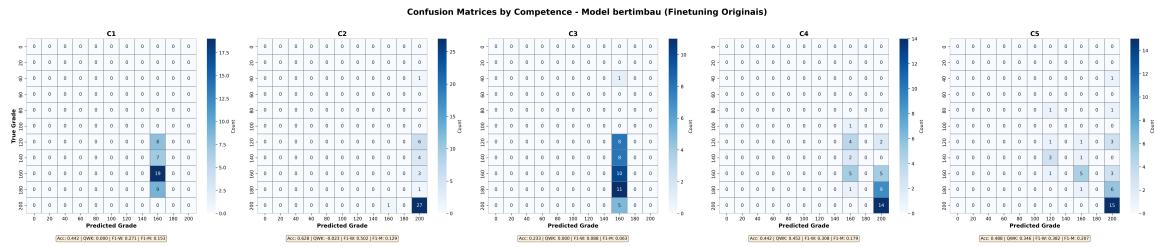


Figura A.9: Matrizes de confusão: BERTimbau – Fine-tuning Exclusivo (C1–C5)

Referências

- [AMORIM e VELOSO 2017] Evelin AMORIM e Adriano VELOSO. “A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese”. In: *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, abr. de 2017, pp. 94–102. URL: <https://aclanthology.org/E17-4010> (citado nas pgs. 1, 2, 11).
- [ATTALI e BURSTEIN 2006] Yigal ATTALI e Jill BURSTEIN. “Automated essay scoring with e-rater® v. 2”. *The Journal of Technology, Learning and Assessment* 4.3 (2006) (citado na pg. 1).
- [BARBOSA *et al.* 2025] André BARBOSA, Igor Cataneo SILVEIRA e Denis Deratani MAUÁ. “An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese”. *Journal of the Brazilian Computer Society* 31.1 (out. de 2025), pp. 858–871. DOI: [10.5753/jbcs.2025.5817](https://doi.org/10.5753/jbcs.2025.5817). URL: <https://journals-sol.sbc.org.br/index.php/jbcs/article/view/5817> (citado nas pgs. 3, 7, 12, 22, 25, 28).
- [BEIGMAN KLEBANOV e MADNANI 2020] Beata BEIGMAN KLEBANOV e Nitin MADNANI. “Automated evaluation of writing – 50 years and counting”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. por Dan JURAFSKY, Joyce CHAI, Natalie SCHLUTER e Joel TETREAULT. Online: Association for Computational Linguistics, jul. de 2020, pp. 7796–7810. DOI: [10.18653/v1/2020.acl-main.697](https://doi.org/10.18653/v1/2020.acl-main.697). URL: <https://aclanthology.org/2020.acl-main.697/> (citado na pg. 1).
- [CAVALCANTI *et al.* 2025] Rodrigo CAVALCANTI *et al.* “Diplomatrix-br: um corpus paralelo de redações de autoria humana e de llms no concurso de diplomacia brasileira”. In: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. SBC. 2025, pp. 192–205 (citado nas pgs. 11, 13).
- [DEVLIN *et al.* 2019] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE e Kristina TOUTANOVA. “Bert: pre-training of deep bidirectional transformers for language understanding”. In: *North American Chapter of the Association for Computational Linguistics*. 2019 (citado nas pgs. 3, 7, 22).

- [DOEWES *et al.* 2023] Afrizal DOEWES, Nughthoh Arfawi KURDHI e Akrati SAXENA. “Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring”. In: *Proceedings of the 16th International Conference on Educational Data Mining*. Ed. por Mingyu FENG, Tanja KÄRSER e Partha TALUKDAR. Bengaluru, India: International Educational Data Mining Society, jul. de 2023, pp. 103–113. ISBN: 978-1-7336736-4-8. doi: [10.5281/zenodo.8115784](https://doi.org/10.5281/zenodo.8115784) (citado nas pgs. 8, 22, 25).
- [FONSECA *et al.* 2018] Erick Rocha FONSECA, Ivo MEDEIROS, Dayse KAMIKAWACHI e Alessandro BOKAN. “Automatically grading brazilian student essays”. In: *Proceedings of International Conference on Computational Processing of the Portuguese Language*. 2018, pp. 170–179 (citado na pg. 22).
- [ISHIOKA e KAMEDA 2006] Tsunenori ISHIOKA e Masayuki KAMEDA. “Automated japa-nese essay scoring system based on articles written by experts”. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 2006, pp. 233–240 (citado na pg. 1).
- [LEAL *et al.* 2024] Sidney Evaldo LEAL, Magali Sanches DURAN, Carolina Evaristo SCARTON, Nathan Siegle HARTMANN e Sandra Maria ALUÍSIO. “NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese”. *Language Resources and Evaluation* 58.1 (2024), pp. 73–110. url: <https://doi.org/10.1007/s10579-023-09693-w> (citado nas pgs. 3, 12).
- [LEMAIRE e DESSUS 2001] Benoit LEMAIRE e Philippe DESSUS. “A system to assess the semantic content of student essays”. *Journal of Educational Computing Research* 24.3 (2001), pp. 305–320 (citado na pg. 1).
- [MARINHO, ANCHIÉTA *et al.* 2021] Jeziel MARINHO, Rafael ANCHIÉTA e Raimundo MOURA. “Essay-br: a brazilian corpus of essays”. In: *Anais do III Dataset Showcase Workshop*. Sociedade Brasileira de Computação, 2021, pp. 53–64. doi: [10.5753/dsw.2021.17414](https://doi.org/10.5753/dsw.2021.17414) (citado nas pgs. 1, 2, 11).
- [MARINHO, ANCHIÉTA *et al.* 2022] Jeziel MARINHO, Rafael ANCHIÉTA e Raimundo MOURA. “Essay-br: a brazilian corpus to automatic essay scoring task”. *Journal of Information and Data Management* 13.1 (2022), pp. 65–76. doi: [10.5753/jidm.2022.2340](https://doi.org/10.5753/jidm.2022.2340). url: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/2340> (citado na pg. 22).
- [MARINHO, CORDEIRO *et al.* 2022] Jeziel MARINHO, Fábio CORDEIRO, Rafael ANCHIÉTA e Raimundo MOURA. “Automated essay scoring: an approach based on enem competencies”. In: *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*. 2022, pp. 49–60 (citado na pg. 22).

