

Avaliação Automatizada de Coerência e Coesão Textual com Modelos Transformers

Ryan Camargo Luna¹, Lovania Roehrig Teixeira¹, Caroline Vieira Neves¹,
Eliane de Bortoli Fávero¹, Dalcimar Casanova¹

¹ Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco – Pato Branco – PR – Brasil

ryancamaargo@hotmail.com

lovaniateixeira@gmail.com, carolinevieiraneves1994@gmail.com

elianedb@utfpr.edu.br, dalcimar@gmail.com

Abstract. *This study investigates the use of Artificial Intelligence to automatically evaluate coherence (Competency 3) and cohesion (Competency 4) in argumentative essays. Using the Essay-BR Corpus (ENEM-style texts), we applied fine-tuning to the BERTimbau Transformer model to predict isolated scores for these competencies. The agreement with human raters, evaluated using the Quadratic Weighted Kappa (QWK), showed competitive results, particularly for cohesion (QWK=0.7856). The research suggests the practical potential of this model as a support tool for text correction, aiming to assist teaching activities and providing insights for the advancement of Natural Language Processing applied to Brazilian education.*

Keywords: *Natural Language Processing; Automated Essay Scoring; Transformers; Coherence and Cohesion; Informatics in Education.*

Resumo. *Este estudo investiga o uso de Inteligência Artificial para avaliar automaticamente a coerência (Competência 3) e coesão (Competência 4) em redações dissertativo-argumentativas. Utilizando o Essay-BR Corpus (textos padrão ENEM), aplicou-se fine-tuning no modelo Transformer BERTimbau para prever as notas destas competências de forma isolada. A concordância com avaliadores humanos, aferida pelo Kappa Quadrático Ponderado (QWK), revelou resultados competitivos, destacando-se a avaliação da coesão (QWK=0,7856). Essa análise sugere o potencial prático do modelo como ferramenta de apoio à atividade de correção de textos, visando auxiliar o trabalho docente e oferecendo reflexões para o avanço do Processamento de Linguagem Natural aplicado à Informática na Educação brasileira.*

Palavras-chave: *Processamento de Linguagem Natural; Correção Automática; Transformers; Coerência e Coesão; Informática na Educação.*

1. Introdução

O ensino-aprendizagem de produções textuais e a sua correção estão entre as atividades que mais demandam tempo e esforço de um professor de Língua Portuguesa. Nessas etapas, o docente assume papéis complementares: orienta o processo de escrita e atua como avaliador para diagnosticar deficiências. Avaliar a produção textual requer analisar

aspectos gramaticais (ortografia, concordância) e estruturais (coerência, coesão, desenvolvimento do tema) [Conceição 2004]. O uso de instrumentos automatizados de auxílio apresenta-se como uma alternativa para otimizar esse processo, permitindo aos docentes dedicar maior atenção à orientação pedagógica individualizada do aluno.

Com o avanço da Inteligência Artificial (IA), sistemas de *Automated Essay Scoring* (AES) têm auxiliado na correção textual. Ferramentas comerciais clássicas [CHUNG et al. 2009] focavam primordialmente na ortografia e, gradativamente, passaram a identificar desvios lexicais e de concordância [NAU et al. 2017]. Trabalhos recentes [RODRIGUES 2023] apontam a importância da avaliação global em exames como o Exame Nacional do Ensino Médio (ENEM), embora frequentemente sem isolar aspectos granulares de ordem discursiva.

Dentre os critérios de avaliação textual, os aspectos estruturais de coesão e coerência destacam-se pela complexidade de automação, pois envolvem relações semânticas e dependências sequenciais de longo alcance [Marcuschi 2008]. Embora a literatura acadêmica apresente avanços na avaliação automática desses critérios [SOMASUNDARAN et al. 2014, LIN et al. 2011, Li and Jurafsky 2017], ainda existem desafios para a adoção dessas soluções em larga escala, especialmente devido às particularidades morfológicas de diferentes idiomas.

Observa-se, portanto, uma oportunidade na literatura para o desenvolvimento de modelos neurais específicos para a avaliação profunda desses aspectos em português. No contexto nacional, o ENEM avalia milhões de textos anualmente, exigindo rigor na Competência 3 (que transcende a coerência textual básica, englobando a seleção, relação, organização e interpretação de fatos e opiniões) e na Competência 4 (focada nos mecanismos linguísticos de coesão).

Diante desse cenário, a questão de pesquisa que norteia este estudo é: *Em que medida arquiteturas neurais baseadas em Transformers, pré-treinadas em idioma nativo, conseguem avaliar isoladamente as competências de coerência (C3) e coesão (C4) em redações do ENEM?*

Para responder a esta questão, a hipótese desta pesquisa é de que o modelo neural BERTimbau possui representações latentes adequadas para mapear as dependências sintáticas e os encadeamentos interfrasais do Português Brasileiro, viabilizando a predição de notas de coerência e coesão com níveis de concordância substanciais.

O objetivo deste trabalho é avaliar o desempenho preditivo do modelo BERTimbau nas competências C3 e C4 utilizando o *Essay-BR Corpus*, comparando os resultados com linhas de base clássicas. Ressalta-se que o intuito da pesquisa é realizar uma validação técnica computacional. Dessa forma, busca-se investigar como tais tecnologias podem contribuir futuramente para a Informática na Educação, servindo não como substitutas do professor, mas como instrumentos de triagem e apoio para o fornecimento de *feedback* formativo ágil.

