CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

AVALIAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA A CORREÇÃO AUTOMÁTICA DE REDAÇÕES SEGUNDO AS COMPETÊNCIAS DO ENEM

Vanessa Santos Soares

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Bezerra da Silva, D.Sc.
Gustavo Paiva Guedes e Silva, D.Sc.

Rio de Janeiro,
Junho de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S676 Soares, Vanessa Santos
Avaliação de modelos de aprendizado de máquina para a correção automática de redações segundo as competências do ENEM / Vanessa Santos Soares. — 2026.
80f. + apêndice : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 73-80
Orientador: Eduardo Bezerra da Silva
Coorientador: Gustavo Paiva Guedes e Silva

1. Avaliação educacional. 2. Língua portuguesa - Correção de textos. 3. Aprendizado do computador. 4. Exame Nacional do Ensino Médio (Brasil) - Avaliação. 4. Redação acadêmica. I. Silva, Eduardo Bezerra da. (Orient.). II. Silva, Gustavo Paiva Guedes e. (Coorient.). III. Título.

CDD 371.26

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

Avaliação de modelos de aprendizado de máquina para a correção automática de redações segundo as competências do ENEM

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Vanessa Santos Soares

Aprovada por:

Presidente, Eduardo Bezerra da Silva, D.Sc. (CEFET/RJ)

Gustavo Paiva Guedes e Silva, D.Sc.

Diego Moreira de Araújo Carvalho (CEFET/RJ)

Geraldo Bonorino Xexéo (PESC/UFRJ)

Rio de Janeiro,

Junho de 2026

DEDICATÓRIA

Dedico esta dissertação à minha família, que me ensinou a sonhar e a persistir.
Aos que caminharam comigo ao longo dessa jornada, celebrando cada avanço e me lembrando, nos momentos de dúvida, que eu era capaz.
Nada disso teria sido possível sozinha.

AGRADECIMENTOS

Agradeço, antes de tudo, a Deus, pela força nos dias difíceis e pela serenidade nos momentos de dúvida. Foi a fé que me ajudou a seguir quando o caminho parecia mais pesado.

Aos meus pais, por todo amor e incentivo ao longo da minha vida. Vocês sempre acreditaram em mim, até quando eu hesitei. Essa conquista também é de vocês.

Aos meus avós, pelo carinho e pelos valores que me acompanham em cada etapa.

Aos meus amigos, que compreenderam minhas ausências e celebraram cada pequena vitória comigo. Vocês tornaram essa jornada mais leve.

Em especial, à minha melhor amiga, Elen, por estar ao meu lado em todos os momentos. Obrigada por me ouvir e por nunca soltar a minha mão, nem nos dias mais difíceis.

Ao Centro Brasileiro de Pesquisas Físicas (CBPF) e à Diretoria de Pesquisa e Pós-Graduação (DIPPG) do CEFET/RJ, pelo suporte computacional e pela infraestrutura utilizada nos experimentos.

Aos membros da banca examinadora, pelas contribuições que enriqueceram este trabalho.

E a todos que, de alguma forma, contribuíram para esta conquista, meu muito obrigada.

RESUMO

Avaliação de modelos de aprendizado de máquina para a correção automática de redações segundo as competências do ENEM

Vanessa Santos Soares

Orientadores:
Eduardo Bezerra da Silva, D.Sc.
Gustavo Paiva Guedes e Silva, D.Sc.

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Com o crescimento do ensino remoto e a aplicação de exames de larga escala como o ENEM, a automatização da correção de textos discursivos tornou-se uma necessidade crescente. Este trabalho investiga diferentes estratégias de aprendizado de máquina para avaliação automática de redações em língua portuguesa, tomando como referência as cinco competências avaliativas do ENEM. Para isso, foram analisadas 9.599 redações coletadas do portal Vestibular Brasil Escola, abrangendo 102 temas publicados entre 2009 e 2024. São comparadas duas linhas de abordagem: (i) métodos tradicionais baseados em features linguísticas extraídas dos textos, e (ii) modelos de linguagem pré-treinados com fine-tuning. O desempenho dos modelos é avaliado por meio da métrica *Quadratic Weighted Kappa*, que mede a concordância com corretores humanos. Espera-se demonstrar que modelos pré-treinados oferecem ganhos expressivos em robustez e confiabilidade, superando abordagens baseadas em engenharia de atributos. O estudo contribui para o avanço da área de Automatic Essay Scoring em português, oferecendo um benchmark e análise comparativa que podem apoiar futuras pesquisas e aplicações educacionais.

Palavras-chave:
Avaliação automática de redações; Aprendizado de máquina; Fine-tuning; Língua portuguesa; ENEM; Redações;

Rio de Janeiro,

Junho de 2026

ABSTRACT

Avaliação de modelos de aprendizado de máquina para a correção automática de redações segundo as competências do ENEM

Vanessa Santos Soares

Advisors:
Eduardo Bezerra da Silva, D.Sc.
Gustavo Paiva Guedes e Silva, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

With the rise of remote education and the implementation of large-scale examinations such as ENEM, the automation of discursive text grading has become an increasing necessity. This work investigates different machine learning strategies for the Automatic Essay Scoring (AES) of Portuguese language texts, based on the five assessment competencies defined by ENEM. To this end, 9,599 essays were analyzed, collected from the Vestibular Brasil Escola portal, covering 102 topics published between 2009 and 2024. Two main approaches are compared: (i) traditional methods based on linguistic features extracted from the texts, and (ii) pre-trained language models with fine-tuning. Model performance is evaluated using the Quadratic Weighted Kappa (QWK) metric, which measures the degree of agreement with human raters. The study aims to demonstrate that pre-trained models provide significant gains in robustness and reliability, outperforming approaches based on feature engineering. This research contributes to the advancement of Automatic Essay Scoring in Portuguese by offering a benchmark and comparative analysis that can support future studies and educational applications.

Keywords:
Automatic Essay Scoring; Machine Learning; Fine-tuning; Portuguese Language; ENEM; Essays;

Rio de Janeiro,

Junho de 2026

Sumário

Declaração de Uso de Ferramentas de IA 1

I Introdução **2**

I.1 Contextualização 2

I.2 Motivação 3

I.3 Objetivos 4

I.4 Metodologia 5

I.5 Organização 6

II Fundamentação Teórica **7**

II.1 Avaliação Automática de Textos e Contexto Educacional 7

II.2 A Estrutura Avaliativa do ENEM 8

II.3 Representação de Texto em Processamento de Linguagem Natural 10

II.3.1 Pré-processamento Textual 10

II.3.2 Representações Clássicas de Texto 11

II.3.3 Representações Distribuídas de Texto 12

II.4 Aprendizado de Máquina para Avaliação de Redações 14

II.4.1 Formulação como Regressão ou Classificação Ordinal 14

II.4.2 Features Manuais e Representações Automáticas 14

II.4.3 Modelos Clássicos em AES 15

II.4.4 Métricas de Desempenho em AES 15

II.5 Modelos de Linguagem e *fine-tuning* 16

II.5.1 Arquitetura *Transformer* 16

II.5.2 Modelo XLM-RoBERTa 17

II.5.3 *Fine-tuning* para tarefas específicas 17

II.5.4 LoRA 17

II.6 Classificação ordinal e função de custo WKLoss 18

II.7 Considerações sobre Balanceamento e Qualidade dos Dados 18

II.8 Síntese e Conexão com os Capítulos Seguintes 19

III Trabalhos Relacionados **21**
III.1 Seleção das publicações 21
III.2 Critérios de inclusão, exclusão e priorização 22
III.2.1 Critérios de inclusão 22
III.2.2 Critérios de exclusão 22
III.2.3 Critérios de priorização 22
III.3 Resultados 23
III.3.1 Publicações selecionadas 24
III.3.2 Comparação 26

IV Metodologia **29**
IV.1 Formulação do Problema 29
IV.2 Base de dados 30
IV.3 Abordagens e arquiteturas utilizadas 37
IV.4 Avaliação e métricas de desempenho 42

V Experimentos **43**
V.1 Corpus de Redações para Treinamento e Avaliação 43
V.2 Configurações experimentais 44
V.3 Tempo de treinamento dos modelos 46
V.4 Plano de Experimentos 48
V.4.1 Modelos Avaliados 49
V.4.2 Procedimentos Experimentais 49
V.4.3 Questões da pesquisa 51
V.5 Resultados e Análises 51
V.5.1 Panorama e desempenho geral das abordagens 52
V.5.2 Efeitos do tamanho do texto e limitação de contexto 55
V.5.3 Comparação com abordagens de referência (*baseline*) 58
V.5.4 Respostas às Questões de Pesquisa 59
V.5.5 Análise de Erros por Matrizes de Confusão. 60
V.5.6 Análise de Casos 63
V.6 Discussão e síntese dos Resultados 66
V.6.1 Comparação com Trabalhos Relacionados 68

VI Conclusão **70**

Referências **73**

A Estrutura dos Dados **80**

B Exemplo de Registro **82**

C Exemplo de Registro **83**

D Resultados detalhados por abordagem e competência **86**

D.1 Competência 1 - Domínio da modalidade escrita formal 87

D.2 Competência 2 - Compreender a proposta e aplicar conceitos das várias áreas de conhecimento para desenvolver o texto dissertativo-argumentativo em prosa 88

D.3 Competência 3 - Selecionar, relacionar, organizar e interpretar informações em defesa de um ponto de vista 89

D.4 Competência 4 - Conhecimento dos mecanismos linguísticos necessários para a construção da argumentação 90

D.5 Competência 5 - Proposta de intervenção com respeito aos direitos humanos 91

E Ferramenta Quintana **92**

E.1 Visão geral da ferramenta 92

E.2 Arquitetura e integração com os modelos 92

E.3 Fluxo de uso e visualização das avaliações 94

E.4 Disponibilização e reprodutibilidade 95

Lista de Figuras

III.1 Fluxo do processo de revisão sistemática segundo a metodologia PRISMA. 24

IV.1 Fluxo de processamento dos dados 34
IV.2 Distribuição das notas das competências e da nota geral 36
IV.3 Fluxo do treinamento e validação das arquiteturas de predição 41

V.1 Evolução da métrica *Quadratic Weighted Kappa* (QWK) na competência C-I 54
V.2 Evolução da métrica QWK na competência C-II 55
V.3 Evolução da métrica QWK na competência C-III 56
V.4 Evolução da métrica QWK na competência C-IV 57
V.5 Evolução da métrica QWK na competência C-V 58
V.6 Matriz de confusão para a Competência I 61
V.7 Matriz de confusão para a Competência II 62
V.8 Matriz de confusão para a Competência III 63
V.9 Matriz de confusão para a Competência IV 64
V.10 Matriz de confusão para a Competência V 65

E.1 Tela de listagem de temas na Quintana. Cada linha corresponde a um tema cadastrado pelo professor, incluindo título, descrição e ações para inserção de novas redações associadas. 93
E.2 Tela de submissão de redações na Quintana. O aluno seleciona o tema, escreve ou cola o texto na área principal e solicita a avaliação automática. 94
E.3 Detalhes de uma redação avaliada na Quintana, com exibição do texto completo e acesso às notas atribuídas em cada competência. 95
E.4 Visualização das notas por competência na Quintana. Cada competência recebe uma pontuação específica, permitindo uma análise detalhada do desempenho do aluno. 96
E.5 Tela de feedback pedagógico na Quintana. Os comentários são organizados por competência e auxiliam o estudante a compreender as razões da nota e como aprimorar sua escrita. 96

E.6 Tela de detalhes de um tema de redação. A partir dessa interface, o professor pode consultar as redações associadas e acompanhar o desempenho da turma. 97

Lista de Tabelas

III.1 Resumo dos trabalhos relacionados 28

IV.1 Características da base de dados 32
IV.2 Conversão de notas em rótulos discretos 35
IV.3 Competências utilizadas pelo ENEM 35

V.1 Configurações gerais de treinamento 44
V.2 Ambiente computacional utilizado nos experimentos 45
V.3 Tempo médio de execução por abordagem e modelo. 47
V.4 Melhores resultados por competência 52
V.5 Respostas às questões de pesquisa 59
V.6 Comparação com trabalhos relacionados em AES para o ENEM 69

D.1 Resultados - competência C1 87
D.2 Resultados - competência C2 88
D.3 Resultados - competência C3 89
D.4 Resultados - competência C4 90
D.5 Resultados - competência C5 91

Lista de Abreviações

ACC	Acurácia	42
AES	*Automatic Essay Scoring*	2, 3, 4, 6, 7, 14, 15, 16, 17, 18, 22, 23, 24, 25, 26, 27, 37, 43, 52, 58, 67
ASAP	*Automated Student Assessment Prize*	3, 15, 23, 25
BOW	*Bag of Words*	11, 12
ENEM	Exame Nacional Do Ensino Médio	4, 6, 7, 8, 9, 14, 17, 18, 19, 23, 26, 29, 30, 70, 71, 72, 80, 86, 92
LORA	*Low-Rank Adaptation*	5, 6, 17, 18, 37, 38, 39, 40, 41, 45, 67, 69, 70, 71
PLMS	*Pre-trained Language Models*	16, 70, 71
PLN	Processamento De Linguagem Natural	10, 11, 12, 16, 24
QWK	*Quadratic Weighted Kappa*	5, 6, 15, 16, 18, 19, 38, 39, 40, 41, 42, 43, 44, 45, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 67, 68, 69, 86, 87, 88, 89, 90, 91
RMSE	Raiz Quadrada Do Erro Médio	5, 6, 15, 40, 41, 42
TF-IDF	*Term Frequency–Inverse Document Frequency*	12, 26, 40, 43, 58, 67, 70
WKLOSS	*Weighted Kappa Loss*	18, 19, 45, 48, 51, 52, 55, 59, 60, 67

Declaração de Uso de Ferramentas de Inteligência Artificial

O autor declara que utilizou ferramentas de Inteligência Artificial generativa exclusivamente como apoio à produção textual desta dissertação, incluindo sugestões de redação, revisão linguística, esclarecimento de conceitos técnicos e organização de ideias. Todas as decisões conceituais, análises, interpretações, escolhas metodológicas, resultados experimentais e conclusões apresentadas são de responsabilidade integral do autor.

Nenhum conteúdo gerado por ferramentas de IA foi incorporado sem revisão crítica. O uso dessas ferramentas ocorreu conforme princípios de integridade acadêmica, transparência e responsabilidade, e não substituiu o trabalho intelectual necessário à elaboração desta pesquisa.

Capítulo I Introdução

I.1 Contextualização

A área de avaliação automática de redações (*Automatic Essay Scoring* (AES)) tem ganhado crescente relevância nas últimas décadas, especialmente diante dos desafios impostos por avaliações educacionais de larga escala. No cenário brasileiro, destaca-se o Exame Nacional do Ensino Médio (ENEM), realizado anualmente por milhões de candidatos e utilizado como principal mecanismo de acesso às universidades públicas federais e a programas de financiamento estudantil. O exame combina questões objetivas de múltipla escolha com uma redação dissertativa-argumentativa, avaliada por professores especializados segundo critérios e competências previamente definidos.

Embora o ENEM seja o caso mais emblemático no contexto brasileiro, a necessidade de avaliação automática de textos dissertativo-argumentativos não se restringe a esse exame. Concursos públicos, vestibulares estaduais e avaliações discursivas em cursos a distância também demandam correções em larga escala, frequentemente enfrentando desafios semelhantes relacionados a custo, tempo e padronização. Assim, o ENEM é adotado neste trabalho como um caso representativo e estruturado, mas as estratégias investigadas possuem potencial de aplicação mais amplo em diferentes contextos avaliativos.