REFERÊNCIAS

- [MAZZA ZAGO e AGNOLETTI DOS SANTOS PEDOTTI 2024] Ricardo MAZZA ZAGO e Luciane AGNOLETTI DOS SANTOS PEDOTTI. “Bertugues: a novel bert transformer model pre-trained for brazilian portuguese”. *Semina: Ciências Exatas e Tecnológicas* 45 (dez. de 2024), e50630. doi: [10.5433/1679-0375.2024.v45.50630](https://doi.org/10.5433/1679-0375.2024.v45.50630). URL: <https://ojs.uel.br/revistas/uel/index.php/semexatas/article/view/50630> (citado nas pgs. 3, 22).
- [MELLO *et al.* 2024] Rafael Ferreira MELLO *et al.* “PROPOR’24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays”. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*-Vol. 2. Ed. por Pablo GAMALLO *et al.* 2024, pp. 1–5. URL: <https://aclanthology.org/2024.propor-2.1/> (citado nas pgs. 11, 22).
- [OLIVEIRA *et al.* 2025] Hilário OLIVEIRA *et al.* “A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in brazilian portuguese”. *Data in Brief* 60 (2025), p. 111526. ISSN: 2352-3409. doi: [10.1016/j.dib.2025.111526](https://doi.org/10.1016/j.dib.2025.111526). URL: <https://www.sciencedirect.com/science/article/pii/S2352340925002586> (citado na pg. 11).
- [OPENAI *et al.* 2024] OPENAI *et al.* *OpenAI o1 System Card*. 2024. arXiv: [2412.16720 \[cs.AI\]](https://arxiv.org/abs/2412.16720). URL: <https://arxiv.org/abs/2412.16720> (citado na pg. 7).
- [PAGE 1966] Ellis B. PAGE. “The imminence of... grading essays by computer”. *The Phi Delta Kappan* (1966), pp. 238–243. ISSN: 00317217. URL: <http://www.jstor.org/stable/20371545> (acesso em 15/12/2022) (citado na pg. 1).
- [RADFORD *et al.* 2018] Alec RADFORD, Karthik NARASIMHAN, Tim SALIMANS, Ilya SUTSKEVER *et al.* “Improving language understanding by generative pre-training” (2018) (citado na pg. 7).
- [SILVEIRA, BARBOSA, COSTA *et al.* 2025] Igor Cataneo SILVEIRA, André BARBOSA, Daniel Silva Lopes da COSTA e Denis Deratani MAUÁ. “Investigating Universal Adversarial Attacks Against Transformers-Based Automatic Essay Scoring Systems”. In: *Intelligent Systems*. Ed. por Aline PAES e Filipe A. N. VERRI. Cham: Springer Nature Switzerland, 2025, pp. 169–183. ISBN: 978-3-031-79032-4. URL: <https://sol.sbc.org.br/index.php/bracis/article/view/33592> (citado nas pgs. 3, 13, 27).
- [SILVEIRA, BARBOSA e MAUÁ 2024] Igor Cataneo SILVEIRA, André BARBOSA e Denis Deratani MAUÁ. “A new benchmark for automatic essay scoring in Portuguese”. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. 2024, pp. 228–237 (citado nas pgs. 1, 2, 7, 11, 17).
- [SONG *et al.* 2016] Wei SONG *et al.* “Learning to identify sentence parallelism in student essays”. In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*. 2016, pp. 794–803 (citado na pg. 1).

- [SOUZA *et al.* 2020] Fábio SOUZA, Rodrigo NOGUEIRA e Roberto LOTUFO. “BERTimbau: pretrained BERT models for Brazilian Portuguese”. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020 (citado nas pgs. 3, 22).
- [VASWANI *et al.* 2017] Ashish VASWANI *et al.* “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017 (citado na pg. 5).