Avaliação Automática de Redações do Enem: Uma Análise Comparativa entre Engenharia de Características e Transformers

Pâmela Camilo Chalegre, Vitor da Rocha Machado, Valéria Delisandra Feltrim
Departamento de Informática
Universidade Estadual de Maringá
Maringá – PR – Brasil
{ra134241, ra132769, vdfeltrim}@uem.br

Resumo

A Avaliação Automática de Redações (AES) é um desafio central em avaliações educacionais de larga escala, como o Exame Nacional do Ensino Médio (Enem), no qual redações são avaliadas em múltiplas competências. Este trabalho apresenta uma análise comparativa de representações textuais para a AES em nível de competência no português brasileiro. Foram avaliados modelos baseados em características utilizando TF-IDF, métricas linguísticas extraídas com o NILC-Metrix e uma combinação híbrida de ambos, além de modelos baseados em transformers. Os experimentos foram conduzidos sobre o *corpus* Enem-AES, considerando formulações de classificação e de regressão. Os resultados indicam que formulações de regressão são, em geral, mais adequadas do que as de classificação multiclasse, pois acomodam melhor a estrutura ordinal das notas. Modelos baseados em transformers alcançaram uma concordância maior em competências relacionadas ao uso da linguagem e à coesão textual, enquanto representações baseadas em características demonstraram um desempenho comparável em competências associadas à pertinência temática. Apesar de alcançarem alta acurácia sob o critério de tolerância do Enem, todas as abordagens demonstraram dificuldade em prever notas extremas, principalmente devido ao desbalanceamento do *corpus*. Dessa forma, conclui-se que as metodologias são complementares e que sistemas híbridos são promissores para a AES.

1 Introdução

A avaliação de redações é uma tarefa central em contextos educacionais, especialmente em exames de larga escala, nos quais a correção manual demanda alto custo e está sujeita a variações interavaliador. No Brasil, o Exame Nacional do Ensino Médio (Enem) constitui um dos principais exemplos desse cenário, avaliando anualmente milhões de redações por meio de um esquema baseado em cinco competências distintas, cada uma pontuada em uma escala ordinal discreta (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025).

Nesse contexto, a Avaliação Automática de Redações (AES – *Automatic Essay Scoring*) tem sido amplamente investigada como uma alternativa para apoiar processos avaliativos em larga escala. Em exames padronizados internacionais, como TOEFL, GRE e GMAT, sistemas automáticos já são empregados de forma complementar à correção humana (Shermis et al., 2013). Para o português brasileiro, entretanto, o avanço da área foi historicamente limitado pela escassez de *corpora* públicos, pela heterogeneidade das escalas de pontuação adotadas e pela ausência de análises sistemáticas da qualidade das anotações humanas (Amorim e Veloso, 2017; Fonseca et al., 2018; Marinho et al., 2022a).

Recentemente, Silveira et al. (2024) propuseram o Enem-AES, um novo *benchmark* público para AES em português, construído a partir de redações dissertativo-argumentativas no modelo Enem e acompanhado de uma análise da distribuição das notas e da concordância interavaliador. Um dos resultados desse trabalho é que a AES se beneficia da modelagem como um problema de regressão ordinal, dada a natureza ordenada e discreta da escala de pontuação do Enem, bem como a sensibilidade das métricas de avaliação ao desbalanceamento das classes. Esse *benchmark* estabelece, assim, um protocolo experimental consistente para a comparação de diferentes abordagens de modelagem.

No que se refere às estratégias de modelagem, a literatura em AES para o português brasileiro apresenta dois paradigmas principais. O primeiro baseia-se na engenharia de características explícitas, extraindo métricas linguísticas e estatísticas dos textos para alimentar modelos supervisionados clássicos, como regressão linear, máquinas de vetores de suporte e métodos de *ensemble* (Amorim e Veloso, 2017; Fonseca et al., 2018). Nessa linha, ferramentas como o NILC-Metrix (Leal et al.,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 738–748

2023), que disponibilizam um amplo conjunto de métricas linguísticas, abrangendo níveis morfológico, sintático, semântico e discursivo, favorecem análises interpretáveis dos critérios avaliativos.

O segundo paradigma fundamenta-se em modelos de aprendizado profundo baseados na arquitetura Transformer (Vaswani et al., 2017), em particular variantes do BERT (Devlin et al., 2019) adaptadas ao português brasileiro. Esses modelos produzem representações contextuais capazes de capturar dependências semânticas e discursivas de longo alcance, tendo demonstrado resultados promissores em tarefas de AES, especialmente quando ajustados ao domínio do Enem (Silveira et al., 2024).

Ainda nesse paradigma, estudos recentes têm utilizado *Large Language Models* (LLMs) no contexto de AES. Esses modelos têm demonstrado potencial para avaliar textos e gerar *feedback* a partir do processamento de grandes volumes de dados textuais. Contudo, questões relacionadas a custos operacionais e a aspectos de equidade têm sido apontadas como desafios em aberto para a adoção dessas tecnologias em contextos educacionais com recursos limitados (Lobo et al., 2025). Portanto, abordagens baseadas em algoritmos clássicos e em modelos executados localmente permanecem como alternativas viáveis, pois demandam menor poder computacional e apresentam maior acessibilidade.