Devido ao elevado número de participantes, a correção manual das redações demanda alto custo financeiro, esforço logístico e tempo considerável. Além disso, esse processo está sujeito a fatores como fadiga dos avaliadores, vieses humanos e inconsistências entre corretores, o que compromete a padronização e a confiabilidade das notas atribuídas. Soma-se a isso o fato de que, para alcançar um bom desempenho, os estudantes precisam praticar continuamente a escrita e receber um retorno qualificado de especialistas. Contudo, a oferta de correções detalhadas e individualizadas em larga escala é inviável. Nesse contexto, ferramentas de correção automática apresentam-se como alternativas promissoras, capazes de estabilizar critérios de avaliação, fornecer um retorno imediato aos alunos e reduzir a carga de trabalho docente. Contudo, modelos treinados em padrões textuais dominantes podem penalizar variações legítimas de estilo e contribuir para uma pasteurização da escrita [Kearney et al., 2025; Ramirez, 2025].

Considerando que o ENEM mobiliza anualmente milhões de participantes e envolve a contratação e coordenação de milhares de corretores distribuídos nacionalmente, o processo de avaliação

constitui também um desafio de ordem econômica e logística. A necessidade de padronização, dupla correção e auditoria eleva significativamente os custos operacionais associados à correção manual. Nesse contexto, a investigação de mecanismos automatizados de apoio à avaliação não se restringe à busca por eficiência técnica, mas dialoga com a necessidade de maior sustentabilidade operacional, transparência e padronização do processo avaliativo em larga escala.

Além das limitações individuais dos avaliadores, há também desafios relacionados à gestão do processo avaliativo em larga escala. A coordenação de milhares de corretores exige mecanismos rigorosos de padronização, supervisão e auditoria, uma vez que discrepâncias sistemáticas podem comprometer a equidade do processo avaliativo [Shermis and Burstein, 2013]. Nesse sentido, ferramentas de avaliação automática não devem ser compreendidas como substitutas do julgamento humano, mas como instrumentos de apoio à gestão, capazes de auxiliar na detecção de inconsistências, oferecer critérios mais estáveis e contribuir para maior transparência e rastreabilidade das decisões avaliativas.

As primeiras iniciativas de correção automática de textos remontam à década de 1960, com o *Project Essay Grader* (PEG) de Ellis Page, que utilizava métricas como comprimento do texto, características gramaticais e vocabulário para atribuir escores [Page, 1966]. Posteriormente, sistemas como o *Intelligent Essay Assessor* (IEA), baseado em *Latent Semantic Analysis* [Landauer et al., 1998], e o *e-rater*, desenvolvido pelo ETS [Burstein et al., 1998], consolidaram o campo e passaram a ser usados em contextos de avaliação em larga escala.

O avanço recente de técnicas baseadas em redes neurais profundas, modelos de linguagem pré-treinados e processamento de linguagem natural impulsionou o desenvolvimento de modelos mais sofisticados para AES, ampliando sua aplicação em contextos educacionais. Entretanto, observa-se que a maioria desses estudos concentra-se em textos escritos em inglês, utilizando grandes bases de dados públicas, como o *Automated Student Assessment Prize* (ASAP) [Hamner et al., 2012]. Em contraste, ainda há escassez de pesquisas e de *corpora* disponíveis para a língua portuguesa, o que dificulta o desenvolvimento de modelos robustos e aplicáveis ao contexto brasileiro [Amorim and Veloso, 2017].

Apesar desses avanços históricos, persistem lacunas importantes, especialmente quando se trata da língua portuguesa. Dessa forma, investigações voltadas para a avaliação automática de redações em português tornam-se especialmente relevantes, tanto para reduzir a lacuna existente na literatura quanto para explorar soluções alinhadas às especificidades do ENEM.

I.2 Motivação

Apesar dos avanços contínuos na área de avaliação automática de redações (AES), a maioria das pesquisas permanece concentrada em textos escritos em língua inglesa [Mayfield and Black, 2020].

Esse cenário faz com que investigações em outros idiomas, como o português, sejam mais escassas e enfrentem desafios adicionais, especialmente no que se refere à disponibilidade de dados anotados. A construção de pesquisas robustas em AES pressupõe a existência de bases de redações em volume considerável, devidamente avaliadas por especialistas humanos, para possibilitar a replicação da confiabilidade do processo avaliativo [Pack et al., 2024]. Contudo, atualmente são poucas as coleções públicas de redações em português que podem ser utilizadas nesse contexto.

Esse quadro revela uma lacuna significativa, visto que o português, além de ser uma língua de grande riqueza linguística e morfológica, oferece amplo espaço para pesquisa, ainda pouco explorado. Assim, pesquisas em AES voltadas para o português não apenas contribuem para o avanço científico em um idioma de grande representatividade global, mas também podem ter efeitos diretos na melhoria de processos educacionais e avaliativos de alta relevância social, como acontece no contexto do Exame Nacional do Ensino Médio (ENEM).

Diante desse cenário, surge a necessidade de investigar abordagens específicas de modelagem voltadas à realidade linguística e avaliativa do português brasileiro.

I.3 Objetivos

O presente trabalho tem como objetivo investigar e comparar empiricamente diferentes estratégias de modelagem para a avaliação automática de redações em língua portuguesa, considerando a predição das competências avaliativas inspiradas na matriz do ENEM, o desempenho preditivo dos modelos e o custo computacional associado às abordagens testadas.

A investigação foi orientada pelas seguintes questões de pesquisa:

- **QP1**: O aumento do porte dos modelos de linguagem implica, necessariamente, melhor desempenho na avaliação automática de redações em português?
- **QP2**: Modelos especializados em português oferecem melhor equilíbrio entre desempenho preditivo e custo computacional do que modelos multilíngues de porte semelhante ou superior?
- **QP3**: Formulações que incorporam a natureza ordinal das notas melhoram o desempenho em relação a formulações categóricas tradicionais?
- **QP4**: O desempenho dos modelos varia entre as competências avaliativas, indicando diferentes níveis de dificuldade conforme a natureza linguística, discursiva ou argumentativa de cada dimensão?
- **QP5**: A limitação de contexto dos modelos baseados em *transformers* afeta o desempenho da avaliação automática, especialmente quando parte da redação precisa ser truncada?

Essas questões foram definidas em nível conceitual. No Capítulo V, elas são operacionalizadas por meio das escolhas experimentais possíveis no escopo deste trabalho, incluindo os modelos, conjuntos de dados e estratégias de treinamento avaliados.

Os objetivos específicos são:

- Construir, estruturar e documentar um corpus de redações em língua portuguesa com notas por competência, respeitando as restrições éticas e de redistribuição associadas à origem dos textos;
- Comparar modelos de linguagem pré-treinados monolíngues e multilíngues na tarefa de predição das cinco competências avaliativas;
- Avaliar abordagens de modelagem por competência, multi-saída, em cascata e com formulação ordinal, analisando seus efeitos sobre o desempenho por competência;
- Investigar o uso de técnicas de adaptação eficiente de parâmetros, em especial *Low-Rank Adaptation* (LoRA), como estratégia para viabilizar o *fine-tuning* em diferentes modelos e configurações experimentais;
- Analisar o impacto do limite de 512 *tokens* sobre o desempenho, comparando o conjunto completo, submetido a truncamento, com o conjunto filtrado;
- Avaliar as abordagens com métricas utilizadas na literatura de avaliação automática de redações, especialmente QWK, acurácia e Raiz quadrada do erro médio (RMSE), considerando também o custo computacional;
- Desenvolver um protótipo de ferramenta Web para demonstrar a aplicação dos modelos de predição por competência em um contexto educacional.

I.4 Metodologia

Para o desenvolvimento desta pesquisa, utilizamos um *corpus* composto por 9.599 redações, distribuídas em 102 temas distintos, extraídas da plataforma Brasil Escola, na qual cada redação foi revisada por avaliadores especializados. O *corpus* foi estruturado e pré-processado de modo a viabilizar a utilização desses dados no treinamento e avaliação dos modelos propostos.

O levantamento bibliográfico que fundamenta esta pesquisa foi conduzido com base na metodologia PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), reconhecida por sua robustez e rigor na condução de revisões sistemáticas da literatura. Essa abordagem orientou o processo de busca, aplicação de critérios de inclusão e exclusão e seleção dos estudos mais

relevantes, assegurando a abrangência e a consistência da revisão. Os procedimentos específicos, incluindo as strings de busca e os critérios utilizados, são detalhados no Capítulo III.

A partir desse embasamento teórico e da base de dados construída, esta pesquisa investigou diferentes abordagens de modelagem para a avaliação automática de redações em língua portuguesa, tomando como ponto de partida os critérios utilizados no ENEM. Foram conduzidos experimentos com modelos de linguagem pré-treinados e diferentes formulações da tarefa, incluindo classificação ordinal, modelos multi-saída, abordagens em cascata e variações com *fine-tuning* eficiente via LoRA. O objetivo foi comparar empiricamente essas alternativas em termos de concordância com as notas humanas, desempenho por competência, custo computacional e sensibilidade ao limite de contexto dos modelos.

Por fim, os resultados foram avaliados por meio de métricas quantitativas, com destaque para a *Quadratic Weighted Kappa* (QWK), amplamente utilizada na literatura de AES. Essa métrica permitiu mensurar o grau de concordância entre as previsões dos modelos e as avaliações humanas, possibilitando a comparação entre as diferentes abordagens e a interpretação de seus comportamentos em cada competência analisada.

I.5 Organização

Os demais capítulos desta dissertação estão organizados como segue. No Capítulo II, são discutidos os conceitos essenciais de *Automated Essay Scoring* (AES), incluindo métricas de avaliação como o *Quadratic Weighted Kappa* (QWK). O Capítulo III apresenta trabalhos relacionados, tanto internacionais quanto nacionais, com ênfase em abordagens aplicadas à língua portuguesa e ao exame do ENEM. O Capítulo IV descreve o corpus utilizado, os procedimentos de pré-processamento e as estratégias de modelagem propostas, como arquiteturas multi-saída, modelos por competência e classificadores ordinais em cascata. No Capítulo V, são apresentados os experimentos realizados, acompanhados da análise dos resultados obtidos em termos de QWK, RMSE e acurácia. Por fim, o Capítulo VI traz as conclusões do trabalho, destacando contribuições, limitações e sugestões para pesquisas futuras.

Além dos capítulos, esta dissertação inclui quatro apêndices que complementam a apresentação dos resultados e a reprodutibilidade do estudo. O Apêndice A descreve a estrutura completa da base de dados utilizada, enquanto o Apêndice C apresenta um exemplo integral de registro após o pré-processamento. O Apêndice D reúne os resultados experimentais completos por competência e por abordagem, oferecendo transparência às análises discutidas no Capítulo V. Por fim, o Apêndice E documenta a ferramenta *Quintana*, desenvolvida no âmbito desta pesquisa e contemplada no edital CAPES/InovaEducação, descrevendo sua arquitetura, módulos, fluxos de interação e a integração com os modelos de avaliação propostos.

Capítulo II Fundamentação Teórica

II.1 Avaliação Automática de Textos e Contexto Educacional

A avaliação automática de redações, ou Automated Essay Scoring (AES), busca aproximar o julgamento computacional daquele realizado por avaliadores humanos [Mayfield and Black, 2020]. A área se consolidou ainda na década de 1960, com o *Project Essay Grader* (PEG) [Page, 1966], e evoluiu com sistemas que incorporaram análise semântica e modelos estatísticos, como o *Intelligent Essay Assessor* (IEA), baseado em *Latent Semantic Analysis* [Foltz et al., 1999], e o e-rater, amplamente utilizado pelo ETS [Burstein et al., 1998]. Esses trabalhos definiram os primeiros referenciais para AES e influenciaram diversas pesquisas posteriores, em especial para a língua inglesa [Hamner et al., 2012].

Avaliar textos dissertativo-argumentativos, no entanto, é uma tarefa complexa. A produção escrita envolve organização global do texto, coerência, argumentação e domínio da norma padrão, aspectos difíceis de capturar por meio de sinais superficiais. Ao mesmo tempo, o julgamento humano apresenta limitações conhecidas, como subjetividade, divergências interpretativas e queda de consistência em função da fadiga, principalmente quando o volume de redações é elevado [Dikli, 2006]. Essas instabilidades afetam diretamente a confiabilidade das notas e tornam a tarefa especialmente desafiadora em contextos de larga escala.

Nesse cenário, modelos automáticos têm sido investigados como alternativas para apoiar o trabalho dos avaliadores e reduzir inconsistências. Estudos mostram que esses sistemas conseguem aplicar critérios de forma mais regular e processar grandes quantidades de textos em menos tempo [Burstein et al., 1998; Lim et al., 2021]. Ainda assim, a adoção de AES não é trivial: os modelos precisam representar fenômenos discursivos complexos, lidar com vieses presentes nos dados de treinamento e se adaptar a diferentes perfis sociolinguísticos.

O contexto brasileiro reforça essas dificuldades. O ENEM mobiliza milhões de candidatos por ano e apresenta grande diversidade de estilos de escrita, temas e repertórios socioculturais. Além disso, sua matriz avaliativa distribui a nota em cinco competências distintas, o que exige que os modelos considerem múltiplas dimensões da produção textual. Pesquisas nacionais já exploraram a predição de notas do exame com técnicas tradicionais de aprendizado de máquina e atributos linguísticos [Amorim and Veloso, 2017], destacando a importância de compreender como cada

dimensão avaliativa se manifesta no texto e influencia o desempenho dos modelos.

O aumento do uso de plataformas educacionais e a necessidade de oferecer feedback rápido também estimulam a busca por soluções automáticas no país. Nessas aplicações, a proposta não é substituir o avaliador humano, mas tornar o processo mais escalável, consistente e acessível. Para isso, é essencial desenvolver modelos robustos ao português brasileiro, sensíveis à variação linguística e capazes de representar nuances discursivas relevantes para a avaliação.

Dessa forma, a avaliação automática de textos se apresenta como um campo que integra educação, linguística e processamento de linguagem natural. Este trabalho se insere nesse contexto investigando diferentes estratégias de modelagem e analisando sua adequação às demandas do ENEM, com foco em compreender seus limites e potencialidades em cenários de grande escala.

II.2 A Estrutura Avaliativa do ENEM

A redação do Exame Nacional do Ensino Médio (ENEM) é avaliada com base em uma matriz composta por cinco competências, cada uma pontuada entre 0 e 200 pontos, totalizando até 1000 pontos. Segundo INEP [2025], essa matriz busca orientar a produção textual dos participantes e oferecer critérios claros e consistentes para a avaliação realizada por múltiplos corretores. Por combinar aspectos linguísticos, discursivos e argumentativos, essa estrutura tornou-se referência para pesquisas em avaliação automática de textos em português, servindo como guia para o desenvolvimento de abordagens de classificação por competências e predição da nota total.

Com base no que está definido na cartilha oficial de avaliação do ENEM, a seguir descrevem-se as cinco competências avaliadas.

- **Competência I - Domínio da modalidade escrita formal da língua portuguesa (C-I).** Avalia o uso adequado da modalidade escrita formal, considerando ortografia, concordância, pontuação e estruturação sintática. A Cartilha do ENEM destaca que desvios frequentes ou que prejudiquem a compreensão reduzem significativamente o desempenho do participante [INEP, 2025].