2. Avaliação da Produção Textual

A atividade de avaliação textual é essencial para o desenvolvimento da competência linguística desde a educação básica até ao ensino superior. Trata-se de uma tarefa na qual o docente atua como mediador, propiciando condições para que o estudante uti-

lize a linguagem efetivamente em diversos contextos sociais. O ensino da produção textual é desafiador, pois demanda habilidades que vão desde a organização sequencial de ideias até ao atendimento aos objetivos comunicativos de gêneros textuais específicos [Marcuschi 2008, Conceição 2004, RODRIGUES et al. 2020].

De acordo com a literatura [SERAFINI 1989], a correção textual praticada pelos docentes apresenta três tendências principais: a *indicativa*, na qual o professor apenas aponta os erros ortográficos e de pontuação; a *resolutiva*, em que o docente reformula o texto para a norma culta; e a *classificatória*, que utiliza códigos para indicar problemas e dar autonomia de reescrita ao aluno. A estas, soma-se a correção *textual-iterativa* [RUIZ 2010], que se assemelha a um diálogo em que o docente escreve comentários longos, buscando desenvolver as habilidades de escrita dos alunos e, assim, a coerência, a coesão e a sua reflexão sobre a própria produção textual [SILVA et al. 2013].

O grande desafio do trabalho docente é, justamente, despertar as habilidades que permitam aos alunos redigir textos não apenas adequados à estrutura do gênero, mas que apresentem um discurso coerente e coeso, significativo à sua finalidade social [MENEGASSI and GASPAROTTO 2016]. Um texto bem escrito transmite os seus sentidos de forma clara, apresenta estrutura lógica e coesão entre as ideias, facilitando a compreensão e o engajamento do leitor [Marcuschi 2008].

Num viés social, a proficiência na escrita e a capacidade de veicular intenções com clareza auxiliam no desenvolvimento humano e educacional [FREIRE 1979]. Por serem indispensáveis, competências como a coesão e a coerência são rigorosamente testadas em exames de larga escala, como o ENEM, cujas particularidades avaliativas serão abordadas na seção seguinte.

2.1. Coerência e Coesão em Redações do ENEM

A avaliação de textos no ENEM baseia-se em cinco competências (0 a 200 pontos cada, totalizando 1.000 pontos) [INEP 2020]. Destacam-se para este estudo a **Competência 3 (C3)**, que avalia a seleção e organização lógica dos argumentos (coerência), e a **Competência 4 (C4)**, focada nos mecanismos linguísticos e estruturais (coesão). As demais (C1, C2 e C5) avaliam, respectivamente, o domínio da norma-padrão, a compreensão do tema e a proposta de intervenção.

Apesar de haver 5 competências, o presente estudo concentra-se na análise e na qualificação de modelos de IA para a avaliação da Coerência e da Coesão, o que justifica o foco exclusivo nas competências C3 e C4. A avaliação textual nestas rubricas fundamenta-se fortemente na capacidade do candidato de estruturar as suas ideias de forma lógica e articulada. Embora a coerência e a coesão sejam interdependentes na construção do sentido, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) avalia esses fenômenos de forma segmentada.

A Coerência textual, avaliada primordialmente pela C3, diz respeito ao nível semântico e cognitivo de desenvolvimento e articulação das ideias e conceitos do texto. Trata-se da logicidade das ideias, da ausência de contradições e da progressão temática. Na C3, o candidato é julgado pela sua habilidade de “selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista” [INEP 2020]. Um texto coerente é aquele que faz sentido para o leitor e que sustenta a sua tese do início ao fim, sem falhas argumentativas. Sob a ótica do Processamento de

Linguagem Natural (PLN), avaliar a C3 é um desafio complexo, pois exige que o modelo computacional, como o BERT, transcenda a análise de palavras isoladas e consiga capturar o significado global e a validade lógica dos argumentos apresentados ao longo dos parágrafos. É importante ressaltar que a Competência 3 do ENEM, embora centrada na coerência, exige do aluno a capacidade de selecionar e interpretar fatos e opiniões, configurando-se como uma habilidade de construção argumentativa mais ampla do que a mera coesão sintático-semântica.

Por outro lado, a Coesão textual, foco central da C4, refere-se à superfície do texto, ou seja, aos mecanismos linguísticos e gramaticais que conectam as palavras, orações e parágrafos. A C4 exige que o candidato demonstre “conhecimento dos mecanismos linguísticos necessários para a construção da argumentação” [INEP 2020]. Isso inclui o uso adequado de conjunções, pronomes, preposições e sinônimos para evitar repetições e garantir a fluidez da leitura. Enquanto a coerência (C3) é o esqueleto lógico da redação, a coesão (C4) funciona como os instrumentos de articulação que unem esse esqueleto.

2.2. Avaliação Automática de Redações

Embora a literatura apresente estudos voltados à análise de coerência em tarefas sintéticas, nota-se uma carência de trabalhos fundamentados em modelos para a avaliação de coerência e de coesão em redações escolares reais. Por serem critérios dotados de alta subjetividade e de complexidade estrutural, a aplicação de ferramentas de IA nesta área configura-se como uma abordagem inovadora, que pode auxiliar o docente em suas tarefas e no ensino-aprendizagem de produção textual, elemento fundamental para o desenvolvimento de cidadãos competentes no uso da linguagem [LINSINGEN 2017].

Historicamente, os primeiros trabalhos de avaliação automática focados nesses critérios utilizavam tarefas sintéticas, como a reordenação de frases ou a inserção de sentenças alheias para distinguir textos originais de exemplos modificados [LAPATA 2003, BARZILAY and LAPATA 2005, ELSNER and CHARNIAK 2011, Li and Jurafsky 2017]. Embora fundamentais para o avanço da área, tais abordagens baseadas em comparações binárias não atribuíam pontuações diretas a produções textuais reais. Com a evolução do campo, tornou-se necessária a adoção de bases de dados com anotação manual para o treinamento de modelos capazes de inferir notas em escalas complexas, aproximando-se do processo de avaliação humana [SOMASUNDARAN et al. 2014, BURSTEIN et al. 2010].