Apesar do avanço, permanece em aberto a compreensão sistemática de como esses dois paradigmas se comparam quando avaliados sob um mesmo protocolo experimental, especialmente no nível das competências individuais do exame. Diante desse cenário, este trabalho investigou, de forma comparativa e controlada, o desempenho de abordagens baseadas em representações explícitas e em representações contextuais na tarefa de avaliação automática das competências do Enem. Seguindo o *benchmark* proposto por Silveira et al. (2024), os experimentos foram conduzidos tratando cada anotação humana como uma instância independente e respeitando a divisão dos dados por comando. O problema foi analisado sob diferentes formulações: classificação, regressão e regressão ordinal.

Assim, as principais contribuições deste trabalho são: (i) uma comparação sistemática entre modelos baseados em representações TF-IDF, métricas NILC-Metrix e híbridas, e modelos baseados em transformers, conduzida sob um protocolo experimental unificado; (ii) uma análise empírica do impacto das diferentes formulações do problema na predição das cinco competências do Enem; e (iii) uma discussão dos padrões de erro observados, com foco nas dificuldades associadas ao desbalanceamento das classes e à predição de notas extremas. Os resultados obtidos reforçam a complementaridade entre os paradigmas analisados e indicam direções promissoras para o desenvolvimento de sistemas híbridos de AES em português.

2 Trabalhos Relacionados

Um dos primeiros trabalhos a explorar AES em redações do Enem foi o de Amorim e Veloso (2017), o qual apresenta um sistema estruturado a partir de modelos baseados em características manuais. O *corpus* foi construído com redações do *UOL Essay Database*, totalizando 1.840 textos distribuídos em 96 temas. O sistema foi avaliado nas competências e na nota final, utilizando o Kappa Ponderado Quadrático (QWK), atingindo valores entre 0,13 e 0,36. Ressalta-se que as redações foram avaliadas em incrementos de 50 pontos, diferentemente da escala oficial do Enem, que adota incrementos de 40 pontos.

Em linha semelhante, Fonseca et al. (2018) realizaram uma análise comparativa entre engenharia de características e métodos de aprendizado profundo, utilizando *BiLSTMs* com *embeddings GloVe*. Foi empregado um *corpus* privado com aproximadamente 56 mil textos, provenientes de uma plataforma online. Nas avaliações por competência, o método baseado em características manuais obteve valores de QWK entre 0,50 e 0,67, enquanto o modelo neural apresentou resultados entre 0,50 e 0,63. Para a nota final, ambas famílias de modelos alcançaram valores similares, em torno de 0,75, indicando desempenhos comparáveis e o potencial de abordagens híbridas.

Marinho et al. (2022a) propuseram a expansão e disponibilização do *corpus* público Essay-BR, visando mitigar a escassez de recursos abertos para AES em língua portuguesa. O *corpus* é composto por 6.579 redações, distribuídas em 151 temas, coletadas dos *websites* Brasil Escola e Educação UOL, e avaliadas segundo a escala oficial adotada pelo Enem. Utilizando técnicas empregadas por Amorim e Veloso (2017) e Fonseca et al. (2018) nesse novo *corpus*, os valores QWK reportados situaram-se entre 0,34 e 0,51.

De forma complementar, Marinho et al. (2022b) apresentaram estratégias baseadas na criação de modelos preditivos específicos e direcionados para

cada uma das cinco competências consideradas na avaliação de redações do Enem. Utilizando um subconjunto de 5.730 redações provenientes do *corpus* Essay-BR, os autores estudaram a utilização de técnicas baseadas em características *ad-hoc*, *embeddings* estáticos, como o *Doc2Vec* e *Word2Vec*, e redes neurais recorrentes do tipo *Long Short-Term Memory* (LSTM) para a avaliação automática de redações. Os resultados demonstraram que as características *ad-hoc* foram mais eficientes para as competências 1 e 2, enquanto as redes LSTM obtiveram o melhor desempenho nas competências 3, 4 e 5, alcançando uma concordância moderada, com QWK entre 0,44 e 0,56, e reforçando a importância de tratar a correção de cada competência como uma tarefa de predição independente. Um diferencial desse estudo foi a comparação direta entre o desempenho de *embeddings* estáticos e o uso de características *ad-hoc* para o Enem.

Posteriormente, Silveira et al. (2024) identificaram limitações no *corpus* Essay-BR e construíram um novo conjunto de dados público composto por 3.586 redações, provenientes das mesmas fontes utilizadas por Marinho et al. (2022a). Além disso, o estudo contou com dois avaliadores humanos, que anotaram as redações coletadas de uma das fontes de dados. A concordância entre esses avaliadores superou a concordância com as notas originais da plataforma, evidenciando possíveis inconsistências nas avaliações online. Os preditores construídos basearam-se no modelo BERTimbau (Souza et al., 2020), alcançando valores de QWK entre 0,23 e 0,53, indicando o potencial dos modelos BERT, especialmente em formulações de regressão ordinal.

Em um estudo subsequente, Silveira et al. (2025) investigaram a robustez de sistemas de AES baseados em transformers frente a ataques adversariais universais. Os autores analisaram modelos baseados em BERT, Phi-3 e Gemini utilizando o conjunto de dados Enem-AES. Como parte da metodologia, foi treinado um regressor linear baseado em um subconjunto de características linguísticas extraídas pelo sistema NILC-Metrix. Essas características foram utilizadas para identificar padrões exploráveis pelos modelos e para construir estratégias de ataque adversarial. Os resultados indicaram que, embora modelos neurais apresentem bom desempenho em termos de concordância com avaliadores humanos, eles ainda podem ser sensíveis a pistas superficiais presentes nos textos. Também é interessante notar que, comparativamente, o regressor linear apresentou resultados competitivos em duas das cinco competências.