- **Competência II - Compreender a proposta e aplicar conceitos das várias áreas do conhecimento para desenvolver o texto dissertativo-argumentativo em prosa (C-II).** Verifica se o participante compreende a proposta de redação e desenvolve uma argumentação consistente e pertinente ao tema. A cartilha enfatiza que a mobilização de repertórios socioculturais deve ser contextualizada, evitando abordagens genéricas ou memorizações prévias [INEP, 2025]. Estudos de avaliação automática, como o de Amorim and Veloso [2017], mostram que essa competência depende fortemente da capacidade do modelo de capturar relações semânticas globais e alinhamento temático.

- **Competência III - Selecionar, relacionar, organizar e interpretar informações em defesa de um ponto de vista (C-III).** Avalia a coerência e a progressão argumentativa do texto. Conforme a cartilha, essa competência considera como o participante seleciona e articula informações relevantes para sustentar seu ponto de vista [INEP, 2025].

- **Competência IV - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação (C-IV).** Examina o uso adequado de operadores argumentativos, conectores, pronomes e demais recursos de coesão que garantem fluidez e continuidade entre as partes do texto. A cartilha destaca que o uso artificial, mecânico ou excessivamente repetitivo desses elementos compromete a clareza do texto [INEP, 2025].

- **Competência V - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos (C-V).** Avalia a capacidade do participante de formular uma proposta de intervenção completa, articulada ao desenvolvimento do texto e composta por agente, ação, modo/meio, finalidade e detalhamento. A proposta deve obrigatoriamente respeitar os direitos humanos, sob pena de atribuição de nota zero nessa competência [INEP, 2025].

A redação do ENEM caracteriza-se como um texto dissertativo-argumentativo altamente estruturado, cuja organização segue um padrão relativamente consolidado ao longo dos anos. Embora o INEP [2025] não imponha formalmente um número fixo de parágrafos, observa-se ao longo dos anos a consolidação de uma estrutura prototípica composta por quatro parágrafos: introdução com apresentação de tese; dois parágrafos de desenvolvimento argumentativo; e conclusão com proposta de intervenção estruturada. Essa regularidade cria padrões discursivos relativamente estáveis, o que tem implicações diretas para sistemas de avaliação automática, que podem capturar indícios estruturais associados à organização textual.

Essa regularidade estrutural contribuiu para a consolidação de modelos recorrentes de organização textual e introduz padrões discursivos que podem ser explorados por sistemas de avaliação automática de textos, especialmente aqueles sensíveis à estrutura global e à organização argumentativa [Shermis and Burstein, 2013]).

A estrutura avaliativa do ENEM tem implicações diretas para o desenvolvimento de modelos de avaliação automática, pois fornece critérios explícitos que podem ser operacionalizados como tarefas individuais ou combinadas. Estudos como o de Amorim and Veloso [2017] mostram que diferentes dimensões do texto podem ser modeladas de forma independente, refletindo exatamente a granularidade oferecida pela matriz oficial. Assim, as competências funcionam não apenas como critérios de julgamento humano, mas também como um referencial que orienta a formulação de

abordagens de aprendizagem de máquina voltadas para a predição da nota total e das avaliações específicas.

II.3 Representação de Texto em Processamento de Linguagem Natural

II.3.1 Pré-processamento Textual

O pré-processamento textual é o ponto de partida de qualquer sistema de Processamento de Linguagem Natural (PLN). É nessa etapa que o texto bruto ganha uma forma mais estável, permitindo que os modelos identifiquem padrões linguísticos com maior precisão. Mesmo com o avanço de arquiteturas profundas capazes de capturar dependências complexas, esse passo continua fundamental: quanto mais limpo e coerente o corpus, mais confiáveis tendem a ser as representações aprendidas.

Modelos clássicos de representações distribuídas já reforçavam essa dependência. Abordagens como *Word2Vec* [Mikolov et al., 2013] e *GloVe* [Pennington et al., 2014] exigem um vocabulário bem estruturado para gerar vetores consistentes. Mesmo métodos mais modernos, como *FastText* [Bojanowski et al., 2017] e ELMo [Peters et al., 2018], que lidam melhor com morfologia e contexto, continuam sensíveis à qualidade e à consistência do texto de entrada.

Tokenização e Tratamento do Vocabulário

A tokenização converte o texto contínuo em unidades manipuláveis (palavras, subpalavras ou caracteres) e funciona como a porta de entrada para todo o *pipeline* de PLN. A forma como essa segmentação é feita determina o vocabulário disponível e, em consequência, o tipo de regularidades linguísticas que o modelo será capaz de aprender.

Modelos baseados exclusivamente em palavras, tais como Word2Vec [Mikolov et al., 2013] e GloVe [Pennington et al., 2014], dependem de uma segmentação consistente: pequenas variações ortográficas ou erros de digitação tendem a inflar artificialmente o vocabulário e aumentar a esparsidade, especialmente em domínios nos quais palavras raras são comuns, como textos estudantis, mídias sociais ou línguas morfologicamente ricas, caso do português.

Para lidar com essas limitações, surgiram abordagens que incorporam unidades sublexicais. O FastText [Bojanowski et al., 2017], por exemplo, representa cada palavra como a soma de seus n-grams, permitindo capturar regularidades internas mesmo quando a ocorrência de uma palavra no corpus é mínima. Assim, variações morfológicas deixam de ser tratadas como tokens completamente distintos, e palavras raras passam a ter representações mais informativas.

Modelos contextualizados expandem ainda mais essa estratégia ao adotarem tokenizadores como *WordPiece* e *SentencePiece*, que segmentam termos em subpalavras frequentes. Essa aborda-

gem, presente em arquiteturas como ELMo [Peters et al., 2018] e posteriormente nos Transformers [Vaswani et al., 2017], reduz drasticamente a esparsidade do vocabulário sem sacrificar nuances linguísticas essenciais.

Em suma, tokenizar é decidir o nível de granularidade com que o modelo enxerga a língua. Em idiomas com grande variação morfológica, estratégias baseadas em fragmentos lexicais são particularmente valiosas para lidar com flexões, erros ortográficos e palavras raras.

Normalização e limpeza

A normalização complementa a tokenização ao reduzir ruídos e padronizar elementos que variam demais no texto. Entre as práticas mais comuns estão a substituição de URLs e menções por tokens especiais, a normalização de caixa e a remoção de caracteres problemáticos.

Embora simples, essas operações influenciam diretamente o comportamento dos modelos. Trabalhos como *FastText* [Bojanowski et al., 2017] e ELMo [Peters et al., 2018] mostram que normalizações excessivas podem eliminar pistas morfológicas importantes, especialmente em tarefas que dependem da forma original das palavras. Assim, o nível de limpeza deve ser escolhido com cuidado, equilibrando redução de ruído e preservação de informações linguísticas relevantes.

II.3.2 Representações Clássicas de Texto

Antes da popularização dos modelos distribuídos e contextualizados, a maioria das aplicações em PLN dependia de representações baseadas em contagem, que descrevem os textos a partir da frequência com que palavras ocorrem em cada documento. Embora simples, essas abordagens foram fundamentais para tarefas como classificação, recuperação da informação e análise exploratória de textos.

Bag of Words (BoW)

O modelo *Bag of Words* (BoW) é uma das representações mais tradicionais e diretas: cada documento é tratado como um conjunto de palavras, ignorando completamente a ordem em que aparecem. Cada termo do vocabulário corresponde a uma dimensão, e o documento passa a ser representado por um vetor que indica quantas vezes cada palavra ocorre [Harris, 1954].

Essa simplicidade tornou o BoW popular em tarefas de classificação textual e sistemas de busca nas décadas anteriores aos modelos neurais. No entanto, sua própria natureza limita a capacidade de capturar relações semânticas ou sintáticas entre palavras.

Term Frequency-Inverse Document Frequency (TF-IDF)

O *Term Frequency–Inverse Document Frequency* (TF-IDF) [Spärck Jones, 1972] é uma extensão do modelo Bag of Words que busca destacar termos mais informativos no corpus. A ponderação combina a frequência do termo no documento com sua raridade global, conforme a Equação II.1. Nessa equação, tf(t,d) indica a frequência do termo t no documento d, e $\text{idf}(t) = \log\left(\frac{N}{\text{df}(t)}\right)$, sendo N o número total de documentos e df(t) a quantidade de documentos que contêm o termo.

$$\text{tfidf}(t,d) = \text{tf}(t,d) \times \text{idf}(t) \tag{II.1}$$

Por equilibrar termos frequentes e discriminativos, o TF-IDF permanece uma representação eficiente e amplamente empregada em tarefas de classificação e recuperação textual.

Limitações das representações baseadas em contagem

Apesar de amplamente utilizadas, abordagens como BoW e TF-IDF apresentam limitações importantes. Ambas ignoram completamente a ordem e o contexto das palavras, o que impede a captura de relações sintáticas e de significado. Além disso, produzem vetores de alta dimensionalidade e extremamente esparsos, dificultando a generalização e tornando termos semanticamente relacionados indistinguíveis [Harris, 1954; Mikolov et al., 2013]. Essas restrições motivaram o surgimento de representações distribuídas mais compactas e expressivas, capazes de modelar semântica e similaridade de maneira muito mais eficaz.

II.3.3 Representações Distribuídas de Texto

A evolução das representações textuais em PLN marcou uma mudança profunda na forma como modelos computacionais capturam significado. Em vez de tratar palavras como unidades independentes, como faziam BoW e TF-IDF, as representações distribuídas definem cada termo como um vetor denso em um espaço contínuo, permitindo modelar relações semânticas e sintáticas de maneira muito mais expressiva. Essa abordagem se apoia diretamente na hipótese distribucional de que "palavras que ocorrem em contextos semelhantes tendem a ter significados semelhantes" [Harris, 1954].

Embeddings de Palavras com Vetores Fixos

Os primeiros avanços relevantes em representações distribuídas vieram dos *embeddings* com vetores fixos, nos quais cada palavra recebe um único vetor independente do contexto. O *Word2Vec* [Mikolov et al., 2013] inaugurou essa abordagem por meio dos modelos *Skip-gram* e CBOW, capazes de capturar regularidades semânticas a partir de padrões de coocorrência. Em seguida, o GloVe [Pen-

nington et al., 2014] combinou estatísticas locais e globais do corpus para produzir vetores mais estáveis e informativos. Já o FastText [Bojanowski et al., 2017] ampliou essas ideias ao representar palavras como composições de n-grams, o que melhora significativamente o tratamento de morfologia, variações ortográficas e palavras raras, que são desafios particularmente presentes no português. Apesar da eficácia, esses métodos compartilham uma limitação estrutural: atribuem uma única representação para cada palavra, impossibilitando distinguir sentidos diferentes em contextos distintos.

Embeddings Contextuais e Modelos Multilíngues

Os *embeddings* contextuais surgiram para superar a principal limitação dos vetores fixos: a incapacidade de representar sentidos distintos de uma mesma palavra [Ilić et al., 2018]. O ELMo [Peters et al., 2018] inaugurou essa mudança ao gerar vetores dependentes do contexto por meio de camadas profundas de BiLSTMs, permitindo que cada ocorrência lexical recebesse uma representação própria [Ilić et al., 2018]. A transição se consolidou com a arquitetura Transformer [Vaswani et al., 2017], cujos mecanismos de atenção capturam dependências globais entre tokens, dando origem a modelos como BERT e suas variantes, capazes de produzir representações altamente expressivas e sensíveis à estrutura sintática e discursiva do texto.

No cenário multilíngue, modelos como o XLM-RoBERTa [Conneau et al., 2019] ampliam essas capacidades ao serem pré-treinados em larga escala em mais de cem idiomas. A literatura demonstra que abordagens baseadas em fragmentos lexicais são particularmente eficazes para línguas morfologicamente ricas, a exemplo do tcheco e do russo; por analogia, o português também se beneficia dessa estratégia, especialmente em casos de flexões verbais, concordância e formação de derivados [Bojanowski et al., 2017]. Essa modelagem sublexical melhora a representação de palavras raras e reduz substancialmente o problema de termos fora do vocabulário [Bojanowski et al., 2017].

Além disso, estudos com modelos de grande porte mostram que o pré-treinamento massivo tende a melhorar o desempenho mesmo em tarefas com poucos exemplos rotulados, característica essencial em cenários de baixo recurso [Devlin et al., 2019]. Modelos como BERT demonstraram que aumentar a escala do treinamento (em dados ou parâmetros) produz ganhos consistentes em tarefas com supervisionamento limitado, o que favorece aplicações em português que contam com menos dados anotados. Outro ponto importante é que esses modelos eliminam a necessidade de arquiteturas altamente especializadas para cada tarefa: o *fine-tuning* de modelos multilíngues geralmente requer apenas uma camada adicional e poucos parâmetros treináveis, simplificando a adaptação para diferentes problemas em português [Devlin et al., 2019].

Em conjunto, essas características (robustez morfológica, melhor desempenho em cenários de

poucos dados e facilidade de adaptação) tornam os modelos multilíngues pré-treinados especialmente adequados para aplicações na língua portuguesa. Por isso, arquiteturas como o XLM-RoBERTa vêm se consolidando como escolhas eficientes e de alto desempenho em sistemas de avaliação automática de textos, incluindo aqueles desenvolvidos neste trabalho.

II.4 Aprendizado de Máquina para Avaliação de Redações

A tarefa de avaliação automática de textos tem sido tradicionalmente formulada como um problema supervisionado, no qual o sistema aprende a mapear redações para notas produzidas por avaliadores humanos. Essa perspectiva acompanha a própria origem da área, ainda na década de 1960, quando Page [1966] propôs o *Project Essay Grader*. Desde então, diferentes abordagens supervisionadas foram exploradas, indo de modelos lineares e baseados em features linguísticas até métodos mais recentes fundamentados em representações distribuídas e arquiteturas neurais profundas.

II.4.1 Formulação como Regressão ou Classificação Ordinal

Uma redação pode ser tratada como uma entrada textual X, associada a uma nota Y que pode assumir valores contínuos ou inteiros em uma escala ordinal. Em muitas avaliações educacionais, como no próprio ENEM, as notas são inteiros organizados em uma escala progressiva; por isso, a modelagem como regressão é bastante comum [Yannakoudakis et al., 2011]. No entanto, observa-se na literatura um crescimento no uso de modelos de classificação ordinal, especialmente em tarefas onde os rótulos seguem uma estrutura ordenada e a distância entre as classes possui significado. Trabalhos voltados para AES apontam que métodos puramente classificatórios não exploram essa estrutura e tratam erros próximos e distantes de forma equivalente, o que acaba por reduzir a sensibilidade da modelagem [Phandi et al., 2015a; He and Garcia, 2009]. Em contrapartida, estudos mais recentes mostram que abordagens ordinais conseguem incorporar explicitamente essa hierarquia entre categorias, resultando em predições mais consistentes com a gradação de qualidade textual esperada em contextos avaliativos [Uto, 2020].

II.4.2 Features Manuais e Representações Automáticas

As primeiras gerações de sistemas de AES baseavam-se em um conjunto manual de indicadores linguísticos, tais como extensão do texto, contagem de erros gramaticais, diversidade lexical e medidas de coerência extraídas por técnicas de análise semântica. Exemplos desse paradigma incluem o Intelligent Essay Assessor (IEA), fundamentado em Latent Semantic Analysis [Foltz et al., 1999], e o e-rater, que combina métricas linguísticas com modelos estatísticos [Attali and Burstein, 2006].

Apesar de úteis, essas features demandam conhecimento linguístico especializado e frequentemente apresentam limitações. Com o avanço dos modelos baseados em representações distribuídas, esse cenário mudou significativamente. Vetores densos como Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014] e FastText [Bojanowski et al., 2017] passaram a capturar automaticamente relações semânticas e morfológicas, dispensando parte da engenharia manual de atributos. Essa transição foi crucial para AES, pois essas representações reduzem a necessidade de engenharia de atributos e permitem que o modelo aprenda padrões discursivos mais profundos.