No cenário internacional recente, as pesquisas avançaram para a atribuição direta de pontuações de coerência e coesão utilizando bases de dados anotadas manualmente. O trabalho de Lai e Tetreault [LAI and TETREAULT 2018], por exemplo, estabeleceu um marco ao avaliar a coerência em textos de escritores não profissionais em inglês. Contudo, a transposição direta desses modelos esbarra em complexidades morfológicas e sintáticas próprias de cada idioma. No contexto brasileiro, a literatura voltada ao ENEM frequentemente é focada na nota global ou em critérios isolados, como a classificação de fuga ao tema [Pinho et al. 2022]. Em uma revisão recente, Mello et al. [MELLO et al. 2023] avaliaram a coerência temática em português, porém limitando-se a modelos clássicos de aprendizado de máquina.

O campo de avaliação automática de redações (AES) tem passado por uma transição paradigmática recente. Enquanto estudos fundamentais estabeleceram as bases da atenção via *Transformers* [Vaswani et al. 2017], a literatura contemporânea tem

migrado para o uso intensivo de Grandes Modelos de Linguagem (*Large Language Models* - LLMs). Revisões sistemáticas recentes [Gao and Wang 2024] indicam que, embora LLMs apresentem capacidades impressionantes de generalização, o ajuste fino (*fine-tuning*) em *corpora* específicos e de domínio nativo — como o caso do BERTimbau aplicado ao *Essay-BR* — continua sendo uma estratégia robusta para atingir alta concordância métrica em rubricas avaliativas complexas, como a coerência e a coesão.

3. Materiais e Métodos

3.1. Base de Dados e Critérios de Exclusão

Para o desenvolvimento deste estudo, utilizou-se o *Essay-BR Corpus* em sua versão estendida [Marinho et al. 2022]. Este é atualmente o maior conjunto de redações do ENEM disponível para pesquisa acadêmica em Língua Portuguesa, coletadas entre dezembro de 2015 e abril de 2020.

Para garantir o rigor metodológico, foram aplicados critérios de exclusão que descartaram textos com menos de 100 palavras, presença de URLs ou elementos atípicos que inviabilizariam a análise discursiva. Após a limpeza, o *corpus* resultou em 6.577 produções dissertativo-argumentativas. A Tabela 1 consolida as métricas quantitativas e a distribuição de frequência absoluta das notas para as competências C3 e C4.

Tabela 1. Análise Quantitativa e Distribuição de Frequência do Essay-BR Corpus

Métrica Estatística	Competência 3 (C3)	Competência 4 (C4)
Média $\pm$ Desvio Padrão (Nota)	115, 56 $\pm$ 38, 17	134, 99 $\pm$ 45, 38
Média $\pm$ Desvio Padrão (Tokens)	390, 27 $\pm$ 114, 76	390, 27 $\pm$ 114, 76
Frequência por Nota	Qtd. de Textos	Qtd. de Textos
Nota 0	185	207
Nota 40	164	65
Nota 80	1.606	885
Nota 120	3.054	2.459
Nota 160	1.377	1.822
Nota 200	191	1.139
Total de Textos	6.577	6.577

A análise revela um profundo desbalanceamento, com a vasta maioria dos textos concentrada nas faixas intermediárias (120 e 160 pontos). Este cenário é um parâmetro crítico para o *fine-tuning*, pois a escassez de exemplos nas extremidades da escala impõe um viés analítico ao algoritmo, dificultando a penalização rigorosa de redações deficitárias. Para ilustrar o formato estrutural dos dados brutos que alimentarão o algoritmo, a Tabela 2 apresenta uma amostra representativa do *corpus*. Cada instância contém um identificador único, o corpo do texto e a anotação detalhada das competências avaliadas pelo exame nacional (cujas definições teóricas de C3 e C4 foram detalhadas na Seção 2.1), assumindo valores discretos em múltiplos de 40 pontos.

3.2. Pré-processamento e Preparação dos Textos

Diferentemente de abordagens tradicionais de Processamento de Linguagem Natural, que aplicam etapas extensivas de limpeza (como remoção de *stopwords* ou *stemming*), adotou-

Tabela 2. Amostra Representativa da Base de Dados

ID	Título	Redação (Trecho Inicial)	C3	C4	Nota Final
56	[Sem Título]	É explícito que o sistema prisional...	120	120	520
57	[Sem Título]	Ao fazer uma análise do cenário atual...	80	120	520
56	Justiça injusta...	A realidade dos presídios é bem mais...	80	120	520
60	Perplexidade...	A reforma da previdência social...	80	120	520
62	Responsabilidade...	Ao se falar sobre o discurso de ódio...	80	120	520

se um pré-processamento estritamente mínimo para preservar a estrutura linguística. Elementos como conjunções e pontuação são cruciais para as competências C3 e C4. A segmentação foi delegada nativamente ao algoritmo *WordPiece* do modelo BERT, convertendo o texto em representações vetoriais densas que preservam o contexto semântico original.

A segmentação foi delegada nativamente ao tokenizador *WordPiece* do modelo BERT, convertendo o texto em representações vetoriais densas que preservam o contexto semântico original, respeitando o limite arquitetural da janela de contexto de 512 tokens.