Abordando tecnologias mais recentes, Barbosa et al. (2025) realizaram uma análise empírica abrangente de LLMs para a tarefa de AES *cross-prompt* em português. A partir de um *corpus* proveniente do conjunto de dados utilizado em (Silveira et al., 2024), formado por 385 redações, o estudo analisou 12 modelos com diferentes estratégias de ajuste fino e arquiteturas, abrangendo estratégias baseadas em características, em modelos BERT, em pequenos modelos de linguagem e LLMs com *zero-shot learning*. Os resultados estabeleceram o estado da arte para o conjunto de dados, alcançando valores de QWK entre 0,60 e 0,73. Os autores concluíram que nenhuma arquitetura é universalmente superior, de forma que os modelos BERT demonstraram-se ideais pelo equilíbrio entre custo computacional e desempenho, enquanto LLMs com *zero-shot* com janelas de contexto amplas destacaram-se em competências mais subjetivas e argumentativas.

Explorando um contexto educacional diferente, Lobo et al. (2025) investigaram o uso de LLMs na avaliação automática de redações narrativas escritas por alunos do ensino fundamental no Brasil, por meio do Gemini 2.0 Flash, GPT-4o, Claude 3.7 e Mistral Large. A pesquisa utilizou um *corpus* de 1.235 redações, avaliadas em quatro competências específicas. O estudo revelou que o modelo Claude 3.7 obteve a maior concordância com os avaliadores humanos, atingindo valores QWK entre 0,30 e 0,62, enquanto o Gemini 2.0 Flash destacou-se na taxa de acertos exatos. Diante dessa variação de desempenho, os autores concluíram que a eficácia da avaliação automática depende do alinhamento entre a escolha do modelo e as prioridades pedagógicas do avaliador, de forma a equilibrar a necessidade de uma consistência geral nas notas com a precisão rigorosa em critérios específicos.

Diferenciando-se desses trabalhos, o presente estudo investiga, de forma comparativa, abordagens baseadas em engenharia de características, incluindo características superficiais e NILC-Metrix, e modelos baseados em transformers com base no *benchmark* de Silveira et al. (2024). Ambas as abordagens foram avaliadas sob o mesmo protocolo, garantindo a comparabilidade direta dos resultados.

3 Materiais e Métodos

3.1 *Corpus*

Os experimentos foram conduzidos utilizando o Enem-AES (Silveira et al., 2024), em particular, o

subconjunto denominado *Source A With Graders*[1]. A escolha desse subconjunto se deve ao mesmo apresentar uma distribuição de notas mais equilibrada entre as competências, por contar com as anotações humanas e por ter sido utilizado nos experimentos reportados por Silveira et al. (2024). Esse subconjunto é composto por redações produzidas entre os anos de 2015 e 2020 e foi coletado do portal Educação UOL. Cada redação está associada a um comando (proposta de redação) e foi avaliada segundo as cinco competências do Enem, com notas discretas no conjunto {0, 40, 80, 120, 160, 200}, além das notas finais.

Seguindo o protocolo estabelecido por Silveira et al. (2024), neste estudo, cada anotação humana disponível foi tratada como uma instância independente, permitindo modelar explicitamente a variabilidade interavaliador inerente ao processo de correção manual.

Durante a análise dos dados, foi identificada e removida uma instância cuja escala de pontuação das competências variava de 0 a 2, em desacordo com o padrão adotado pelo Enem. Após esse procedimento, o conjunto final totalizou 1.164 instâncias.

O particionamento dos dados adotado seguiu o definido no *benchmark* original. O *corpus* está dividido em conjuntos de treino, validação e teste, com 757, 198 e 209 instâncias, respectivamente, correspondendo a aproximadamente 65%, 17% e 18% do total. A divisão foi realizada de modo que todas as redações associadas a um mesmo comando permanecessem no mesmo subconjunto, prevenindo vazamento de informações entre as partições e garantindo a comparabilidade dos resultados entre diferentes abordagens.

3.2 Abordagem Baseada em Engenharia de Características

Na abordagem baseada em engenharia de características, as redações são representadas por vetores numéricos explícitos que capturam propriedades linguísticas e estatísticas do texto, servindo como entrada para modelos supervisionados de regressão, regressão ordinal e classificação multiclasse.

3.2.1 Representações Textuais

Foram consideradas três representações distintas. Como *baseline*, adotou-se a representação TF-IDF (*Term Frequency-Inverse Document Frequency*), uma técnica clássica de *bag-of-words* que pondera a importância dos termos com base em sua frequência no documento e sua raridade no *corpus* (Jurafsky e Martin, 2026).

A configuração do vetorizador foi adaptada para cada competência, adequando a representação textual às particularidades linguísticas avaliadas. Na C1, priorizaram-se n-gramas de caracteres, uma vez que esse nível de representação pode capturar variações morfológicas e desvios ortográficos associados ao domínio da norma padrão. Já nas competências relacionadas ao conteúdo e à argumentação (C2, C3 e C5), empregaram-se n-gramas de palavras, com remoção de *stopwords* para dar maior peso ao conteúdo temático. Na C4, que avalia coesão textual, optou-se por manter conectivos e marcadores discursivos, preservando sinais de encadeamento textual. Em todas as competências, aplicou-se seleção supervisionada de características para reduzir a dimensionalidade: utilizou-se informação mútua na formulação de regressão e o teste χ^2 na formulação de classificação.