II.4.3 Modelos Clássicos em AES

Antes da adoção generalizada de redes neurais, modelos tradicionais dominaram o campo. Técnicas como regressão linear, máquinas de vetores de suporte [Cortes and Vapnik, 1995] e Random Forests [Breiman, 2001] foram amplamente empregadas em diferentes corpora educacionais. Esses modelos se beneficiavam diretamente da qualidade dos atributos linguísticos manuais e, embora simples, alcançaram resultados competitivos em diversos cenários.

No contexto específico da avaliação ordinal, a regressão ordinal de McCullagh [1980] ofereceu uma formulação estatística robusta para lidar com rótulos ordenados, influenciando abordagens posteriores utilizadas em AES. Trabalhos como Phandi et al. [2015b] exploram modelos híbridos para melhorar a previsibilidade em escalas ordinais, antecipando tendências que seriam refinadas por arquiteturas neurais nos anos seguintes.

II.4.4 Métricas de Desempenho em AES

Como AES envolve notas contínuas ou ordinais, as métricas utilizadas devem capturar não apenas acertos exatos, mas também o grau de discrepância entre o sistema e os avaliadores humanos. Entre as métricas mais comuns estão o RMSE, a correlação de Pearson e o F1-score, cada uma destacando aspectos específicos do comportamento do modelo.

No entanto, a métrica predominante na área é a QWK, introduzida por Cohen [Cohen, 1968] e consolidada como padrão a partir de competições como a ASAP em 2012 [asa, 2012]. A QWK mede o grau de concordância entre duas avaliações ponderando os erros pela distância entre as categorias, o que a torna especialmente adequada para notas ordinais.

A acurácia, apesar de intuitiva, trata todos os erros da mesma forma. Errar uma redação que deveria receber nota 120, atribuindo-lhe 100 ou 0, resulta em penalidades idênticas. A QWK, por outro lado, penaliza mais severamente erros maiores e recompensa previsões próximas da nota correta, refletindo melhor o comportamento esperado de avaliadores humanos. Trabalhos como Cicchetti and Feinstein [1990] discutem detalhadamente essas diferenças, mostrando que métricas simples podem mascarar discordâncias estruturais.

Em avaliações educacionais baseadas em escalas ordinais, nem todos os erros possuem o mesmo impacto. Prever uma nota ligeiramente diferente da esperada representa discrepância menor do que atribuir um valor drasticamente distante. A QWK incorpora essa noção ao ponderar os erros de acordo com a distância entre as categorias, sendo amplamente utilizada em contextos de AES por refletir melhor a concordância entre avaliadores humanos e sistemas automáticos [Cohen, 1968].

A escolha da QWK neste trabalho alinha-se a essa tradição metodológica e à necessidade de considerar a natureza ordinal das notas atribuídas no ENEM.

No contexto específico de AES, Napoles et al. [2016a] reforçam que métricas sensíveis à ordem e à magnitude dos erros são essenciais para reproduzir a consistência entre avaliadores humanos. Isso explica por que estudos recentes em avaliação automática, tanto clássicos quanto neurais, utilizam a QWK como principal métrica de comparação.

II.5 Modelos de Linguagem e *fine-tuning*

Os avanços recentes em PLN têm sido impulsionados pelo uso de modelos de linguagem pré-treinados (*Pre-trained Language Models* (PLMs)), que se tornaram fundamentais em tarefas que exigem interpretação contextual de textos, como a avaliação automática de redações. Esses modelos são treinados previamente em grandes volumes de dados textuais e aprendem representações linguísticas profundas capazes de capturar relações sintáticas, semânticas e discursivas de forma mais precisa do que técnicas tradicionais. Essa etapa de pré-treinamento permite que o modelo seja posteriormente adaptado para tarefas específicas com uma quantidade muito menor de dados rotulados, alcançando resultados superiores em cenários com recursos limitados [Devlin et al., 2019].

II.5.1 Arquitetura *Transformer*

A maioria dos PLMs atuais se baseia na arquitetura *Transformer*, proposta por Vaswani et al. [2017]. Diferentemente de arquiteturas recorrentes, o *Transformer* opera inteiramente por mecanismos de autoatenção, que comparam cada token do texto com todos os outros para modelar dependências de curto e longo alcance, e em diferentes níveis de abstração. Essa característica torna o processamento paralelo mais eficiente e permite ao modelo capturar relações de contexto de maneira mais flexível.

A arquitetura é composta por blocos empilhados de encoders e/ou decoders. Nos modelos voltados para classificação, como BERT e XLM-R, utiliza-se apenas a pilha de encoders, cuja função é transformar a sequência de tokens em representações contextualizadas. Outro ponto central é a contextualização bidirecional: durante o pré-treinamento, o modelo aprende a prever partes mascaradas da frase considerando simultaneamente o contexto anterior e posterior, o que melhora de forma expressiva a compreensão semântica de textos longos e complexos [Devlin et al.,

2019].

II.5.2 Modelo XLM-RoBERTa

Entre os modelos amplamente utilizados em tarefas envolvendo textos em português, destaca-se o XLM-RoBERTa (XLM-R), proposto por Conneau et al. [2019]. Trata-se de uma extensão multilíngue da família RoBERTa [Liu et al., 2019], treinada em 2,5 TB de dados cobrindo mais de 100 idiomas. A estratégia de pré-treinamento inclui masked language modeling (MLM) em larga escala, com vocabulário baseado em sentence piece e unidades sublexicais, permitindo lidar melhor com morfologia rica e variações lexicais comuns ao português.

O XLM-R apresenta capacidade robusta de transferência entre idiomas, o que o torna especialmente útil em cenários com escassez de dados anotados em português, como nos corpora de redações do ENEM. Estudos recentes em AES mostram que modelos dessa família tendem a capturar padrões discursivos, estruturais e de coesão textual de forma mais consistente, mesmo quando treinados com conjuntos relativamente pequenos de exemplos rotulados [Mayfield and Black, 2020].

II.5.3 *Fine-tuning* para tarefas específicas

Após o pré-treinamento, os modelos são ajustados por *fine-tuning* para tarefas supervisionadas específicas, como classificação ordinal, regressão de notas ou classificação por competências. Nessa etapa, adiciona-se uma camada final que transforma as representações contextualizadas em predições. O treinamento envolve a atualização dos pesos do modelo com base em uma função de perda apropriada e em hiperparâmetros como taxa de aprendizado, batch size, número de épocas e estratégias de regularização.

O processo de *fine-tuning* é sensível à disponibilidade de dados e ao custo computacional. Como mostrou Mayfield and Black [2020], pequenas variações nesses hiperparâmetros podem resultar em ganhos significativos em tarefas de AES, especialmente quando o objetivo é minimizar a variabilidade das previsões e aumentar a consistência com avaliadores humanos.

II.5.4 LoRA

Apesar de seu excelente desempenho, modelos como BERT e XLM-R possuem alta complexidade computacional, tornando o *fine-tuning* completo caro e, em muitos casos, inviável em ambientes com recursos limitados. A técnica LoRA, introduzida por Hu et al. [2021], surge como uma solução eficiente ao permitir a adaptação de modelos grandes sem a necessidade de atualizar todos os seus parâmetros.

A ideia central é decompor as atualizações aprendidas em matrizes de baixa dimensão. Isso reduz drasticamente o número de parâmetros treináveis e o consumo de memória durante o treina-

mento. Na prática, a LoRA insere pequenas projeções treináveis nas camadas de atenção enquanto mantém congelados os pesos originais do modelo. Essa abordagem permite treinar modelos grandes em GPUs com memória reduzida, sem perdas significativas de desempenho, o que a torna particularmente relevante para experimentos em AES com múltiplas abordagens, como classificação ordinal, regressão e modelos multi-saída.

II.6 Classificação ordinal e função de custo WKLoss

Muitas tarefas de avaliação automática de redações utilizam escalas ordinais, como as notas de 0 a 200 atribuídas às competências do ENEM. Em cenários desse tipo, funções de perda tradicionais de classificação tratam todas as confusões da mesma maneira, ignorando que errar por um ponto é menos grave do que errar por cem. Para lidar com essa limitação, diferentes trabalhos têm investigado funções de perda derivadas da métrica QWK, que incorporam explicitamente a estrutura ordinal das classes [de La Torre et al., 2018].

A *Weighted Kappa Loss* (WKLoss) segue esse princípio ao transformar a QWK em uma função diferenciável, penalizando as predições conforme a distância entre o valor previsto e o valor verdadeiro [Uto, 2020]. Sua formulação geral é apresentada na Equação II.2.

$$\mathcal{L}_{WK} = \frac{\sum_{i,j} W_{ij}\, O_{ij}}{\sum_{i,j} W_{ij}\, E_{ij}}, \tag{II.2}$$

Na equação acima, O representa a matriz de confusão observada, E a matriz de confusão esperada, e W a matriz de penalização quadrática. Esta última é definida pela Equação II.3.

$$W_{ij} = \frac{(i-j)^2}{(K-1)^2}, \tag{II.3}$$

Nessa equação, K é o número total de classes. Essa estrutura garante que discrepâncias pequenas recebam penalizações leves, enquanto erros mais amplos geram punições proporcionalmente maiores. Essa característica é particularmente importante em avaliações educacionais, nas quais a proximidade entre notas carrega significado pedagógico e impacta diretamente a coerência das predições. Ao alinhar o processo de otimização à métrica empregada na avaliação final, a WKLoss tende a produzir modelos mais estáveis, sensíveis à progressão ordinal das notas e mais consistentes com o comportamento esperado de avaliadores humanos.

II.7 Considerações sobre Balanceamento e Qualidade dos Dados

A distribuição das notas em bases reais de redações costuma ser assimétrica, com forte concentração nas faixas intermediárias e poucas ocorrências de notas muito baixas ou muito altas. Esse desbalanceamento afeta diretamente o processo de aprendizagem: modelos supervisionados tendem

a aproximar predições do centro da distribuição e apresentam menor sensibilidade às classes menos representadas [Phandi et al., 2015a; Uto, 2020]. O problema é ainda mais crítico em tarefas ordinais, nas quais erros distantes têm maior impacto em métricas como a QWK [Cohen, 1968].

A literatura apresenta diferentes estratégias para lidar com essa assimetria, como *undersampling*, *oversampling* e *class weighting* [He and Garcia, 2009; Chawla et al., 2002]. Em modelos baseados em *Transformers*, o ajuste de pesos por classe costuma ser preferido por manter a integridade do corpus e evitar redundâncias indesejadas [Devlin et al., 2019; Souza et al., 2020]. Para tarefas ordinais, perdas adaptadas como a WKLoss também têm sido exploradas por incorporarem a distância entre classes ao processo de otimização [Uto, 2020].

Além da distribuição das notas, é necessário considerar o ruído decorrente da própria correção humana. Estudos mostram que divergências entre avaliadores, subjetividade nos critérios e efeitos de fadiga introduzem inconsistências que limitam o desempenho máximo possível dos modelos [Napoles et al., 2016b; Taghipour and Ng, 2016]. Esses efeitos são particularmente relevantes em competências que exigem julgamento discursivo, como coerência e argumentação.

No presente trabalho, não foram aplicadas técnicas de balanceamento, optando-se por preservar a distribuição original das notas. Essa escolha segue trabalhos que mantêm o comportamento realista do corpus e evitam a introdução de vieses artificiais [Phandi et al., 2015a; Amorim and Veloso, 2017]. Ao manter o cenário natural, o modelo é avaliado em condições mais próximas ao uso prático, incluindo o impacto combinado do desbalanceamento e do ruído avaliativo.

II.8 Síntese e Conexão com os Capítulos Seguintes

Os tópicos discutidos ao longo deste capítulo permitem consolidar a base conceitual necessária para compreender os desafios e as decisões metodológicas adotadas nesta pesquisa. Foram apresentados os elementos centrais envolvidos na avaliação automática de redações: a estrutura das competências do ENEM, as particularidades da representação textual, os modelos de aprendizado empregados em tarefas de classificação ou regressão ordinal e as principais métricas utilizadas na área, como a QWK. Também foram discutidas questões práticas relacionadas à qualidade dos dados, ao desbalanceamento das distribuições de notas e ao impacto do ruído nas anotações humanas.

Em conjunto, esses conceitos formam um mapa de dependências: (i) as competências do ENEM definem o objeto de avaliação e orientam quais aspectos linguísticos, discursivos e argumentativos devem ser capturados; (ii) a representação textual estabelece como o conteúdo das redações é transformado em características processáveis; (iii) os modelos de aprendizado operam sobre essas representações, variando de abordagens tradicionais a arquiteturas baseadas em Transformers, com diferentes estratégias para lidar com questões como ruído e assimetria das classes; (iv) e, por fim, as métricas de avaliação determinam como o desempenho é medido, priorizando coerência com a

escala de notas e com o comportamento esperado de avaliadores humanos.

Essa articulação conceitual sustenta diretamente as escolhas apresentadas no Capítulo IV. A metodologia se apoia nos fundamentos estabelecidos aqui para justificar as etapas de pré-processamento, as estratégias de modelagem adotadas, o uso de funções de perda adequadas à escala ordinal e a decisão de manter a distribuição original das notas. Do mesmo modo, os experimentos exploram empiricamente como essas decisões se traduzem em desempenho, analisando o impacto das representações, das arquiteturas e das métricas discutidas nesta fundamentação.

Capítulo III Trabalhos Relacionados

Segundo de Sousa et al. [2024], a avaliação automática de redações vêm ganhando cada vez mais força para língua portuguesa por conta das redações do ENEM, principalmente devido ao alto número de redações escritas todo ano durante a realização do exame. Este trabalho vai ao encontro dessa necessidade ao desenvolver e avaliar abordagens que visam a melhoria da qualidade e consistência da correção de redações, além da economia de custo, trabalho e tempo [Liew and Tan, 2024].

Este capítulo está estruturado em quatro seções principais. A Seção III.1 descreve o processo de seleção das publicações, detalhando a base de dados utilizada, as combinações de termos de busca e a aplicação da metodologia PRISMA para condução da revisão sistemática. A Seção III.2 apresenta os critérios de inclusão, exclusão e priorização adotados, que orientaram a escolha dos estudos analisados. Em seguida, a Seção III.3 expõe os resultados obtidos a partir da aplicação desses critérios, incluindo o fluxograma do processo de triagem, o conjunto final de trabalhos selecionados, além da discussão de suas respectivas contribuições.

III.1 Seleção das publicações

Para realização desta revisão, foi selecionada a base de dados Scopus. A escolha desta deve-se ao tamanho de seu acervo e sua credibilidade. Durante as pesquisas, não houve restrição temporal, e os tipos de documentos buscados ficaram restritos a artigos publicados em revistas ou trabalhos apresentados em conferências. Os termos de busca utilizados foram combinados conforme mostra a lista a seguir:

- (“automated” OR “automatic”) AND “essay” AND “scoring” AND (“brazilian” OR “portuguese”)
- (“automated essay scoring” OR “automatic essay evaluation” OR “essay grading” OR “automatic writing assessment”) AND (“portuguese” or “brazilian”)
- ((“automated essay scoring” OR “text scoring”) AND (“BERT” OR “RoBERTa” OR “transformers” OR “Lora” OR “QWK”))

A seleção dos artigos foi conduzida conforme as diretrizes da metodologia PRISMA [Page et al., 2021]. Além dos artigos selecionados, o processo de busca foi complementado com a técnica *snowballing*, que permitiu a identificação de publicações adicionais.