Cabe ressaltar que, diferentemente de tarefas tradicionais de classificação de textos, optou-se por não empregar técnicas de aumento de dados (*Data Augmentation*) para mitigar o desbalanceamento das classes. Em tarefas de AES focadas em coerência e coesão, qualquer alteração sintética no texto original — como a substituição por sinônimos, inserção ou supressão de frases — modifica fundamentalmente a estrutura discursiva, gerando um novo texto cuja nota original perderia a validade. Portanto, a integridade da rubrica exigiu o treinamento estritamente com os dados reais.

Uma análise do *corpus* revelou que a extensão média dos textos é de 390,27 tokens ($\pm 114,76$). Como a arquitetura possui uma janela de contexto limite de 512 tokens, aplicou-se o truncamento (*truncation*) para as instâncias excedentes. Dado que a média se posiciona abaixo deste limite, o truncamento afetou apenas *outliers* marginais, preservando a integridade do fluxo argumentativo na maioria das produções.

3.3. Treinamento e Seleção do Modelo

A seleção e treinamento do modelo constitui etapa crítica para garantir que a arquitetura neural seja capaz de capturar as sutilezas linguísticas e as relações de dependência presentes nas redações do ENEM.

Optou-se pela utilização do **BERTimbau**, uma versão da arquitetura BERT pré-treinada exclusivamente com um vasto *corpus* do Português Brasileiro. A escolha justifica-se por seu desempenho competitivo observado em relação à literatura em tarefas de compreensão de linguagem natural no idioma nativo, superando modelos multilíngues genéricos. Foram selecionadas para experimentação as variantes *Base* (110 milhões de parâmetros) e *Large* (335 milhões de parâmetros), visando identificar o equilíbrio ideal entre a capacidade representacional e o risco de *overfitting*.

3.4. Pipeline Experimental e *Fine-Tuning*

Para garantir a rigorosa reprodutibilidade do experimento, o *pipeline* foi fixado com uma semente aleatória (*random seed* = 42). Os dados foram divididos de forma estratificada

em conjuntos de treino (70%, 4.604 textos), validação (15%, 987 textos) e teste (15%, 986 textos inéditos). O fluxo experimental foi estruturado em duas abordagens:

- **Modelos Clássicos (*Baselines*):** Implementação de algoritmos tradicionais de Aprendizado de Máquina — Regressão Linear, Random Forest e SVM — baseados em características textuais extraídas, estabelecendo as linhas de base para comparação.
- **Ajuste Fino do BERTimbau:** Realização do *fine-tuning* completo nas arquiteturas BERTimbau Base e Large, em que as camadas da rede tiveram seus pesos recalibrados para especializar o modelo na tarefa de classificação das notas.

3.4.1. Otimização de Hiperparâmetros

A otimização dos hiperparâmetros na fase de *fine-tuning* foi conduzida pela biblioteca *Optuna* com otimização Bayesiana, tendo como função objetivo a maximização do QWK no conjunto de validação. O processo de otimização focou na definição de parâmetros críticos para a convergência e estabilidade da rede neural, tais como a taxa de aprendizado (*learning rate*), o tamanho do lote (*batch size*), o decaimento de peso (*weight decay*) e o número de épocas de treinamento. A função “objetivo” do algoritmo de busca foi configurada para maximizar estritamente o Kappa Quadrático Ponderado (QWK) obtido na base de validação, garantindo que o modelo fosse penalizado adequadamente por erros de predição distantes do gabarito.

Adicionalmente, a utilização desta ferramenta viabilizou a implementação de mecanismos de poda (*pruning*), que interrompem precocemente as rodadas de treinamento (*trials*) cujo desempenho inicial se mostre pouco promissor. A aplicação combinada dessas técnicas reduziu significativamente o custo computacional, mitigou o risco de *overfitting* e eliminou a dependência de escolhas empíricas manuais, assegurando a robustez experimental e a reprodutibilidade dos modelos finais gerados para as competências C3 e C4.

3.5. Métricas de Avaliação

Para mensurar o desempenho preditivo e o grau de concordância entre as predições do sistema e o gabarito humano, a avaliação foi fundamentada no QWK - κ . Esta métrica é o padrão-ouro em tarefas de AES, pois penaliza erros de forma não linear de acordo com a distância entre as classes ordinais. O cálculo do coeficiente κ e sua matriz de penalidade quadrática são definidos pela equação:

$$\kappa = 1 - \frac{\sum_{i,j=1}^N w_{i,j} O_{i,j}}{\sum_{i,j=1}^N w_{i,j} E_{i,j}}, \quad \text{onde} \quad w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (1)$$

Nesta formulação, $O_{i,j}$ representa os elementos da matriz de confusão observada (frequência de predição i para a nota real j) e $E_{i,j}$ representa os elementos da matriz de contagens esperadas sob a hipótese de independência estatística. O peso $w_{i,j}$ aplica a penalidade quadrática baseada na distância entre a nota atribuída pelo humano (i) e a nota inferida pelo modelo (j), sendo $N = 6$ o número total de classes de pontuação do ENEM. Complementarmente, foram utilizados o MAE, Acurácia e o F1-Score para detalhar a performance global e a capacidade discriminativa do modelo.

4. Resultados e Discussão

4.1. Avaliação na Base de Teste com Modelos Clássicos

Antes da aplicação de arquiteturas de aprendizagem profunda baseadas no *Transformer*, estabeleceu-se uma linha de base (*baseline*) experimental rigorosa. Para tanto, a tarefa de avaliação das redações foi modelada empregando-se algoritmos clássicos de *Machine Learning* associados à extração de características tradicionais, combinando TF-IDF (*Term Frequency-Inverse Document Frequency*) com métricas linguísticas de superfície.