A segunda representação consistiu das métricas extraídas pelo sistema NILC-Metrix (Leal et al., 2023). O sistema disponibiliza 200 métricas organizadas em 14 categorias, abrangendo níveis morfológico, sintático, semântico e discursivo. Após o alinhamento dos vetores aos textos, procedeu-se à limpeza de instâncias com valores ausentes. Em seguida, foram removidos os atributos com baixa variância e aplicou-se padronização por *z-score*, seguido de seleção supervisionada de características.

Por fim, a representação híbrida combinou informações superficiais de vocabulário e indicadores linguísticos. Os atributos TF-IDF e as métricas NILC-Metrix foram processados em fluxos independentes e posteriormente concatenados. Em ambos os fluxos, aplicou-se seleção supervisionada de características, escolhendo-se, em cada fluxo, as 50 características mais informativas de acordo com a tarefa. Para isso, utilizou-se informação mútua nos modelos de regressão; para a classificação, aplicou-se o teste χ^2 aos atributos TF-IDF e o teste F da ANOVA às métricas padronizadas do NILC-Metrix. O vetor híbrido resultante totalizou 100 dimensões.

3.2.2 Algoritmos de Aprendizado de Máquina

Seguindo a metodologia proposta por Silveira et al. (2024), para as tarefas de classificação e regressão ordinal, as notas das competências (0, 40, 80, 120, 160, 200) foram mapeadas para uma escala ordinal de 0 a 5. O problema de AES para essa abordagem foi, então, modelado sob três perspectivas: regres-

[1] https://huggingface.co/datasets/kamel-usp/aes_enem_dataset/viewer/sourceAWithGraders

são, classificação e regressão ordinal.

Para as abordagens de regressão e classificação, foram treinados e avaliados diversos algoritmos da biblioteca *scikit-learn*[2]. Para a regressão, foram testados modelos lineares (*LinearRegression, Ridge, Lasso*), baseados em vizinhos (*KNeighborsRegressor*), máquinas de vetores de suporte (*SVR*), modelos de *ensemble* (*RandomForestRegressor, GradientBoostingRegressor, HistGradientBoostingRegressor*), árvores de decisão (*DecisionTreeRegressor*) e redes neurais (*MLPRegressor*).

Para a classificação, optou-se pelo uso do *Support Vector Classifier* (SVC), dada sua eficácia em espaços de alta dimensionalidade. Seus hiperparâmetros foram otimizados por meio de busca em grade com validação cruzada (*GridSearchCV*).

Para a regressão ordinal, foi utilizada a biblioteca `mord` (Pedregosa et al., 2017), especificamente os algoritmos *LogisticAT*, *LogisticIT*, *OrdinalRidge* e *Least Absolute Deviation* (LAD). No total, foram avaliados 42 modelos de regressão e três modelos de classificação.

3.3 Abordagem Baseada em Transformers

Na abordagem baseada em transformers, cada redação é representada por *embeddings* contextuais extraídos de um modelo pré-treinado, que codifica automaticamente propriedades linguísticas e semânticas. Diferentemente das abordagens baseadas em características linguísticas explícitas, esses vetores não são construídos manualmente, mas podem ser ajustados para a tarefa de avaliação automática de redações por meio de *fine-tuning*.

Neste estudo, os preditores basearam-se no modelo BERTimbau (Souza et al., 2020). O ajuste fino foi realizado com o auxílio da biblioteca *SimpleTransformers*[3], que abstrai o treinamento supervisionado de modelos baseados em transformers. Utilizou-se a variante *BERTimbau-base*[4], uma vez que a adoção da versão *large* não se mostrou vantajosa frente ao tamanho reduzido do *corpus*, conforme discutido por Silveira et al. (2024).

Nessa abordagem, duas formulações do problema foram consideradas: classificação, com rótulos correspondentes aos valores discretos da escala de avaliação, e regressão, com uma única saída para a predição das notas. Novamente, cada competência foi tratada como uma tarefa independente. As notas originais foram mapeadas para o intervalo de 0 a 5 e a saída contínua dos regressores foi posteriormente arredondada para o rótulo de classe inteiro mais próximo.

Os modelos foram treinados utilizando os conjuntos de treino e validação, com experimentação de diferentes combinações de hiperparâmetros para cada competência. Para mitigar *overfitting*, empregou-se *early stopping* monitorado pela perda no conjunto de validação. As configurações avaliadas variaram o número de épocas, a taxa de aprendizado, o decaimento de pesos, o tamanho de lote (treino e validação) e a frequência de avaliação durante o treinamento. No total, foram treinados 70 modelos de classificação e 31 modelos de regressão nas cinco competências. A configuração detalhada dos experimentos, bem como a implementação da extração de características, encontra-se disponíveis no repositório do projeto[5].

3.4 Métricas de Avaliação

A avaliação dos modelos foi conduzida considerando a natureza ordinal das notas atribuídas às competências e seguindo as recomendações do *benchmark* proposto por Silveira et al. (2024).

A principal métrica utilizada foi Kappa Ponderado Quadrático (QWK – *Quadratic Weighted Kappa*), amplamente adotado em tarefas de AES por medir o grau de concordância entre as notas previstas e as notas de referência, penalizando mais severamente discrepâncias de maior magnitude. Valores de QWK próximos a zero indicam concordância próxima ao acaso, enquanto valores mais elevados refletem maior concordância entre predição e anotação humana.

Também foi calculada a Raiz do Erro Quadrático Médio (RMSE – *Root Mean Squared Error*), que quantifica o desvio médio das predições, atribuindo maior peso a erros maiores. Essa métrica foi inicialmente calculada na escala ordinal de 0 a 5 e posteriormente reconvertida para a escala original do Enem (0–200), multiplicando-se os valores de erro por um fator de 40 pontos. Essa conversão permitiu uma interpretação mais direta do desvio médio das predições em termos da pontuação oficial do exame.