III.2 Critérios de inclusão, exclusão e priorização

A definição dos critérios de inclusão, exclusão e priorização teve como objetivo garantir rigor metodológico, transparência no processo de seleção e alinhamento com os objetivos desta pesquisa.

III.2.1 Critérios de inclusão

1. Estudos que abordem a tarefa de correção automática de redações (AES) com foco em análise textual, atribuição de notas ou modelagem de aspectos avaliativos.
2. Estudos aplicados a redações em português brasileiro, com interesse especial nos que abordam diretamente o Exame Nacional do Ensino Médio (ENEM).
3. Artigos com texto completo disponível em inglês ou português.
4. Estudos que investiguem múltiplos aspectos de correção, como análise de competências, análise holística, ou atributos linguísticos/discursivos das redações.

III.2.2 Critérios de exclusão

Foram excluídos da análise:

1. Estudos cujo foco não estivesse diretamente relacionado à tarefa de AES.
2. Trabalhos que não apresentem relação direta com o tema de Automatic Essay Scoring (AES).
3. Artigos cuja versão completa não esteja disponível para acesso por meio do Portal de Periódicos da CAPES.

O critério 3 foi adotado para garantir que os estudos analisados pudessem ser lidos integralmente e avaliados de forma consistente. Entretanto, essa escolha também constitui uma limitação da revisão, pois trabalhos potencialmente relevantes podem ter sido excluídos por restrições de acesso ao texto completo, e não necessariamente por baixa aderência ao tema investigado.

III.2.3 Critérios de priorização

1. Pesquisas voltadas para a língua portuguesa, com foco no português do Brasil.
2. Estudos que empregam Large Language Models (LLMs) ou arquiteturas baseadas em BERT.

3. Trabalhos direcionados ao idioma inglês, espanhol ou francês.

4. Demais estudos que, embora relevantes, não se enquadram nas categorias anteriores.

Observa-se que a literatura internacional em AES concentra a maior parte dos avanços metodológicos recentes na língua inglesa, especialmente no que se refere ao uso de modelos pré-treinados, conjuntos de dados amplamente utilizados (como o ASAP) e formulações inovadoras da tarefa, incluindo regressão ordinal e integração de múltiplos objetivos de otimização. Assim, os trabalhos em inglês foram considerados prioritários para fins de comparação metodológica, não por exclusão de outras línguas, mas em função da centralidade desse idioma na produção científica da área.

Adicionalmente, foram realizadas buscas exploratórias utilizando termos equivalentes em espanhol e francês. Entretanto, não foram identificados estudos consolidados que abordassem a avaliação automática de redações dissertativo-argumentativas em larga escala, com modelagens estruturadas por competência, comparáveis ao contexto do ENEM. Tal constatação reforça a relevância de investigações voltadas ao português brasileiro.

III.3 Resultados

A metodologia PRISMA [Page et al., 2021] foi aplicada conforme descrito na Figura III.1, e a pesquisa concluída em 18/10/2025. As publicações resultantes da aplicação dessa metodologia são detalhadas na seção seguinte (Seção III.3.1)

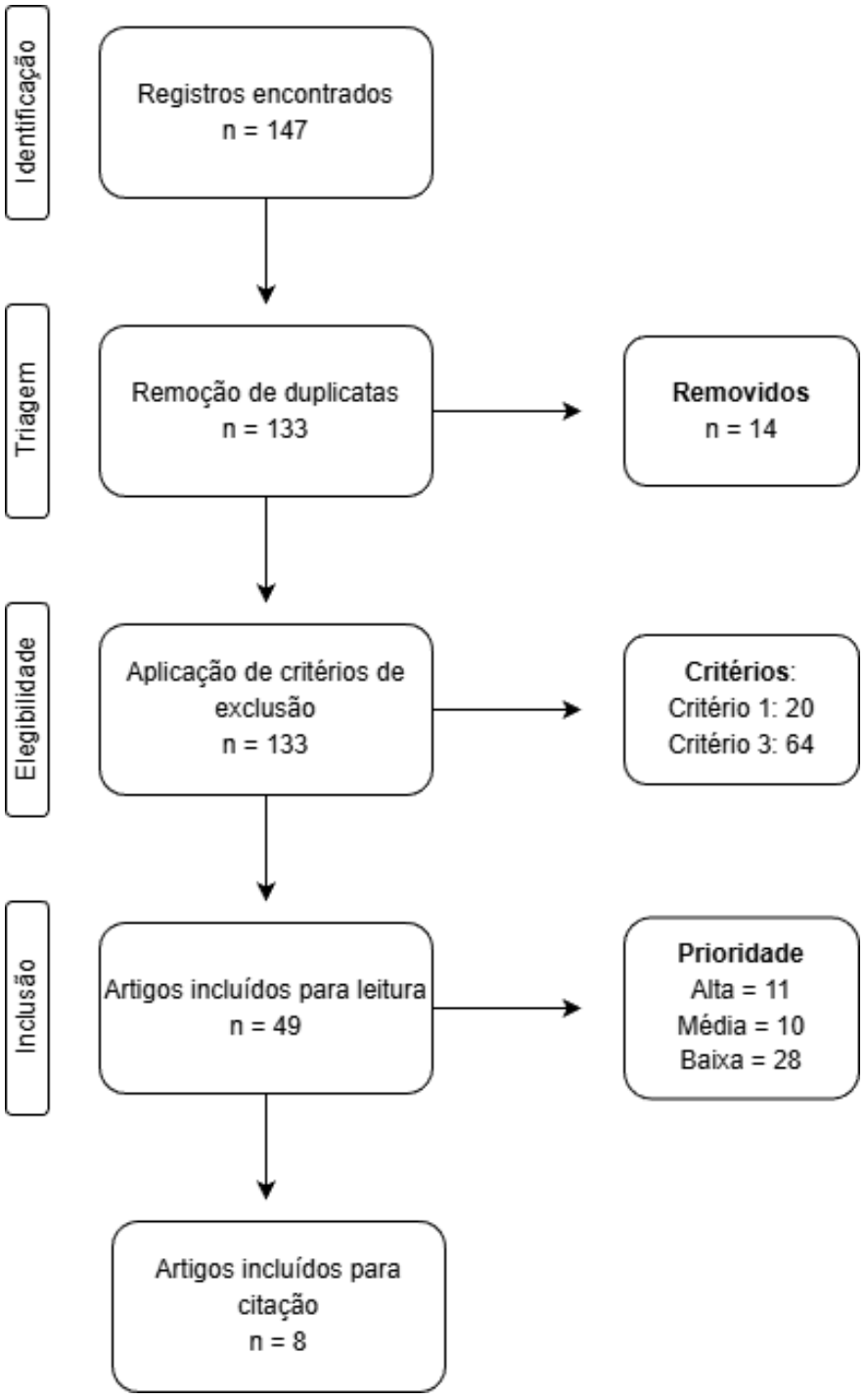


Figura III.1: Representação do fluxo da revisão sistemática realizada conforme a metodologia PRISMA, detalhando as etapas de identificação, triagem, avaliação de elegibilidade e inclusão final dos estudos considerados na revisão.

III.3.1 Publicações selecionadas

Tarefas de AES envolvem a atribuição de notas a redações e são reconhecidas por demandarem tempo e esforço significativos, configurando-se como um desafio para educadores e instituições [Liew and Tan, 2024]. Nesse sentido, conforme apontado por de Sousa et al. [2024], trata-se de uma das áreas educacionais mais relevantes dentro de PLN, cujos estudos têm tradicionalmente recorrido a

técnicas de aprendizado de máquina baseadas em engenharia de atributos.

Trabalhos como Amorim and Veloso [2017]; Haendchen Filho et al. [2018]; Marinho et al. [2022] exemplificam essa vertente, aplicando aprendizado de máquina com engenharia de atributos em conjuntos de redações escritas em língua portuguesa. Os estudos de Amorim and Veloso [2017] e Marinho et al. [2022] buscam definir o conjunto mais adequado de atributos para avaliar cada uma das competências de forma independente, evitando que o desempenho em uma competência influencie as demais. Por sua vez, Haendchen Filho et al. [2018] concentram-se na seleção de atributos voltados à competência II, responsável por avaliar a aderência da redação ao tema proposto, um dos aspectos mais desafiadores para os estudantes por exigir a compreensão plena da proposta de escrita. Além da seleção de atributos, essas abordagens exploram o balanceamento de classes, a formulação do problema como tarefas de classificação e regressão, e o cálculo de similaridade entre o texto produzido e o texto-base fornecido no enunciado.

Embora eficientes, os modelos tradicionais baseados em aprendizado de máquina e engenharia de atributos demandam maior intervenção manual e apresentam alta dependência da qualidade dos atributos definidos. Além disso, os avanços no desenvolvimento de redes neurais mudaram significativamente a área, contribuindo bastante na melhoria da acurácia de generalização dos modelos [Liew and Tan, 2024].

Com o avanço das arquiteturas neurais e dos modelos pré-treinados, estudos como Ribeiro et al. [2024]; de Sousa et al. [2024] passaram a investigar o uso de *fine-tuning* em modelos de linguagem em português, como o BERTimbau [Souza et al., 2020] e o Albertina [Santos et al., 2024], para desenvolver sistemas capazes de avaliar automaticamente textos escritos. Os trabalhos citados exploram a correção de textos do gênero narrativo, considerando suas competências específicas e propondo modelos multi-saída para avaliação simultânea. No cenário internacional, as pesquisas em AES concentram-se majoritariamente na língua inglesa, como o estudo de Yang et al. [2020], no qual os autores utilizam modelos pré-treinados BERT com o tradicional conjunto de redações em inglês ASAP. A proposta combina técnicas de *fine-tuning* com algoritmos de regressão e ranking, demonstrando que a integração de múltiplos objetivos de otimização favorece a captura de relações semânticas e estruturas textuais mais profundas.

Considerando as diferenças entre as abordagens baseadas em redes neurais e aquelas fundamentadas em aprendizado de máquina tradicional, estudos como Silveira et al. [2024]; Filho et al. [2021]; Silva et al. [2024]; Fonseca et al. [2018], dedicam-se à comparação entre essas estratégias, destacando vantagens, limitações e cenários de complementaridade. O trabalho Silveira et al. [2024], por exemplo, investigou o problema sob as perspectivas de classificação, regressão e regressão ordinal, concluindo que esta última apresentou os melhores resultados. Além disso, avaliou o *fine-tuning* em diferentes variantes do modelo BERT para o português, constatando que, embora as redes neurais

superem as abordagens de engenharia de atributos, o tamanho do modelo não exerce influência significativa sobre o desempenho final.

De forma semelhante, Filho et al. [2021] compararam abordagens baseadas em engenharia de atributos e redes neurais utilizando a base Essay-BR [Marinho et al., 2021]. Os autores testaram modelos como Gradient Boosting e LSTM, observando que as redes neurais obtiveram melhor desempenho geral, enquanto a engenharia de atributos, combinada ao balanceamento de classes por meio do algoritmo SMOTE, alcançou maior precisão nas notas extremas. Em uma análise mais abrangente, Fonseca et al. [2018] desenvolveram duas abordagens: uma fundamentada em redes neurais e outra em engenharia de atributos, avaliando separadamente cada competência. Os resultados indicaram que a engenharia de atributos apresentou melhor desempenho nas quatro primeiras competências, relacionadas à estrutura e ao domínio linguístico, enquanto as redes neurais se mostraram mais eficazes na quinta competência, voltada à proposta de intervenção. O estudo conclui que ambas as abordagens podem ser aplicadas de forma complementar, resultando na obtenção de melhores resultados para avaliação de redações de forma automática. Silva et al. [2024] investigam a detecção de redações fora do tema, problema diretamente ligado à segunda competência do ENEM e que pode levar modelos automáticos a superavaliar textos bem escritos, mas irrelevantes. Os autores compararam abordagens com atributos manuais, *fine-tuning* do BERTimbau [Souza et al., 2020] e prompt engineering, alcançando melhor desempenho com o modelo ajustado. O estudo reforça a importância de integrar técnicas de aprendizado profundo e análise semântica para aprimorar a robustez dos sistemas de AES em português.

III.3.2 Comparação

A Tabela III.1 apresenta uma visão geral das principais características dos trabalhos mencionados anteriormente. Ela sintetiza aspectos relevantes, distinguindo abordagens baseadas em atributos manuais ou TF-IDF, uso de representações neurais ou *embeddings*, aplicação de *fine-tuning* em modelos pré-treinados e presença de estratégias comparativas. Essa separação evita tratar toda forma de aprendizado de máquina como uma única categoria e facilita a identificação das semelhanças e diferenças entre os trabalhos analisados.

De modo geral, os trabalhos analisados anteriormente se dividem entre o uso de aprendizado de máquina tradicional (com uso de engenharia de atributos) e aplicação de modelos pré-treinados com *fine-tuning*. Ainda assim, é possível observar uma carência de estudos que avaliem diferentes estratégias de modelagem aplicadas às competências atualmente avaliadas no ENEM. O diferencial deste trabalho é observado na realização de uma comparação abrangente entre múltiplas abordagens, incluindo modelos tradicionais e variações com uso de *fine-tuning* com arquiteturas de múltipla saída, classificação ordinal e em cascata. Essa análise busca compreender os impactos de

cada abordagem em termos de desempenho, interpretabilidade e custo, contribuindo para o avanço das pesquisas em AES para língua portuguesa.

Trabalho	Tema principal	Aprendizado de máquina	*Fine-tuning* / Redes neurais	Por competência	Multi-saída	Cascata	Abordagem comparativa
Amorim and Veloso [2017]	Engenharia de atributos por competência em redações do ENEM	✓		✓			
Haendchen Filho et al. [2018]	Identificação de redações fora do tema com atributos linguísticos	✓					
Fonseca et al. [2018]	Comparação entre métodos baseados em atributos e redes neurais	✓	✓	✓			✓
Filho et al. [2021]	Comparação entre engenharia de atributos, LSTM e Gradient Boosting	✓	✓				✓
Silveira et al. [2024]	Avaliação de regressão, classificação e regressão ordinal com BERTimbau		✓	✓			✓
Ribeiro et al. [2024]	Modelagem multi-saída e avaliação por competência em textos narrativos		✓	✓	✓		
de Sousa et al. [2024]	*Fine-tuning* de modelos em português para avaliação de múltiplas competências		✓	✓	✓		
Silva et al. [2024]	Detecção de redações fora do tema com BERTimbau e *prompt engineering*	✓	✓				✓
Yang et al. [2020]	Combinação de perdas de regressão e *ranking* para *fine-tuning* no ASAP	✓	✓				✓
Este trabalho	**Avaliação comparativa de modelos tradicionais e de *fine-tuning* em arquiteturas por competência, multi-saída e em cascata para a predição das notas do ENEM.**	✓	✓	✓	✓	✓	✓

Tabela III.1: Visão geral das principais características dos trabalhos relacionados em avaliação automática de redações. A tabela mantém as categorias originais (aprendizado de máquina, *fine-tuning*) e destaca as abordagens exploradas neste trabalho: modelos por competência, modelos multi-saída e classificação em cascata, apresentando também se cada estudo realiza comparação entre diferentes métodos.

Capítulo IV Metodologia

Neste capítulo, descrevemos a metodologia proposta para investigar e comparar diferentes abordagens de modelagem aplicadas à avaliação automática de redações do ENEM, especificando como os dados foram estruturados, quais arquiteturas foram empregadas e de que forma o desempenho dos modelos foi analisado. A Seção IV.1 apresenta a formulação do problema. Na Seção IV.2, descrevemos a base de dados utilizada. A Seção IV.3 detalha as abordagens investigadas. Por fim, a Seção IV.4 traz os procedimentos de avaliação e as métricas de desempenho adotadas.