Essa etapa teve como objetivo verificar até que ponto os métodos clássicos de representação textual conseguem mapear a complexidade estrutural e semântica das competências de Coerência (C3) e Coesão (C4) do ENEM, servindo de termo de comparação para os modelos contextuais. A Tabela 3 apresenta o desempenho preditivo dos modelos de Regressão Linear, *Random Forest* e *Support Vector Machine* (SVM). Para garantir uma comparação justa com o estado da arte, as previsões contínuas foram discretizadas para os níveis da rubrica oficial (0, 40, 80, 120, 160 e 200), sendo avaliadas pelo Erro Absoluto Médio (MAE) e por métricas de classificação multiclasse e concordância.

Tabela 3. Comparação de desempenho dos Modelos Clássicos na base de teste.

Competência	Modelo Clássico	MAE	F1-Score	QWK (κ)	Acurácia
C3 - Coerência	Regressão Linear	89,61	0,1857	0,0720	14,59 %
C3 - Coerência	Random Forest	23,34	0,5106	0,4454	54,91 %
C3 - Coerência	SVM	24,85	0,3731	0,2017	49,14 %
C4 - Coesão	Regressão Linear	94,19	0,1981	0,1380	19,35 %
C4 - Coesão	Random Forest	27,50	0,4098	0,4873	45,09 %
C4 - Coesão	SVM	31,17	0,3318	0,3242	40,12 %

A análise comparativa revela que a avaliação isolada das competências impõe desafios metodológicos significativos. O modelo de Regressão Linear demonstrou dificuldade acentuada de mapear a representação vetorial (TF-IDF) para as notas, resultando num Erro Absoluto Médio (MAE) extremo, errando as avaliações em mais de duas faixas inteiras de nota (aprox. 90 pontos de desvio) e apresentando níveis de concordância quase nulos (*Kappa Quadrático* inferior a 0,15). O modelo SVM obteve um desempenho intermediário, superando a abordagem linear, mas esbarrando em limitações de generalização.

Em contrapartida, o algoritmo *Random Forest* obteve a melhor performance entre os métodos tradicionais, estabelecendo o teto de desempenho para esta abordagem. Na avaliação da Coerência (C3), apresentou o menor erro (MAE = 23,34) e a maior taxa de concordância (*Kappa Quadrático* = 0,4454). Na Coesão (C4), atingiu o seu melhor nível de concordância (*Kappa Quadrático* = 0,4873). No contexto da escala de notas do exame, um MAE situado entre 23 e 27 pontos indica que o *Random Forest* consegue estimar a nota do candidato com um desvio geralmente inferior a um nível completo da rubrica.

Apesar de o *Random Forest* apresentar resultados razoáveis sob a ótica de um modelo clássico, a abordagem tradicional demonstra limitações fundamentais insuperáveis. Modelos baseados em contagem de frequência de palavras (*bag-of-words*, TF-IDF) perdem a informação sequencial do texto e ignoram o contexto em que as palavras estão inseridas. Consequentemente, falham em capturar as nuances semânticas complexas, o fluxo

de progressão temática e as dependências de longo alcance estritamente exigidas pelas rubricas que englobam Coerência e Coesão do ENEM. Esta restrição metodológica sugere que a extrema complexidade da tarefa e justifica a necessidade essencial de transição para modelos neurais contextuais mais avançados.

4.2. Desempenho do Modelo Ajustado (Fine-Tuning) e Análise de Erros

Para superar as limitações da abordagem clássica, procedeu-se com o *fine-tuning* supervisionado. A fase de ajuste fino permitiu que as camadas de atenção do *Transformer* fossem calibradas para as matrizes de referência específicas do exame. A configuração ótima de hiperparâmetros, selecionada de forma autônoma pelo algoritmo TPE (*Optuna*) para a consolidação do melhor modelo, consistiu em: taxa de aprendizado (*learning rate*) de $3,39 \times 10^{-5}$, tamanho do lote (*batch size*) de 16, decaimento de peso (*weight decay*) de 0,2319 e treinamento ao longo de 3 épocas.

Durante a modelagem, comparou-se a eficácia das variantes BERTimbau *Base* (110 milhões de parâmetros) e *Large* (335 milhões de parâmetros), padronizando a avaliação através de quatro métricas centrais: o Erro Absoluto Médio (MAE) para mensurar o desvio contínuo, e o F1-Score, Acurácia e *Kappa Quadrático* (QWK) para avaliar o acerto das categorias da rubrica. Os resultados detalhados encontram-se na Tabela 4.

Tabela 4. Comparação de desempenho após *Fine-Tuning*: BERTimbau *Base* vs. *Large*

Competência	Arquitetura Neural	MAE	F1-Score	QWK (κ)	Acurácia
C3 - Coerência	BERT Base Fine-Tuned	12,96	0,6477	0,6397	67,58%
C3 - Coerência	BERT Large Fine-Tuned	19,60	0,5441	0,5418	56,61%
C4 - Coesão	BERT Base Fine-Tuned	10,97	0,7113	0,7856	72,57%
C4 - Coesão	BERT Large Fine-Tuned	34,50	0,2071	0,0000	37,77%

A análise empírica revela que a versão *Large* demonstrou uma limitação preditiva severa na competência C4. Esse comportamento é evidenciado pela drástica queda das métricas de classificação, resultando em um QWK nulo e em um Erro Absoluto Médio (MAE) elevado de 34,50 pontos, indicando que o modelo convergiu para a previsão de uma nota constante para todos os casos. Esse comportamento pode ser associado à instabilidade de gradiente e ao *overfitting*, visto que arquiteturas massivas aplicadas a conjuntos de dados com desbalanceamento de classes tendem a apresentar desafios de generalização.