Adicionalmente, seguindo Silveira et al. (2024), foram consideradas a acurácia exata (ACC) e a acurácia no padrão do Enem (ACC-E). Nesta última, a previsão é dada como correta se estiver dentro de

[2] https://scikit-learn.org/stable/

[3] https://simpletransformers.ai/

[4] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[5] https://github.com/hito-boo/aes-enem-transformers-features

uma margem tolerável de 80 pontos em relação à nota real. Para o cálculo de acurácia, as saídas contínuas dos modelos regressores foram discretizadas para a classe ordinal mais próxima da escala Enem.

Além disso, a matriz de confusão associada a cada preditor também foi analisada para identificar padrões recorrentes nos erros e rótulos mais frequentemente previstos. Outras métricas de classificação, como precisão, revocação e *F1-score*, foram coletadas nos modelos de classificação. Em particular, foi considerada a média macro da *F1-score* (F1-Macro), de modo a atribuir o mesmo peso a todas as classes e evidenciar o impacto do desbalanceamento na capacidade dos modelos de distinguir faixas de pontuação menos frequentes.

4 Resultados e Discussão

4.1 Modelos Baseados em Características

Os modelos baseados em engenharia de características (TF-IDF, NILC-Metrix e Híbrida) apresentaram desempenhos distintos ao longo das cinco competências. De forma consistente, os melhores resultados foram obtidos por modelos de regressão ou regressão ordinal, enquanto os classificadores mostraram desempenho inferior, especialmente quando avaliados por métricas sensíveis à ordinalidade e ao desbalanceamento das classes, como o QWK e o F1-Macro.

A Tabela 1 apresenta, para cada competência, o melhor modelo regressor obtido com cada representação. Diferentemente da tarefa de classificação, os regressores não passaram por seleção de hiperparâmetros, sendo treinados com configurações padrão e avaliados no conjunto de teste. Para fins comparativos, destaca-se o modelo com maior valor de QWK, utilizando RMSE e ACC-E como critérios de desempate em casos de valores muito próximos de QWK. Observa-se que os valores de QWK situaram-se, em geral, na faixa de concordância fraca a moderada, comportamento compatível com os níveis de acordo observados entre avaliadores humanos no próprio *benchmark*.

A análise por competência evidencia padrões distintos conforme o tipo de representação adotada. A competência C1 (Norma Culta) apresentou o melhor desempenho global entre os modelos baseados em características, alcançando QWK de 0,39 com a representação híbrida. Esse resultado sugere uma complementaridade entre as métricas linguísticas do NILC-Metrix, que capturam aspectos estruturais e gramaticais, e a representação TF-IDF, que incorpora informações lexicais relevantes para a avaliação da correção linguística.

A superioridade do TF-IDF nas competências C2 (Compreensão da Proposta) e C3 (Argumentação) indica que, por representar diretamente o conteúdo textual produzido pelo estudante, essa representação foi mais informativa quanto à presença de termos relacionados ao tema, variedade de vocabulário e ao uso de conectores argumentativos.

Para C4 (Coesão Textual) e C5 (Proposta de Intervenção), a abordagem Híbrida apresentou desempenho superior ou equivalente às demais representações. No caso da C4, a combinação entre indicadores linguísticos de coesão e informações lexicais favoreceu a identificação de padrões de encadeamento textual. Na C5, cuja avaliação exige a identificação de múltiplos elementos obrigatórios da proposta de intervenção, a integração entre estrutura sintática e vocabulário específico também mostrou-se vantajosa.

Embora as métricas NILC-Metrix incorporem indicadores de complexidade e estrutura linguística, essa representação teve desempenho inferior na maioria das competências. Esse comportamento indica que tais métricas, isoladamente, não são suficientes para capturar critérios avaliativos fortemente dependentes do conteúdo temático e semântico do texto, embora tragam ganhos quando combinadas com TF-IDF. Cabe destacar ainda que tais métricas foram desenvolvidas para outra finalidade, de modo que seu baixo desempenho quando utilizadas de forma isolada era, de certa forma, esperado.

Em uma análise global das métricas, nota-se uma discrepância entre a ACC (acurácia exata) e a ACC-E (acurácia no padrão Enem). Enquanto a ACC variou entre aproximadamente 20% e 47%, a ACC-E manteve-se elevada, entre 79% e 95%. Em paralelo, os valores de RMSE situaram-se, em média, entre 42 e 59 pontos, indicando que os erros tendem a se concentrar em um ou dois níveis adjacentes da escala oficial de pontuação, comportamento compatível com o esperado em tarefas de AES sob forte desbalanceamento de notas.