IV.1 Formulação do Problema

A avaliação automática de textos é uma tarefa desafiadora, especialmente quando se considera a necessidade de analisar múltiplas competências específicas de escrita. Este trabalho propõe a comparação de diferentes abordagens para a tarefa de avaliação de textos escritos em língua portuguesa com base em competências previamente definidas. Tais competências seguem as mesmas premissas avaliativas utilizadas pelo ENEM, o que adiciona uma camada de complexidade ao processo, dado seu caráter subjetivo e a necessidade de interpretá-las a partir do texto escrito pelo estudante. Além disso, a escassez de dados disponíveis em português com seus respectivos dados anotados dificultam o desenvolvimento e a validação de modelos eficazes.

O problema pode ser formulado como uma tarefa de aprendizado supervisionado, sendo tratado como classificação ou regressão. Seja $D = \{(x_i, \mathbf{y}_i)\}_{i=1}^{n}$ o conjunto de dados, onde x_i representa o texto da redação i, e $\mathbf{y}_i = (c_1^{(i)}, c_2^{(i)}, c_3^{(i)}, c_4^{(i)}, c_5^{(i)})$ é o vetor com os valores atribuídos às cinco competências avaliadas daquela redação.

O objetivo é aprender uma função $f : x \rightarrow \mathbf{y}$ capaz de prever os valores individuais das competências a partir do texto x, de modo que a soma $\sum_{j=1}^{5} c_j^{(i)}$ dos valores previstos se aproxime da nota total atribuída manualmente. Essa formulação permite explorar diferentes abordagens de modelagem, tanto na predição de cada competência de forma independente quanto na reconstrução da nota total como uma variável derivada.

A predição de competências individuais, em vez da nota total agregada, permite maior interpretabilidade dos resultados e maior alinhamento com os critérios pedagógicos utilizados na avaliação humana. Essa abordagem também viabiliza a identificação de pontos fortes e fracos em

cada redação, contribuindo para aplicações mais explicativas no contexto educacional. Além disso, a automatização desse processo possibilita um retorno mais rápido ao aluno, favorecendo o aprendizado contínuo e a correção de falhas ao longo do desenvolvimento da escrita.

IV.2 Base de dados

A primeira etapa deste trabalho consistiu na recuperação e no pré-processamento dos dados de redações utilizados nos experimentos. Os textos foram coletados a partir do portal Vestibular Brasil Escola[1], que publica mensalmente temas de redação e permite que estudantes submetam suas produções textuais para correção por profissionais qualificados, seguindo os critérios de avaliação adotados pelo ENEM. Após a correção, as redações são disponibilizadas publicamente, acompanhadas das notas atribuídas e das respectivas competências avaliadas. A escolha dessa plataforma deve-se à sua aderência aos critérios de correção e à qualidade das anotações fornecidas, o que confere maior confiabilidade às notas utilizadas como referência durante o treinamento dos modelos.

Como em qualquer tarefa supervisionada baseada em avaliações humanas, as notas utilizadas neste trabalho devem ser interpretadas como rótulos produzidos a partir de uma prática avaliativa específica, orientada por critérios institucionais e pedagógicos inspirados na matriz do ENEM. Portanto, os modelos treinados não aprendem uma medida objetiva e universal de qualidade textual, mas sim padrões de concordância com as decisões de correção presentes no corpus. Essa característica é importante para interpretar os resultados: altos valores de concordância indicam aproximação ao padrão de avaliação observado nos dados, sem eliminar possíveis variações, vieses ou diferenças de julgamento entre corretores humanos.

Vale ressaltar que a base de dados utilizada neste estudo foi empregada exclusivamente para fins de pesquisa acadêmica. Em conformidade com diretrizes éticas e restrições associadas à origem dos textos, o conjunto completo de redações não pode ser redistribuído publicamente. No entanto, são descritos detalhadamente os procedimentos de coleta, pré-processamento e estruturação, de modo a possibilitar a reprodutibilidade metodológica por outros pesquisadores.

Para essa etapa inicial, foi desenvolvido um script de *crawler* com o objetivo de automatizar a recuperação das redações disponibilizadas na plataforma, possibilitando a montagem do conjunto de dados. Como resultado, foram obtidas 9.599 redações distribuídas em 102 temas distintos, publicados entre maio de 2009 e dezembro de 2024.

Em seguida, os textos passaram por um processo estruturado de limpeza e normalização, com o objetivo de remover ruídos gerados pelo processo de recuperação das redações, padronizar a entrada e preservar elementos linguísticos relevantes para a avaliação automática. Inicialmente, foram removidas todas as marcações *HTML* presentes nos documentos, tais como *tags* de formatação,

[1] `https://vestibular.brasilescola.uol.com.br`

símbolos de navegação do portal, resíduos de tabelas e trechos auxiliares não pertencentes ao conteúdo produzido pelos candidatos. Também foram eliminados caracteres de controle, entidades *HTML* (como ` ` e `&`), espaços duplicados, quebras excessivas de linha e outros artefatos típicos de textos extraídos de páginas web.

Na etapa seguinte, realizou-se a normalização tipográfica, incluindo a padronização de aspas, travessões, hífens e sinais de pontuação, de modo a garantir consistência entre os textos. URLs, emojis e outros elementos não textuais foram substituídos por *tokens* especiais, prevenindo que esses itens interferissem na tokenização ou fossem interpretados como unidades semânticas válidas pelos modelos. Nenhuma operação como lematização, remoção de acentuação, conversão compulsória para caixa baixa ou correção ortográfica foi aplicada, garantindo que o texto preservasse características fundamentais da escrita estudantil, inclusive desvios ortográficos que fazem parte da avaliação e influenciam as notas atribuídas.

Ao final desse processo, cada redação foi estruturada em um registro padronizado contendo: `tema`, `texto`, as notas individuais de `competencia_1_nota` a `competencia_5_nota`, e a `nota_total`. Essa etapa de organização assegura a coerência entre os diferentes formatos de entrada utilizados nas abordagens investigadas. Exemplos completos de registros após o pré-processamento encontram-se nos Apêndices A e C.

Embora o aumento do volume de dados seja uma estratégia natural para aprimorar modelos supervisionados, a ampliação do corpus não deve ser tratada apenas como uma questão quantitativa. Em bases coletadas da web, é necessário considerar a qualidade, a procedência e a autenticidade dos textos, especialmente diante da possibilidade de redações produzidas ou auxiliadas por ferramentas de inteligência artificial generativa. A presença desse tipo de material pode alterar padrões linguísticos do corpus e contaminar o processo de treinamento, fazendo com que o modelo aprenda características que não correspondem integralmente à escrita humana espontânea dos candidatos.

Além da coleta, realizou-se uma análise exploratória dos dados com o objetivo de compreender melhor suas características. Foram avaliados aspectos como o tamanho médio das redações, o número de frases por texto e a variação na extensão dos textos. As principais informações obtidas estão apresentadas na Tabela IV.1.

Além das estatísticas gerais, foi realizada uma análise complementar da relação entre a extensão das redações e os rótulos utilizados na formulação multilabel. Nessa formulação, as notas originais de cada competência, registradas na escala de 0 a 200 pontos, foram convertidas em rótulos ordinais de 0 a 5 pelos intervalos 0, 1-40, 41-80, 81-120, 121-160 e 161-200. Esse mapeamento preserva a ordem dos níveis avaliativos, mas não a escala numérica original da pontuação. Por esse motivo, a soma dos cinco rótulos não é tratada como nota final da redação, mas apenas como uma medida auxiliar para observar a distribuição conjunta dos níveis atribuídos às competências.

Características	**Quantidade**
Total de redações	9.599 redações
Total de temas	102 temas
Tamanho médio das redações	304 palavras
Tamanho máximo de redação	1.599 palavras
Tamanho mínimo de redação	67 palavras
Tamanho médio das frases	26 palavras
Quantidade média de frases por redação	11 frases
Quantidade média de tokens	380 tokens
Redações utilizadas ($tokens \leq 512$)	6.957 redações
Redações removidas ($tokens > 512$)	2.642 redações

Tabela IV.1: Principais estatísticas da base de dados utilizada, composta por 9.599 redações distribuídas em 102 temas do portal *Vestibular Brasil Escola*. As redações possuem em média 304 palavras e 11 frases. Durante a tokenização, observou-se uma média de 380 tokens por texto. Considerando o limite de 512 tokens imposto pelos modelos, 6.957 redações foram mantidas e 2.642 removidas. Esses dados orientaram o pré-processamento e as configurações experimentais.

Com base nessa representação, foi calculada a correlação de Spearman entre o número de palavras da redação e a soma dos rótulos das cinco competências, obtendo-se $\rho = 0{,}487$. O resultado indica uma associação positiva moderada: no corpus analisado, textos mais extensos tendem a aparecer associados a rótulos mais altos. Essa relação, entretanto, não implica causalidade nem permite concluir que aumentar a extensão do texto, isoladamente, leve a melhor avaliação. A associação deve ser interpretada como uma característica descritiva da base, possivelmente relacionada ao fato de redações mais longas oferecerem mais espaço para desenvolvimento argumentativo, repertório sociocultural e detalhamento da proposta de intervenção. Ao mesmo tempo, esse comportamento reforça a necessidade de considerar a extensão textual e a limitação de contexto dos modelos na interpretação dos experimentos, especialmente porque parte das redações ultrapassa o limite de entrada adotado nas arquiteturas utilizadas.

A etapa de tokenização foi realizada com *SentencePiece* [Kudo and Richardson, 2018], empregado internamente pelos modelos baseados em XLM-RoBERTa. Este realiza a segmentação em unidades sublexicais (partes de palavras), dividindo as palavras em partes menores e permitindo lidar de forma eficiente com termos desconhecidos, variações morfológicas e erros ortográficos. Sinais de pontuação não foram removidos no pré-processamento; eles foram preservados após a normalização tipográfica e incorporados à sequência de entrada conforme o vocabulário do tokenizador. Ainda assim, como a representação final é aprendida de forma contextual pelos modelos pré-treinados, desvios locais de ortografia, concordância ou pontuação podem ser atenuados pela fluência global do texto. Essa característica é particularmente relevante para textos produzidos por

estudantes, que frequentemente contêm desvios linguísticos.

Durante o pré-processamento, foi aplicado o limite máximo de 512 tokens por redação, compatível com a capacidade dos modelos de linguagem pré-treinados [Yang et al., 2020]. Textos que ultrapassavam esse limite foram truncados ao final, enquanto aqueles com comprimento inferior foram completados com *padding* até o mesmo tamanho. Essa padronização assegura que todas as entradas apresentem o mesmo número de tokens, requisito fundamental para o processamento em lote durante o treinamento.

Essas informações foram essenciais para o planejamento das abordagens de avaliação automática, uma vez que o limite de tokens impacta diretamente a quantidade de textos que podem ser processados integralmente. Após a análise da distribuição de tokens, foram geradas duas versões do conjunto de dados: a primeira contendo todas as redações coletadas, e a segunda filtrando apenas aquelas cujo tamanho se mantinha dentro do limite suportado pelos modelos. Essa divisão permitiu avaliar se a restrição de tamanho influenciava o desempenho das abordagens propostas. Após o filtro, permaneceram 6.957 redações, enquanto 2.642 foram descartadas.

Como esse procedimento remove aproximadamente 27,5% do corpus original, a comparação entre o conjunto completo e o conjunto filtrado deve considerar a possibilidade de alteração na distribuição das notas. Neste trabalho, o filtro foi definido a partir de uma restrição técnica dos modelos, e não a partir das classes de nota; ainda assim, textos mais longos podem apresentar características discursivas distintas e, consequentemente, concentrar-se de forma diferente nas faixas de pontuação. Dessa forma, os resultados obtidos com o conjunto filtrado são interpretados em conjunto com essa limitação, e uma verificação estatística formal da aderência entre as distribuições de notas dos dois conjuntos é indicada como etapa complementar em trabalhos futuros.

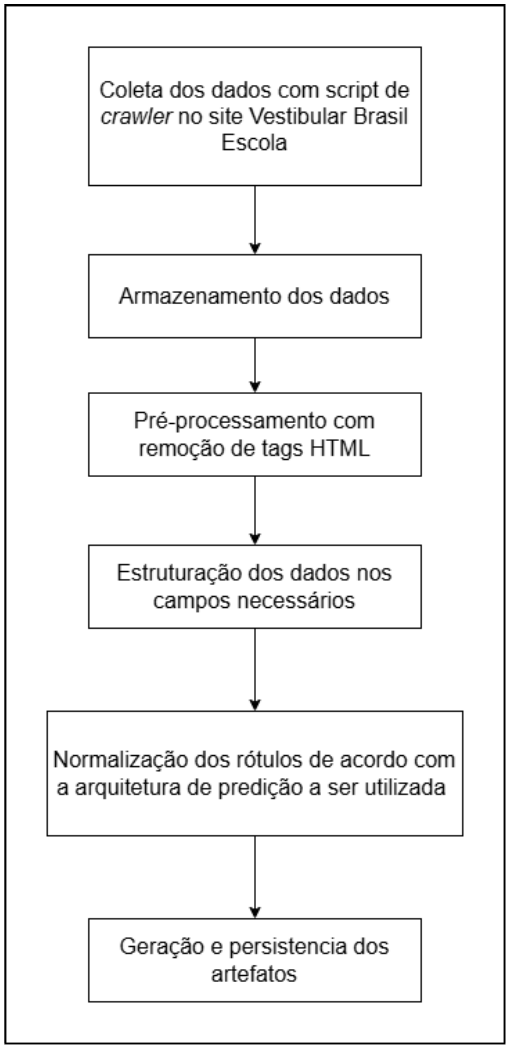


Figura IV.1: Fluxo de processamento dos dados utilizados como entrada para o treinamento dos modelos. O diagrama ilustra as principais etapas da construção do conjunto de dados, iniciando com a coleta automática das redações no portal *Vestibular Brasil Escola* por meio de um *crawler*, seguida pelo armazenamento e pré-processamento dos textos com remoção de marcações *HTML*. Em seguida, os dados são estruturados nos campos relevantes, normalizados conforme as necessidades de cada arquitetura de predição e, por fim, transformados em artefatos persistentes que servem como base para o treinamento e avaliação dos modelos.

Os dados processados foram organizados conforme as arquiteturas de predição investigadas. Nos modelos voltados à predição da nota geral, cada redação foi associada a um rótulo discreto entre 0 e 10, correspondendo a intervalos de 100 pontos na escala de 0 a 1000 (conforme apresentado na Tabela IV.2). Já nos modelos específicos por competência, cada instância continha o texto da redação e a nota atribuída àquela competência, representada em um intervalo de 0 a 5 (de 0 a 200 pontos, em incrementos de 40). Por fim, para os modelos multi-saída, cada redação foi vinculada a um vetor composto por cinco valores, um para cada competência, seguindo o mesmo mapeamento de rótulos.

Após essa fase de estruturação e normalização, os dados foram divididos em 60% para treinamento, 30% para validação e 10% para teste, com a divisão realizada de forma aleatória. Essa

partição foi mantida fixa entre as abordagens principais, de modo que as comparações fossem realizadas sobre os mesmos subconjuntos de treino, validação e teste. Quando uma configuração foi executada mais de uma vez, as repetições tiveram como objetivo estimar a variabilidade do treinamento e do tempo de execução sob o mesmo protocolo experimental, e não constituem validação cruzada. Assim, os desvios padrão reportados devem ser interpretados como variação entre execuções realizadas nas condições experimentais disponíveis, não como estimativa completa da variabilidade associada a diferentes partições do corpus.