Em contrapartida, a versão *Base* demonstrou uma capacidade robusta de aprendizado, alcançando níveis de eficiência na competência C4 com um QWK de 0,7856 e uma precisão de 72,57%. Esse resultado sugere que, no contexto de avaliação automática de redações em português analisado, um modelo mais leve e bem calibrado pode atingir desempenhos superiores a arquiteturas mais paramétricas.

Cabe destacar que a concordância entre avaliadores humanos especialistas em tarefas de correção de redações em larga escala geralmente situa-se em um QWK entre 0,60 e 0,80 [Shermis 2014]. O índice de 0,7856 alcançado pelo BERTimbau Base na avaliação da coesão caracteriza uma *concordância substancial* na literatura da área, aproximando-se das métricas de convergência esperadas entre avaliadores humanos.

Para uma averiguação minuciosa do comportamento preditivo, geraram-se as matrizes de confusão normalizadas (Figura 1), permitindo a observação das taxas de acerto relativas (*Recall*) por classe, independentemente do desbalanceamento volumétrico da base de dados. A análise visual confirma que os modelos concentram a maior parte de seus acertos na diagonal principal nas faixas de notas altas (120 a 200). O Modelo C4 (Figura 1b) demonstra menor dispersão para classes vizinhas, justificando suas métricas superiores. Em contraste, o Modelo C3 (Figura 1a) apresenta maior dificuldade em estabelecer fronteiras rígidas, espalhando mais as predições.

Contudo, a normalização dos dados revela nuances críticas sobre o viés do desbalanceamento: diferentemente do que a observação estritamente absoluta poderia sugerir, a rede neural demonstra considerável eficácia na identificação da classe minoritária de nota 0 (alcançando mais de 81% de acerto na C4). O verdadeiro desafio do modelo encontra-se na nota 40. Observa-se que essa faixa apresenta as menores taxas de retenção na diagonal principal, sendo sistematicamente confundida com a nota 0 ou superestimada para a nota 80. Isso evidencia a dificuldade algorítmica em delinear a tênue fronteira linguística que separa textos completamente rudimentares ou com fuga ao tema (0) de produções muito incipientes (40).

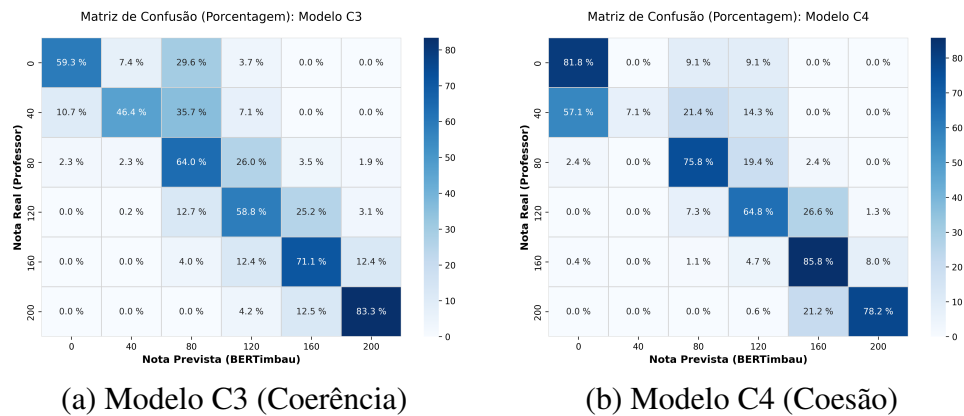


Figura 1. Matrizes de confusão comparando as notas reais e previstas para C3 (a) e C4 (b).

4.3. Comparativo com o Estado da Arte

Para situar a contribuição deste trabalho frente ao estado da arte nacional e internacional, os resultados foram contrastados com pesquisas consolidadas em AES, conforme a Tabela 5.

Tabela 5. Comparação de resultados entre os Modelos Propostos e a literatura

Modelo / Trabalho	Base de Dados	QWK (κ)	Acurácia	MAE	RMSE	Foco / Idioma
Modelo C4 (Este Trab.)	Essay-BR	0,7856	72,57%	11,55	23,99	Coesão (PT-BR)
Modelo C3 (Este Trab.)	Essay-BR	0,6397	67,58%	17,31	30,16	Coerência (PT-BR)
Mello et al. [MELLO et al. 2023]	Essay-BR	-	-	0,52*	0,70*	Coerência (PT-BR)
Lai e Tetreault [LAI and TETREAULT 2018]	Yahoo/Yelp	0,38 a 0,55	65,50%	-	-	Coerência (EN)

* Nota: Os valores de MAE de Mello et al. (2023) referem-se a uma escala de 1 a 5, portanto não são diretamente comparáveis à escala original do ENEM (0-200) utilizada neste estudo.

Ressalta-se que a comparação direta das métricas de erro (MAE e RMSE) entre os presentes modelos e o estudo de Mello et al. [MELLO et al. 2023] é inviabilizada pela diferença de escalas. Enquanto esta pesquisa calcula os desvios na escala original

do exame (0 a 200 pontos), o trabalho anterior utiliza notas normalizadas (1 a 5). Dessa forma, o *MAE* de 0,52 reportado pelos autores reflete apenas uma proporção de sua escala reduzida, não representando uma vantagem preditiva em relação ao BERTimbau.

Ainda assim, a análise do estado da arte evidencia que o *fine-tuning* de grandes modelos de linguagem (*Transformers*) captura as complexidades estruturais do texto com maior profundidade que os algoritmos clássicos de regressão tradicionalmente explorados na literatura nacional [MELLO et al. 2023]. Ademais, mesmo em comparação com estudos consolidados em Língua Inglesa [LAI and TETREAULT 2018], os modelos aqui propostos apresentam alta competitividade, os resultados sugerem que a evidência do BERTimbau para a avaliação automatizada no Português Brasileiro.