Os resultados dos classificadores, sumarizados na Tabela 2, evidenciam as limitações dessa formulação para o problema em questão. Embora, em alguns casos, a ACC apresente valores aparentemente competitivos, os valores de QWK próximos de zero e os baixos valores de F1-Macro indicam um viés acentuado em favor das classes majoritárias. Esse padrão revela que os classificadores tendem a atribuir a nota mais frequente a grande parte

Comp.	Abordagem	Modelo	QWK	RMSE	ACC-E	ACC
C1	**Híbrido**	Gradient Boosting Regressor	**0,39**	**42,51**	0,92	**0,47**
	TF-IDF	OrdinalRidge	0,38	43,04	**0,95**	0,40
	NILC	LogisticAT	0,00	51,69	0,92	0,40
C2	**TF-IDF**	MLP	**0,33**	**49,78**	0,86	0,31
	Híbrido	Linear Regression	0,24	58,26	0,81	0,27
	NILC	Least Absolute Deviation	0,03	59,73	**0,92**	**0,34**
C3	**TF-IDF**	MLP	**0,25**	**48,12**	**0,92**	0,28
	Híbrido	Gradient Boosting Regressor	0,15	53,20	0,85	**0,31**
	NILC	LogisticIT	0,00	56,14	**0,92**	0,28
C4	**Híbrido**	Gradient Boosting Regressor	0,32	**43,62**	**0,93**	0,43
	TF-IDF	MLP	**0,33**	43,72	0,92	0,41
	NILC	LogisticIT	0,00	47,32	0,92	**0,45**
C5	**Híbrido**	Least Absolute Deviation	**0,33**	56,49	**0,94**	**0,28**
	TF-IDF	MLP	0,25	**54,19**	0,86	0,25
	NILC	MLP	0,01	58,19	0,79	0,20

Tabela 1: Desempenho dos melhores regressores por representação e competência.

das instâncias, falhando em distinguir adequadamente as faixas de pontuação menos representadas.

Por fim, a Tabela 3 detalha as predições do melhor modelo baseado em características para a C1. A matriz revela uma grande concentração de acertos para as notas intermediárias (120 e 160) e uma dificuldade do modelo em identificar as notas extremas: as redações com nota 200 foram majoritariamente preditas como 160 e as notas 0 foram superestimadas. Esse comportamento decorre da distribuição assimétrica das notas no *corpus* e reforça a necessidade de estratégias que lidem com o desbalanceamento e a ordinalidade dos dados.

4.2 Modelos Baseados em Transformers

Os modelos baseados em transformers apresentaram desempenho moderado, com superioridade consistente dos modelos de regressão em relação aos classificadores. A Tabela 4 sintetiza os melhores resultados obtidos por competência.

De modo geral, os valores de QWK obtidos pelos regressores situaram-se entre 0,28 e 0,45, indicando níveis de concordância comparáveis aos observados entre avaliadores humanos no *benchmark*. Esse comportamento reforça a adequação da formulação de regressão para a tarefa, uma vez que a saída contínua permite capturar relações de proximidade entre os níveis de pontuação, aspecto fundamental em escalas ordinais como a do Enem.

A análise por competência indica diferenças no desempenho dos modelos. Novamente, a C1 apresentou o maior valor de QWK, alcançando 0,45. Desempenhos relativamente elevados também foram observados nas competências C3 e C4, com valores de QWK de 0,42 e 0,43, respectivamente. Esses resultados sugerem que os modelos conseguem capturar informações relevantes para a avaliação de aspectos gramaticais, discursivos e de organização textual, embora não seja possível isolar, a partir desses experimentos, quais propriedades específicas das representações contribuíram para esse desempenho.

Por outro lado, as competências C2 e C5 apresentaram valores de QWK mais baixos, em torno de 0,33. Essas competências envolvem critérios que dependem da identificação de elementos semânticos específicos e do atendimento a requisitos estruturais explícitos, o que pode representar um desafio adicional em cenários de *fine-tuning* supervisionado com *corpus* de tamanho reduzido e distribuição desbalanceada. Adicionalmente, deve-se considerar que os modelos avaliados não foram formulados explicitamente como regressores ordinais, fator que pode ter limitado a exploração da estrutura ordenada da escala de pontuação.

No que se refere às métricas complementares, observa-se que, apesar dos valores relativamente baixos de ACC e F1-Macro, variando, em média, entre 24% e 49% e entre 15% e 26%, respectivamente, a ACC-E manteve-se elevada em todas as competências, frequentemente acima de 90%. Esse resultado indica que, embora as predições nem sempre coincidam exatamente com a nota atribuída pelos avaliadores humanos, elas tendem a

Comp.	Abordagem	QWK	F1-Macro	ACC-E	ACC
C1	**TF-IDF**	**0,23**	**0,17**	**0,93**	**0,38**
	Híbrido	0,11	0,14	0,92	0,34
	NILC	0,00	0,13	0,92	0,34
C2	**TF-IDF**	**0,09**	0,11	**0,92**	0,33
	Híbrido	0,07	**0,15**	0,90	0,34
	NILC	0,00	0,09	**0,92**	**0,35**
C3	**TF-IDF**	**0,09**	**0,16**	0,87	**0,30**
	NILC	0,02	0,12	**0,91**	**0,30**
	Híbrido	-0,06	0,12	0,85	0,28
C4	**TF-IDF**	**0,16**	0,18	**0,92**	**0,49**
	NILC	0,00	0,10	**0,92**	0,45
	Híbrido	-0,03	**0,19**	0,91	0,39
C5	**TF-IDF**	**0,09**	**0,15**	0,85	**0,22**
	NILC	0,01	0,09	**0,90**	0,19
	Híbrido	-0,08	0,11	0,80	0,14

Tabela 2: Desempenho dos classificadores por representação e competência.

		Predita					
		0	**40**	**80**	**120**	**160**	**200**
Real	**0**	0	1	8	4	3	0
	40	0	0	1	0	1	0
	80	0	2	6	16	2	0
	120	0	0	13	55	15	0
	160	0	0	5	24	34	0
	200	0	0	0	1	9	0

Tabela 3: Matriz de confusão do modelo *Gradient Boosting Regressor* com representação Híbrida para a C1.

situar-se dentro de um intervalo de erro considerado aceitável no contexto do exame. Os valores de RMSE, convertidos para a escala original do Enem, permaneceram na faixa aproximada de 50 pontos, reforçando que os erros médios dos regressores concentram-se em discrepâncias moderadas. Esse comportamento é compatível com a distribuição das notas no *corpus* e com a penalização imposta pelo QWK a erros de maior magnitude.