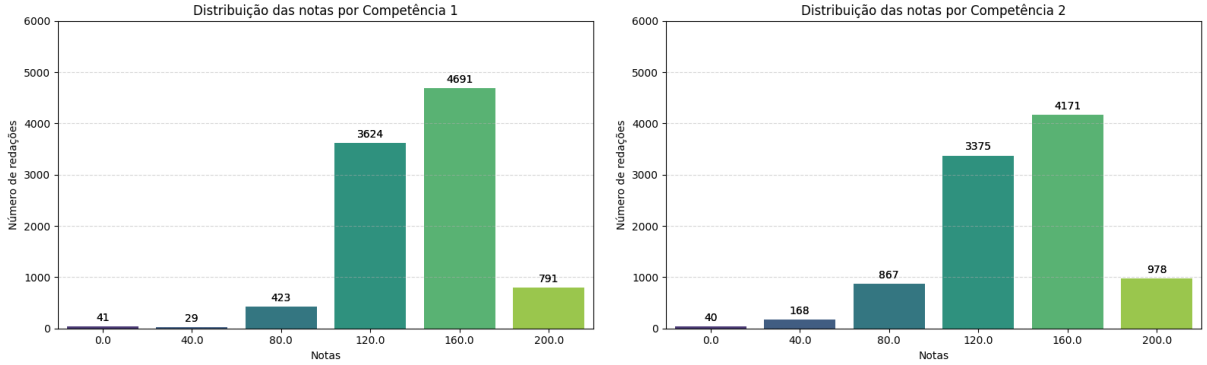
Intervalo de nota	**Label atribuído**
$nota = 0$	0
$0 < nota \leq 100$	1
$100 < nota \leq 200$	2
$200 < nota \leq 300$	3
$300 < nota \leq 400$	4
$400 < nota \leq 500$	5
$500 < nota \leq 600$	6
$600 < nota \leq 700$	7
$700 < nota \leq 800$	8
$800 < nota \leq 900$	9
$900 < nota \leq 1000$	10

Tabela IV.2: Conversão das notas contínuas das redações para rótulos discretos utilizados nos modelos de classificação. Cada intervalo de 100 pontos representa uma classe, variando de 0 a 10, o que permite tratar a previsão da nota total como um problema de classificação multiclasse.

#	**Competências**
1	Domínio da modalidade escrita formal da língua portuguesa.
2	Compreender a proposta e aplicar conceitos das várias áreas do conhecimento para desenvolver o texto dissertativo-argumentativo em prosa.
3	Selecionar, relacionar, organizar e interpretar informações em defesa de um ponto de vista.
4	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
5	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

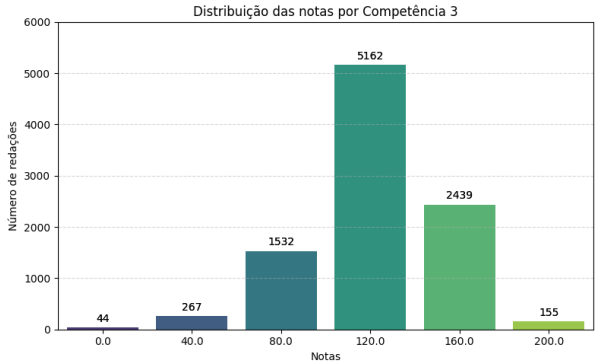
Tabela IV.3: Competências avaliativas do ENEM. Cada competência representa um aspecto específico da produção textual do participante, conforme a matriz oficial de correção.

Foi também realizada uma análise da distribuição das notas de cada uma das competências (listadas na Tabela IV.3), observando-se uma concentração maior de redações nas faixas intermediárias

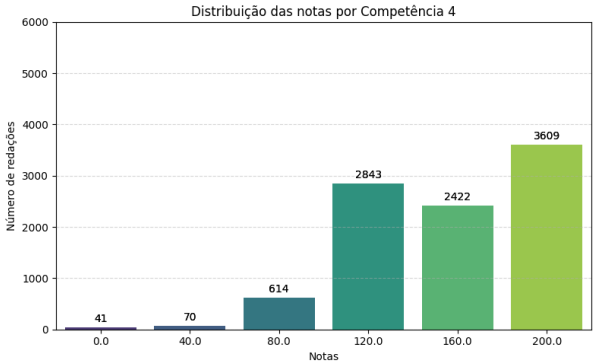


(a) Competência I

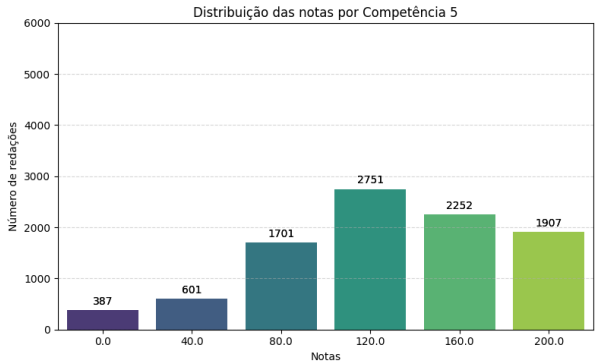
(b) Competência II



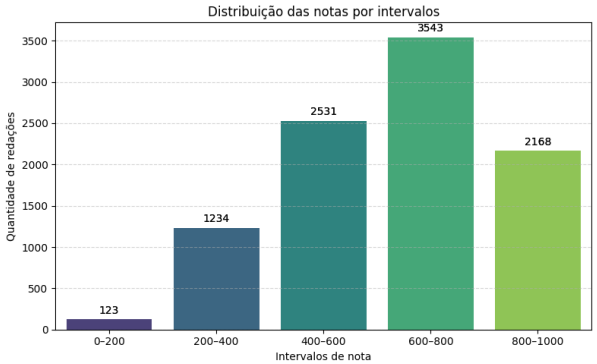
(c) Competência III



(d) Competência IV



(e) Competência V



(f) Nota Geral

Figura IV.2: Distribuição das notas atribuídas às redações em cada uma das cinco competências avaliativas do ENEM e da nota geral. Observa-se uma concentração maior de notas nas faixas intermediárias e superiores, especialmente próximas de 160 e 200 pontos, o que indica predominância de desempenhos médios e altos nas competências relacionadas à estrutura textual e à argumentação. A distribuição da nota geral segue padrão semelhante, com maior frequência de redações entre 600 e 800 pontos, refletindo a tendência geral de concentração em níveis intermediários de proficiência. Esses resultados evidenciam o desbalanceamento característico da base, fator relevante para a etapa de modelagem e treinamento dos modelos preditivos.

e uma quantidade reduzida de textos com notas mais baixas, conforme apresentado na Figura IV.2. Esse desbalanceamento reflete o comportamento típico de bases avaliativas do ENEM e pode influenciar o desempenho dos modelos durante a predição de redações com pontuações extremas. Além disso, tal resultado confirma o padrão descrito na literatura para corpora reais de redações,

geralmente marcados por distribuições assimétricas e forte concentração nas faixas intermediárias.

No processo de construção e preparação da base de dados, as etapas de coleta, limpeza, normalização, tokenização e divisão dos conjuntos de treino, validação e teste foram inteiramente automatizadas. Apenas a verificação final da qualidade dos dados e a análise exploratória descritiva foram conduzidas manualmente, a fim de garantir a consistência e a representatividade das amostras utilizadas nos experimentos.

De modo geral, o conjunto de dados construído apresenta volume, diversidade temática e qualidade suficientes para sustentar experimentos de avaliação automática em língua portuguesa. A combinação de textos autênticos, anotações consistentes e uma estrutura cuidadosamente normalizada oferece uma base sólida para as etapas seguintes desta pesquisa, voltadas à definição das abordagens de modelagem e às estratégias de treinamento dos modelos.

IV.3 Abordagens e arquiteturas utilizadas

Para a realização dos experimentos, foram desenvolvidas diferentes abordagens para avaliação automática de redações, cada uma baseada em premissas distintas quanto à estrutura do modelo, à granularidade da predição e à função de perda empregada. Todas as estratégias partiram de modelos pré-treinados com textos escritos em língua portuguesa (como XLM-RoBERTa, BERTimbau e AlBERTina), adaptado por meio da técnica LoRA, que permite o ajuste eficiente de parâmetros durante o processo de *fine-tuning*. A configuração e execução dos experimentos foram automatizadas por meio de *scripts* de treinamento parametrizáveis, enquanto a análise comparativa dos resultados foi manual, considerando as métricas e gráficos gerados automaticamente ao final das execuções.

A escolha desses modelos busca contemplar diferentes perspectivas em termos de vocabulário, tamanho e cobertura linguística. O XLM-RoBERTa é um modelo multilíngue de grande porte, treinado em um volume expressivo de dados em diversos idiomas, o que tende a gerar representações mais robustas e menos sensíveis a desvios ortográficos e morfológicos, comuns em textos produzidos por estudantes. Já o BERTimbau e o AlBERTina são modelos monolíngues especializados em português, o que permite capturar nuances gramaticais semânticas do idioma de forma mais precisa. Assim, a comparação entre esses três modelos permite avaliar o impacto do porte e da cobertura linguística na tarefa de AES para língua portuguesa.

A técnica LoRA foi escolhida para combinar eficiência e estabilidade durante o *fine-tuning*. No contexto deste trabalho, que envolve múltiplas arquiteturas e modelos, a LoRA reduz significativamente o número de parâmetros treináveis e o consumo de memória, sem comprometer o desempenho. Essa característica viabiliza a execução de diversos experimentos mesmo sob as limitações de hardware disponíveis.

O fluxo geral dos experimentos é apresentado na Figura IV.3, que resume as etapas de configuração, treinamento, validação e geração dos artefatos finais. O processo inicia-se com o modelo pré-treinado, que é ajustado via LoRA e, a partir dele, são exploradas três arquiteturas de predição: (i) o modelo multi-saída, no qual uma única rede realiza simultaneamente a predição das cinco competências; (ii) a abordagem de modelos específicos, que treina um modelo independente para cada competência; e (iii) a cascata ordinal, que organiza o processo de predição em etapas hierárquicas, refinando progressivamente as classes previstas.

Durante o treinamento, são ajustadas diversas configurações experimentais, como o tamanho do *batch*, o otimizador, o *scheduler* da taxa de aprendizado e o uso da técnica de *early stopping*, que contribuem para a estabilidade e a convergência dos modelos. O *early stopping* foi aplicado com paciência entre 10 e 20 épocas, e o critério monitorado variou conforme a estabilidade observada em cada configuração: em parte dos experimentos utilizou-se a função de perda de validação (*validation loss*), enquanto em outros adotou-se diretamente a métrica de QWK calculada no conjunto de validação. Em todos os casos, o checkpoint considerado como modelo final de cada execução foi sempre aquele que apresentou o maior valor de QWK no conjunto de validação, garantindo consistência na comparação entre as abordagens.

Por fim, os artefatos de saída incluem os modelos finais, gráficos de desempenho e registros detalhados das métricas obtidas, elementos que permitem a análise quantitativa e visual das abordagens testadas.

- **A1: Classificação da nota geral com *fine-tuning* padrão.** Essa abordagem busca prever diretamente a nota total da redação, tratada como um problema multiclasse com 11 categorias discretas, tal como mostrado na Tabela IV.2. O modelo pré-treinado em língua portuguesa é ajustado por meio da técnica LoRA, que permite atualizar apenas parte dos parâmetros durante o fine-tuning, reduzindo o custo computacional sem comprometer o desempenho. O treinamento é realizado com backpropagation e avaliado a cada época, registrando-se a função de perda e a métrica QWK. A seleção do checkpoint segue a estratégia de *early stopping* descrita nesta seção, sempre considerando, ao final, o melhor valor de QWK em validação. Essa configuração representa a forma mais direta de ajuste do modelo.

- **A2: Classificação ordinal da nota geral com WKLoss.** Essa abordagem trata a predição da nota total como uma tarefa de classificação ordinal, considerando a relação de ordem entre as classes. O modelo pré-treinado é ajustado com a técnica LoRA, e o treinamento utiliza a função de perda *Weighted Kappa Loss* [de la Torre et al., 2018], que penaliza mais fortemente erros distantes entre as classes reais e previstas. O desempenho é avaliado com

a métrica QWK e controlado por *early stopping*, conforme estratégia descrita anteriormente, buscando melhorar a coerência das previsões em relação à escala de notas.

- **A3: Classificação em cascata com *fine-tuning*.** Inspirada em Frank and Hall [2001], essa abordagem divide a tarefa de predição em três estágios sucessivos. No primeiro, um modelo binário distingue entre notas baixas (0–2) e altas (3–5). Em seguida, dois modelos especializados refinam as previsões: um dedicado às classes de 0 a 2 e outro às classes de 3 a 5. Todos os modelos são baseados em uma arquitetura pré-treinada ajustada com a técnica LoRA, o que permite realizar o *fine-tuning* de forma mais eficiente. Essa estrutura hierárquica busca reduzir ambiguidades nas fronteiras entre classes e melhorar a consistência das previsões, especialmente nas extremidades da escala de notas.

- **A4: Classificação multi-saída com *fine-tuning*.** Nessa configuração, um único modelo é treinado para prever simultaneamente as notas das cinco competências avaliativas. Para isso, é adicionada uma cabeça de saída independente para cada competência, todas conectadas a uma base pré-treinada ajustada com a técnica LoRA. Cada saída gera uma predição específica e sua respectiva perda é calculada com *CrossEntropyLoss*; a perda total corresponde à soma das perdas individuais. Essa arquitetura permite o aprendizado conjunto das competências, aproveitando correlações entre elas e reduzindo o custo computacional em relação ao treinamento de múltiplos modelos separados.

- **A5: Classificação multi-saída com *fine-tuning* e uso de WKLoss.** Nesta configuração, um único modelo possui múltiplas saídas, cada uma responsável pela predição de uma competência. A função de perda total é a soma das perdas individuais calculadas com *WKLoss* [de la Torre et al., 2018]. Essa função de perda é aplicada individualmente, de modo a preservar a distância ordinal entre as classes. Essa granularidade permite analisar separadamente o desempenho do modelo em cada dimensão avaliativa.

- **A6: Classificação por competência com *fine-tuning* padrão.** Essa abordagem realiza o treinamento de um modelo independente para cada competência avaliativa, utilizando o modelo pré-treinado ajustado com a técnica LoRA. O treinamento é feito com a função de perda *CrossEntropyLoss* padrão, sem considerar a relação ordinal entre as classes. O processo envolve o cálculo da métrica QWK a cada época, e a seleção do checkpoint segue o mesmo esquema de *early stopping* adotado nas demais abordagens, tendo como referência final o melhor valor de QWK em validação.

- **A7: Classificação por competência com *fine-tuning* e uso de WKLoss** Diferentemente da abordagem A6, esta configuração adota uma formulação ordinal explícita para a

tarefa de classificação por competência. Para cada competência avaliativa, é ajustado um modelo individual a partir de um modelo pré-treinado com LoRA, sendo a otimização conduzida pela função de perda *WKLoss* [de la Torre et al., 2018]. Essa função penaliza os erros de acordo com sua distância ordinal, refletindo de forma mais adequada a estrutura de pontuação das competências. O desempenho do modelo é acompanhado por meio da métrica QWK, calculada ao final de cada época, a qual também orienta o mecanismo de *early stopping* e a escolha do checkpoint final com melhor desempenho no conjunto de validação.