4.4. Discussão e Implicações Pedagógicas

A vantagem sistemática do Modelo C4 sobre o C3 sugere que fenômenos de coesão, por serem dependentes do léxico e de conectivos explícitos, são mais facilmente mapeados pelo mecanismo de atenção do que a coerência estrutural profunda. Esta última exige inferências sobre a progressão temática e ausência de contradições, tarefas que desafiam as representações latentes atuais.

Embora o foco deste estudo não seja a predição da nota global, os resultados individuais para C3 e C4 possibilitam a criação de ferramentas de *feedback* granular. Diferente de sistemas que fornecem apenas um escore final, o isolamento destas competências permite identificar se as deficiências de um estudante residem na articulação gramatical (C4) ou na organização lógica das ideias (C3). Contudo, a análise diagnóstica das matrizes de confusão expôs um viés algorítmico nas notas baixas (0 e 40), decorrente do desbalanceamento do *Essay-BR Corpus*, o que demanda cautela na aplicação pedagógica para evitar “falsos positivos de proficiência”.

Nesse contexto, a integração do modelo em plataformas de Informática na Educação deve basear-se no *design Human-in-the-Loop* (HITL), atuando não como um substituto, mas como um copiloto educacional. Como instrumento de triagem formativa, o sistema automatiza a avaliação estrutural recorrente e fornece *feedback* imediato, permitindo que o aluno reflita e aplique a reescrita autônoma antes mesmo da entrega final. Isso libera o docente de Língua Portuguesa para exercer uma mediação pedagógica mais profunda, focada na criticidade e no estilo argumentativo, o que viabiliza um acompanhamento mais próximo e individualizado. Nessa dinâmica, as predições do BERTimbau servem como um rascunho analítico a ser validado pelo educador, garantindo que o encaminhamento final permaneça sob jurisdição humana e mitigando os riscos associados aos vieses algorítmicos. Dessa forma, a tecnologia permite escalar o apoio à escrita sem comprometer a qualidade essencial da interação humana.

4.5. Ameaças à Validade e Cuidados Éticos na Educação

Embora os resultados técnicos sejam competitivos, a transposição de modelos neurais para o ambiente educacional exige cautela metodológica e ética. A primeira limitação refere-se ao viés do desbalanceamento de classes do *corpus*. Como os modelos neurais tendem a priorizar as classes majoritárias para minimizar o erro global, observou-se uma dificuldade algorítmica em prever notas extremas inferiores (0 e 40). No contexto pedagógico, isso representa um risco de “falsos positivos de proficiência”, onde alunos

com graves deficiências estruturais poderiam receber avaliações medianas, mascarando a necessidade de intervenção docente. Métricas como o F1-Score (variante macro) foram utilizadas justamente para não mascarar essa deficiência sobre as classes minoritárias.

Outra limitação diz respeito à variância estatística dos resultados experimentais. Devido aos elevados custos computacionais associados ao treinamento de arquiteturas *Transformer*, os experimentos foram conduzidos em uma única divisão de treino, validação e teste, utilizando uma semente aleatória fixa. Trabalhos futuros devem empregar estratégias como a validação cruzada para garantir uma avaliação mais abrangente da estabilidade do modelo perante a variabilidade amostral.

Uma consideração ética relevante reside na opacidade inerente às arquiteturas *Transformer* (modelos "caixa-preta"). Sistemas de avaliação de alto impacto (*high-stakes assessment*) na educação necessitam de explicabilidade; não basta o modelo prever uma nota baixa de coesão, ele precisa indicar ao estudante quais conectivos faltaram. Por fim, ressalta-se que a validade externa do modelo está restrita à tipologia textual dissertativo-argumentativa nos moldes do ENEM, não sendo recomendada a sua generalização direta para outros gêneros textuais, faixas etárias ou contextos escolares sem prévio processo de *transfer learning*.

5. Conclusão e Trabalhos Futuros

Esta investigação demonstrou o potencial do modelo BERTimbau para a avaliação automatizada das competências estruturais nas redações do ENEM. A arquitetura Base alcançou valores de QWK de 0,7856 na Coesão (C4) e 0,6397 na Coerência (C3), superando os resultados obtidos com extratores de características estáticas associados a modelos clássicos de regressão. Esses achados reforçam a viabilidade de dar continuidade a esta linha de pesquisa e indicam um caminho promissor para o desenvolvimento de ferramentas preditivas que auxiliem o trabalho docente.

Apesar dos resultados expressivos, a análise dos erros revelou limitações importantes. O modelo apresentou menor sensibilidade nas pontuações mais baixas, problema agravado pelo desbalanceamento do conjunto de treinamento. Tal viés dificultou a identificação precisa de notas extremas (0 e 40), o que poderia gerar "falsos positivos de proficiência" em aplicações reais.

Como trabalhos futuros, propõe-se a expansão da base de dados, com atenção especial aos extremos da rubrica, combinada ao uso de técnicas de reamostragem e validação cruzada, visando maior estabilidade e generalização dos modelos. Diante do recente avanço da IA Generativa, também se faz necessária uma investigação comparativa com Grandes Modelos de Linguagem (LLMs), avaliando o *trade-off* entre o alto custo computacional dessas ferramentas e o *fine-tuning* de modelos discriminativos locais. Adicionalmente, a incorporação de métodos de *explainable AI* (XAI) configura-se como um avanço fundamental. Mais do que prever notas, essas tecnologias precisam oferecer explicações visuais, como mapas de calor (*heatmaps*) baseados nos pesos de atenção, que destaquem para o estudante exatamente onde a progressão temática foi rompida, transformando a predição algorítmica em um *feedback* verdadeiramente útil para a Informática na Educação.