Por fim, observou-se um padrão recorrente nos erros dos modelos: a maior parte das predições concentra-se nas notas intermediárias, enquanto as notas extremas apresentam maior dificuldade de previsão. Esse comportamento, ilustrado na Tabela 5, reflete o desbalanceamento do *corpus*, que induz os modelos a favorecer as classes majoritárias como estratégia de minimização do erro global.

4.3 Análise Comparativa

Ao realizar uma análise comparativa dos diferentes modelos empregados em cada abordagem, observa-se que os melhores resultados foram obtidos majoritariamente por modelos regressores, tanto na abordagem baseada em engenharia de características quanto nos transformers. Esse resultado reforça a adequação de formulações que respeitam a natureza ordinal das notas do Enem, em consonância com as recomendações do *benchmark* adotado.

Nesse contexto, os modelos baseados em transformers demonstraram maior capacidade de avaliar aspectos relacionados à norma culta, à argumentação e à coesão textual. Essas competências dependem da modelagem de relações contextuais e discursivas ao longo do texto, aspecto no qual representações contextuais profundas se mostram mais eficazes. Por outro lado, o modelo baseado em TF-IDF demonstrou eficácia similar ao do transformer na avaliação da compreensão da proposta, assim como o modelo híbrido na avaliação da proposta de intervenção, tarefas fortemente associadas à identificação de palavras-chave, termos relevantes e à diversidade vocabular. Nesse contexto, vale destacar que a abordagem baseada em engenharia de características apresenta um consumo de recursos computacionais bem menor quando comparada aos modelos baseados em transformers.

Assim, observa-se que a abordagem baseada em características ainda mostra-se competitiva para tarefas cuja avaliação está relacionada à identificação da presença de termos e expressões espe-

	Classificadores				Regressores			
Comp.	**QWK**	**F1-Macro**	**ACC-E**	**ACC**	**QWK**	**RMSE**	**ACC-E**	**ACC**
C1	0,38	0,25	0,93	0,54	0,45	42,39	0,94	0,49
C2	0,33	0,24	0,91	0,40	0,28	53,95	0,95	0,36
C3	0,29	0,21	0,93	0,36	0,42	47,41	0,98	0,35
C4	0,36	0,26	0,92	0,50	0,43	43,01	0,95	0,49
C5	0,31	0,15	0,95	0,24	0,34	51,95	0,98	0,24

Tabela 4: Desempenho dos melhores preditores baseados em transformers por competência.

		Predita					
		0	**40**	**80**	**120**	**160**	**200**
Real	**0**	3	7	26	4	0	0
	40	0	3	22	12	0	0
	80	0	0	26	18	0	0
	120	0	2	24	18	0	0
	160	0	0	8	32	0	0
	200	0	0	1	3	0	0

Tabela 5: Matriz de confusão do regressor melhor avaliado na C5.

cíficas, enquanto os modelos baseados em transformers destacam-se mais em tarefas cuja avaliação depende da compreensão contextual do uso dessas construções ao longo do texto. Nesse sentido, uma abordagem híbrida que combine ambas as metodologias mostra-se promissora, uma vez que diferentes competências envolvem demandas distintas que podem ser atendidas por arquiteturas complementares mais leves.

Por fim, em ambas as abordagens, observa-se uma concentração das predições nas classes intermediárias da escala de pontuação, com maior dificuldade na identificação das notas extremas. Esse efeito é particularmente acentuado nos modelos de classificação e nos regressores com pós-processamento por arredondamento, que tendem a atribuir rótulos centrais com maior frequência. Ainda assim, observa-se que a maioria das predições permanece dentro da margem de tolerância de 80 pontos adotada no Enem, o que explica os elevados valores de acurácia no padrão do exame, mesmo quando a acurácia exata é limitada.

5 Conclusão

Este trabalho apresentou uma análise comparativa entre abordagens baseadas em engenharia de características e em modelos baseados em transformer para a AES no contexto do Enem, considerando a predição das notas por competência sob um protocolo unificado e alinhado ao *benchmark* de Silveira et al. (2024).

Os resultados evidenciam que a avaliação automática de redações do Enem não deve ser tratada como uma escolha binária entre abordagens, mas como um problema de complementaridade. Os modelos baseados em transformers mostraram-se superiores na captura de nuances gramaticais e de coesão, enquanto a abordagem baseada em engenharia de características ainda permanece competitiva para avaliar aspectos como a pertinência temática, além de apresentarem menor custo computacional. Portanto, uma solução promissora consiste na construção de sistemas híbridos que combinem a precisão contextual dos transformers com a interpretabilidade de métricas linguísticas e estatísticas textuais.

Como trabalhos futuros, em um primeiro momento, pretende-se ampliar a análise comparativa das abordagens com uma análise qualitativa dos erros cometidos pelos diferentes preditores. Essa análise pode contribuir para compreender os fatores que favorecem os modelos baseados em engenharia de características e em quais contextos os modelos baseados em transformers demonstram maior capacidade de generalização. Outra direção promissora consiste em explorar o uso de *Large Language Models* (LLMs) para tarefas relacionadas à AES, investigando seu potencial em combinação com modelos mais leves executados localmente.