Referências

- BARZILAY, R. and LAPATA, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.
- BURSTEIN, J., CHODOROW, M., and LEACOCK, C. (2010). Automated essay scoring: A review of the research and a look to the future. Technical report, Educational Testing Service (ETS) Research Report.
- CHUNG, G. Y., ELLIS, G., HUTCHINSON, B., and WHITELAW, C. (2009). Using the web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 890–899, Singapura.
- Conceição, R. I. S. (2004). Correção de texto: um desafio para o professor de português. *Trabalhos em Linguística Aplicada*, 43(2):323–344.
- ELSNER, M. and CHARNIAK, E. (2011). Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA. Association for Computational Linguistics.
- FREIRE, P. (1979). *Educação e Mudança*. Editora Paz e Terra, Rio de Janeiro, 12 edition.
- Gao, Y. and Wang, X. (2024). Automated essay scoring with large language models: A systematic review. *Educational Technology & Society*, 27(2).
- INEP (2020). *A redação no Enem 2020: cartilha do participante*. Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, DF.
- LAI, A. and TETREAULT, J. (2018). Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.
- LAPATA, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan. Association for Computational Linguistics.
- Li, J. and Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.
- LIN, Z., NG, H. T., and KAN, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.
- LINSINGEN, I. V. (2017). Perspectiva educacional cts: aspectos de um campo em consolidação na américa latina. *Ciência e Ensino*, 1(especial).
- Marcuschi, L. A. (2008). *Produção textual, análise de gêneros e compreensão*. Parábola Editorial, São Paulo.

- Marinho, J., Anchiêta, R., and Moura, R. (2022). Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*.
- MELLO, R. F., WENCESLAU, M., CORDEIRO, T., ISOTANI, S., and BITTENCOURT, I. I. (2023). Automated thematic coherence scoring of student essays written in portuguese. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, page 26737.
- MENEGASSI, R. J. and GASPAROTTO, D. M. (2016). Revisão textual-iterativa: aspectos metodológicos. *Revista Domínios da Linguagem*, 10(3).
- NAU, J., FILHO, A. H., PASSERO, G., and CAVACO, V. (2017). Uma ferramenta para identificar desvios de linguagem na língua portuguesa. In *Anais do 11º Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana*, pages 12–16, Uberlândia. Sociedade Brasileira de Computação.
- Pinho, C. M. d. A., Gaspar, M. A., and Sassi, R. J. (2022). Aplicação de técnicas de inteligência artificial para classificação de fuga ao tema em redações. *SciELO Preprints*. Postado em 25 abr. 2022.
- RODRIGUES, C., BRITO, D. M., ALMEIDA, P. E. G., XAVIER, P., and CUSTÓDIO, P. (2020). Importância da avaliação textual-iterativa na escrita de graduandos que cursaram a disciplinas on-line “oficina de língua portuguesa: leitura e produções de texto. *Miguilim*, 9(3).
- RODRIGUES, G. H. P. (2023). Desenvolvimento de um corretor automático e interface web para avaliação de redações do enem com redes neurais profundas. Monografia de Graduação.
- RUIZ, E. D. (2010). *Como corrigir redações na escola: uma proposta textual-iterativa*. Contexto, São Paulo.
- SERAFINI, M. T. (1989). *Como escrever textos*. Globo, São Paulo. Trad. Maria Augusta Bastos de Mattos.
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.
- SILVA, M. C. S., GONÇALVES, M. B., RÊGO, S. B., and DINIZ, M. I. (2013). O trabalho docente no processo de correção textual. In *Anais. V Fórum Nacional de Pedagogia*, Vitória da Conquista – BA.
- SOMASUNDARAN, S., BURSTEIN, J., and CHODOROW, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, 25th International Conference on Computational Linguistics*, Dublin, Ireland.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS 2017: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA.

Cuidados Éticos

Esse estudo utilizou uma base de dados pública (Essay-BR Corpus) na qual todos os textos foram previamente anonimizados, impedindo a identificação dos autores originais. Por essa natureza, e em conformidade com a Resolução CNS nº 510/2016, o estudo dispensa a tramitação e aprovação específicas por Comitê de Ética em Pesquisa (CEP), assegurando-se o respeito irrestrito aos direitos autorais, de imagem e à dignidade dos sujeitos indiretamente envolvidos.

Disponibilidade de Artefatos

Em compromisso com os princípios da Ciência Aberta e objetivando garantir o rigor metodológico e a replicabilidade integral dos achados discutidos, os códigos-fonte utilizados nas etapas de pré-processamento, otimização via Optuna e *fine-tuning* da arquitetura BERTimbau encontram-se estruturados em notebooks computacionais. Todos os artefatos experimentais, bem como os registros de treinamento, estão publicamente disponíveis no repositório oficial do projeto no GitHub. O acesso pode ser realizado através do link:

https://github.com/Ryancamaargo/Enem_Script_Fine_Tuning

Declaração sobre uso de Inteligência Artificial

Durante a preparação deste trabalho, os autores utilizaram ferramentas de Inteligência Artificial Generativa exclusivamente para auxílio na redação e revisão textual. A tecnologia não foi empregada na concepção da pesquisa, análise de dados ou obtenção de resultados. Os autores assumem total responsabilidade pela integridade e originalidade de todo o conteúdo apresentado, incluindo a precisão das informações e a ausência de plágio.