Por fim, a dificuldade recorrente na predição de notas extremas indica que a principal limitação atual reside na escassez de dados rotulados equilibrados. Assim, trabalhos futuros devem priorizar a expansão e curadoria de *corpora* públicos, com foco na obtenção de exemplos nas faixas de pontuação menos representadas, de modo a permitir que os modelos identifiquem com maior segurança notas em todas as faixas do espectro de avaliação.

Limitações

Este trabalho apresenta limitações que devem ser consideradas na interpretação dos resultados. A principal delas refere-se à dimensão e à distribuição do *corpus*. Composto por apenas 1.164 redações, o conjunto impõe restrições especialmente ao treinamento de modelos baseados em aprendizado profundo. Além disso, a distribuição das notas é fortemente desbalanceada, com predominância de faixas intermediárias de pontuação. Esse cenário favorece as classes majoritárias e penaliza o desempenho nas notas extremas, conforme evidenciado pelos baixos valores de F1-Macro.

No caso das formulações de regressão, a necessidade de discretizar as saídas contínuas por meio de arredondamento constitui outra fonte potencial de viés. Esse procedimento, embora necessário para o cálculo de métricas discretas, tende a favorecer notas intermediárias em cenários desbalanceados e contribui para a dificuldade observada na predição de notas extremas. A adoção de uma arquitetura apropriada à regressão ordinal poderia amenizar esse efeito.

Outra limitação está relacionada à natureza subjetiva da correção humana. Embora os avaliadores tenham sido treinados, a avaliação manual está sujeita a inconsistências. Assim, parte dos erros atribuídos aos preditores pode, na realidade, refletir divergências entre avaliadores humanos.

Uma questão metodológica adicional decorre da forma como múltiplas anotações foram tratadas. Cada anotação humana foi considerada como uma instância independente, conforme o protocolo do *benchmark* adotado. Embora essa estratégia permita modelar a variabilidade interavaliador e aumentar o número de instâncias disponíveis para treinamento, ela introduz dependência entre exemplos oriundos de um mesmo texto, o que pode inflar levemente as estimativas de desempenho. Esse efeito foi mitigado pelo particionamento dos dados por comando, mas não pode ser completamente eliminado.

Também há restrições referentes às métricas de avaliação. Embora o QWK e a acurácia no padrão Enem sejam adequados para o contexto deste trabalho e amplamente utilizados na área de AES, em particular o QWK, eles priorizam a concordância estatística em detrimento de aspectos qualitativos, como originalidade ou profundidade argumentativa.

Por fim, destaca-se a delimitação do domínio, uma vez que os experimentos focaram exclusivamente no gênero dissertativo-argumentativo do Enem. Portanto, os resultados deste estudo não devem ser generalizados para outros gêneros textuais e contextos avaliativos, visto que diferentes domínios implicam demandas distintas.

Agradecimentos

Os autores agradecem à Pró-Reitoria de Ensino da Universidade Estadual de Maringá pelo suporte financeiro disponibilizado aos estudantes envolvidos neste projeto.

Referências

Evelin Amorim e Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. Em *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 94–102, Valencia, Spain. Association for Computational Linguistics.

André Barbosa, Igor Cataneo Silveira, e Denis Deratani Mauá. 2025. 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):857–870.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, e Alessandro Bokan. 2018. Automatically grading Brazilian student essays. Em *Proceedings of the 13th International Conference on Computational Processing of Portuguese (PROPOR 2018)*, volume 11122 de *Lecture Notes in Artificial Intelligence*, páginas 170–179, Canela, RS, Brazil. Springer.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 2025. *A Redação no Enem 2025: Cartilha do Participante*. Brasília, DF.

Dan Jurafsky e James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edição. Online manuscript released January 6, 2026.

Sidney E. Leal, Magali S. Duran, Carolina E. Scarton, Nathan S. Hartmann, e Sandra M. Aluísio. 2023. NILC-metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58:73–110.

Jamilla Lobo, Lenon Anthony, Andreza Falcão, Cleon Xavier, Newarney Torrezão, Seiji Isotani, Ig Ibert, Luiz Rodrigues, e Rafael Mello. 2025. Automatic scoring of elementary school essays in Brazilian Portuguese with LLMs: Comparing Gemini, GPT-4o, Claude, and Mistral. Em *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, páginas 167–180, Porto Alegre, RS, Brazil. SBC.

Jeziel Marinho, Rafael Anchiêta, e Raimundo Moura. 2022a. Essay-br: a Brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1):65–76.

Jeziel Marinho, Fábio Cordeiro, Rafael Anchiêta, e Raimundo Moura. 2022b. Automated essay scoring: An approach based on ENEM competencies. Em *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, páginas 49–60, Porto Alegre, RS, Brazil. SBC.

Fabian Pedregosa, Francis Bach, e Alexandre Gramfort. 2017. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35.

Mark D. Shermis, Jill Burstein, e Sharon Apel Bursky. 2013. Introduction to automated essay evaluation. Em Mark D. Shermis e Jill Burstein, editores, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, páginas 1–15. Taylor & Francis, New York, NY.

Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, e Denis Deratani Mauá. 2025. Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. Em *Intelligent Systems*, páginas 169–183, Cham. Springer Nature Switzerland.

Igor Cataneo Silveira, André Barbosa, e Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in Portuguese. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, Galicia, Spain. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, e Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, páginas 403–417, Cham. Springer International Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, e Illia Polosukhin. 2017. Attention is all you need. Em *Advances in Neural Information Processing Systems*, volume 30, páginas 5998–6008.