- **A8: Classificação sem *fine-tuning* (baseline).** Nessa configuração, o modelo pré-treinado é utilizado apenas como extrator de características, sem qualquer atualização de seus pesos. As representações vetoriais geradas são utilizadas diretamente para a etapa de classificação, permitindo avaliar o desempenho obtido exclusivamente a partir do conhecimento prévio do modelo. Essa abordagem serve como baseline para mensurar os ganhos de desempenho introduzidos pelo processo de *fine-tuning* nas demais estratégias.

- **A9: Regressão tradicional com TF–IDF.** Essa abordagem adota uma estratégia tradicional de aprendizado supervisionado, baseada na vetorização dos textos por meio da técnica TF-IDF. Sobre esses vetores, foram treinados modelos de regressão utilizando algoritmo *XGBoost Regressor*, escolhido por sua flexibilidade e interpretabilidade [Chen et al., 2024]. O treinamento foi realizado separadamente para cada competência, e o desempenho foi avaliado por meio das métricas RMSE e QWK. Essa configuração também funciona como um *baseline* clássico, permitindo comparar as abordagens baseadas em representações manuais de texto com aquelas que utilizam modelos de linguagem pré-treinados e *fine-tuning*.

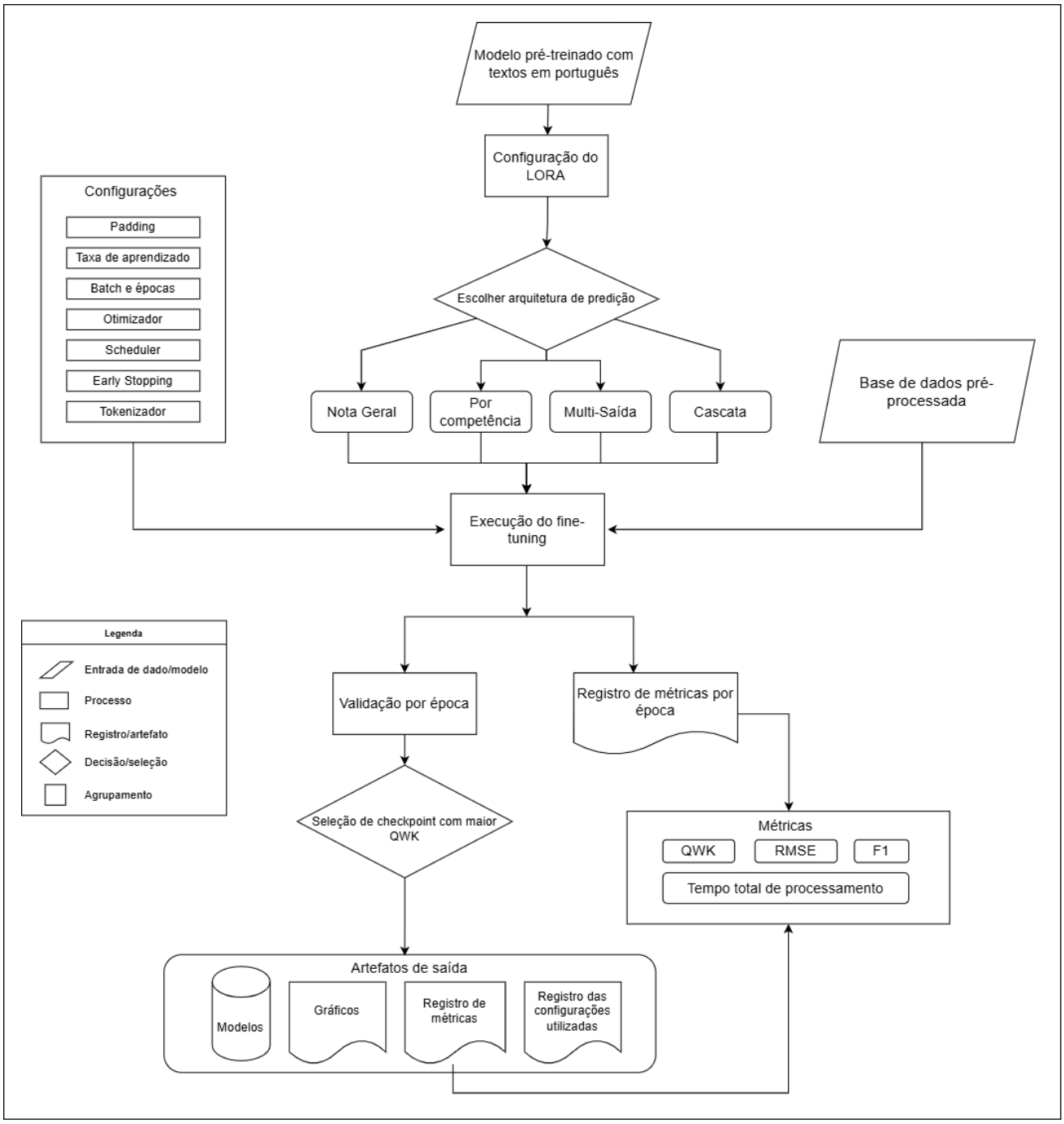


Figura IV.3: Fluxo completo do processo de treinamento e validação das arquiteturas de predição. O diagrama evidencia que a aplicação da técnica LoRA ocorre antes do processo de treinamento propriamente dito: primeiro o modelo pré-treinado é configurado estruturalmente com o LoRA, definindo quais parâmetros serão adaptados; somente depois dessa preparação é iniciado o *fine-tuning*, no qual são aplicados os hiperparâmetros de treinamento, como taxa de aprendizado, tamanho do *batch*, número de épocas e estratégia de *early stopping*. A partir dessa etapa, as diferentes arquiteturas de predição são treinadas sobre a base pré-processada, gerando métricas por época e permitindo a seleção do modelo com melhor desempenho segundo a métrica QWK. Por fim, são produzidos os artefatos finais e computadas as métricas (QWK, RMSE e F1), juntamente com o tempo total de processamento.

A seleção das arquiteturas avaliadas considerou limitações práticas de infraestrutura computacional disponível para treinamento, buscando assim o equilíbrio entre desempenho e viabilidade

computacional. Modelos de maior porte poderiam potencialmente apresentar ganhos adicionais, porém demandariam recursos substancialmente superiores.

IV.4 Avaliação e métricas de desempenho

A avaliação dos modelos propostos considerou tanto o desempenho por competência quanto a qualidade da predição da nota total reconstruída. A principal métrica utilizada foi a QWK, por medir a concordância ponderada entre predições e avaliações humanas e penalizar proporcionalmente a distância entre classes. Por essa razão, a QWK é especialmente adequada a problemas com natureza ordinal e subjetiva, como a atribuição de notas em escalas discretas para redações.

Como métricas complementares, reportamos Acurácia (ACC) e RMSE, a fim de oferecer uma visão mais abrangente do desempenho em termos de acerto categórico e erro médio quadrático, respectivamente.

Por fim, registramos o tempo de processamento por abordagem (treinamento e inferência) para analisar a viabilidade prática de uso.

Capítulo V Experimentos

Este capítulo apresenta os experimentos conduzidos para avaliar diferentes estratégias de modelagem na tarefa de correção automática de redações em português. O foco é comparar abordagens tradicionais baseadas em TF-IDF com modelos de linguagem pré-treinados submetidos a *fine-tuning*, analisando como variam em desempenho, custo computacional e capacidade de generalização. Investigamos também o impacto da formulação da tarefa (classificação padrão, ordinal, multi-saída e cascata), do tipo de modelo adotado (monolíngue ou multilíngue) e da limitação de contexto imposta pelo limite de 512 tokens.

O capítulo está organizado da seguinte forma. Na Seção V.1, retomamos o corpus descrito na Seção IV.2 e detalhamos os dois recortes empregados nos experimentos: o conjunto completo e o conjunto filtrado por limite de tokens. A Seção V.2 descreve as configurações de treinamento adotadas, incluindo hiperparâmetros, critérios de parada antecipada e o ambiente computacional utilizado. A Seção V.3 analisa o custo computacional associado a cada combinação de abordagem e modelo. A Seção V.4 apresenta o plano experimental, os modelos avaliados, os procedimentos adotados e as questões de pesquisa investigadas. Os resultados são discutidos na Seção V.5, que abrange o panorama geral de desempenho, a evolução da métrica QWK ao longo do treinamento, os efeitos da limitação de contexto, a comparação com abordagens de referência, as respostas às questões de pesquisa, a análise de erros por matrizes de confusão e estudos de caso representativos. Por fim, a Seção V.6 sintetiza os principais achados, discute suas implicações para a tarefa de AES e posiciona os resultados obtidos em relação aos trabalhos relacionados.

V.1 Corpus de Redações para Treinamento e Avaliação

O *corpus* apresentado na Seção IV.2 foi utilizado como base para condução de experimentos para avaliação dos modelos propostos. Seu volume e disponibilidade de informações permitiram investigar diferentes estratégias de modelagem voltadas para correção automática de redações. O conjunto foi particionado em subconjuntos de treino, validação e teste. Tal separação permitiu avaliar a capacidade de generalização dos modelos, principalmente em cenários de variação temática e desbalanceamento de notas. Além disso, trabalhamos com dois recortes: Conjunto Completo (textos possivelmente > 512 tokens, truncados ao fim) e Conjunto Filtrado (máximo de 512 tokens,

sem truncamento). Esse desenho isola o efeito da limitação de contexto sobre o desempenho dos modelos.

V.2 Configurações experimentais

Para a realização dos experimentos, adotaram-se configurações padronizadas de treinamento, definidas a partir de testes preliminares voltados à estabilidade e à convergência dos modelos. O processo de ajuste considerou variações controladas nos principais hiperparâmetros, de modo a equilibrar desempenho e custo computacional.

Parâmetro	**Valor / Faixa**	**Descrição**
Número de épocas	70–150	Ajustado conforme convergência e tamanho do modelo
Early stopping	*patience* = 10-20	Evita *overfitting* com base na métrica QWK
Taxa de aprendizado	1e-5 a 2e-5	Variação entre arquiteturas *base* e *large*
Tamanho do *batch*	10–22	Dependente do modelo e da GPU
Otimizador	AdamW	Padrão em todos os experimentos
Técnica LoRA	$r = 8$, $\alpha = 32$	Redução de parâmetros treináveis

Tabela V.1: Resumo dos principais parâmetros utilizados durante o treinamento dos modelos. A tabela apresenta as configurações mais relevantes adotadas nos experimentos, incluindo o número de épocas, a taxa de aprendizado, o tamanho do *batch* e os critérios de *early stopping*. Também são indicados o otimizador utilizado (AdamW) e os parâmetros da técnica de adaptação LoRA, empregada para reduzir o número de parâmetros treináveis e otimizar o uso de memória durante o *fine-tuning*. Essas configurações foram mantidas de forma consistente naquilo que era comum a todas as abordagens, enquanto variações controladas (como o número de épocas ou o critério monitorado pelo *early stopping*) foram ajustadas conforme a estabilidade observada em cada modelo. Esse cuidado garantiu comparações justas sem comprometer a convergência dos experimentos.

A definição das faixas de hiperparâmetros considerou o porte das arquiteturas avaliadas e restrições de memória da GPU utilizada. Modelos large demandaram batches menores, enquanto modelos base permitiram maior paralelização.

O critério de *early stopping* variou conforme a natureza da abordagem empregada. Nos modelos treinados individualmente por competência e nas abordagens de nota geral, o monitoramento foi realizado diretamente com base na métrica QWK no conjunto de validação, por se tratar da métrica principal do estudo e por apresentar comportamento estável nesses cenários.

Já nas abordagens multi-saída, nas quais múltiplas competências são previstas simultaneamente, optou-se por monitorar inicialmente a função de perda de validação, devido à maior variabilidade

observada na QWK conjunta nas primeiras épocas. A seleção final do melhor modelo, entretanto, considerou sempre o maior valor de QWK obtido na validação.

Com o objetivo de avaliar o efeito do critério de otimização sobre o desempenho, a maioria das abordagens foi treinada em duas variações: (i) Cross-Entropy Loss, como formulação categórica padrão, e (ii) WKLoss, buscando alinhar o treinamento à natureza ordinal das notas e à métrica principal adotada (QWK). Essa estratégia foi aplicada de forma consistente nas abordagens avaliadas, permitindo comparação controlada entre funções de custo dentro de uma mesma arquitetura. A exceção foi a abordagem em cascata, para a qual foi considerada apenas a configuração com Cross-Entropy, devido a restrições específicas de implementação/estabilidade do pipeline em múltiplos estágios.

Os valores apresentados na Tabela V.1 foram definidos com base em uma série de execuções exploratórias conduzidas para identificar as combinações que melhor se adaptavam às particularidades das tarefas e dos conjuntos de dados empregados. O uso de *early stopping*, com paciência entre 10 e 20 épocas, permitiu interromper o treinamento quando a métrica de validação não apresentava melhora significativa. O critério monitorado pelo *early stopping* seguiu a seguinte regra: nas abordagens treinadas com saídas únicas (modelos independentes por competência e variantes de nota geral), o monitoramento foi realizado diretamente sobre a métrica QWK no conjunto de validação. Já nas abordagens multi-saída, devido à maior variabilidade da QWK agregada nas primeiras épocas, o monitoramento foi feito inicialmente pela função de perda de validação, mantendo-se, entretanto, a seleção final do melhor *checkpoint* sempre baseada no maior valor de QWK na validação. Já a variação do número de épocas entre 70 e 150 refletiu as diferenças de complexidade entre modelos de distintos tamanhos, especialmente entre as arquiteturas *base* e *large*.

A técnica de adaptação eficiente LoRA foi aplicada a todos os experimentos de *fine-tuning*, reduzindo o número de parâmetros treináveis e, consequentemente, o consumo de memória, sem prejuízo de desempenho. O otimizador AdamW foi escolhido por sua robustez na regularização dos pesos e estabilidade durante o treinamento.

Componente	**Descrição**
GPUs	6× NVIDIA GeForce RTX 3090 (24 GB cada)
CPU	2× AMD EPYC 7282 (16 cores, 32 threads)
Memória RAM	256 GB DDR4
Sistema operacional	Ubuntu 22.04 LTS
CUDA / Driver NVIDIA	12.4 / 550.xx
Uso médio de memória GPU	20–22 GB por modelo

Tabela V.2: Especificações do ambiente de execução dos experimentos.

Os experimentos foram executados em ambiente de alto desempenho composto pelas configurações apresentadas na Tabela V.2. O conjunto de GPUs e CPUs de alta capacidade possibilitou manter a estabilidade mesmo em modelos de grande porte. A utilização média de memória por GPU variou entre 20 GB e 22 GB, dependendo da arquitetura e da complexidade da tarefa.

Vale ressaltar que, antes da consolidação desses parâmetros, foram conduzidas execuções exploratórias para determinar as combinações mais adequadas ao problema em questão, assegurando um equilíbrio entre desempenho, custo computacional e tempo de treinamento.

V.3 Tempo de treinamento dos modelos

A comparação entre os tempos de treinamento permite avaliar o custo computacional associado às diferentes abordagens e arquiteturas utilizadas nos experimentos. A Tabela V.3 apresenta a média, o desvio padrão e o número de execuções realizadas para cada combinação entre abordagem e modelo, considerando tempos medidos em horas de relógio de parede. Os códigos das abordagens seguem a nomenclatura definida no Capítulo IV. Como as configurações seguiram um protocolo experimental comum, ainda que com quantidades distintas de execuções por abordagem, as diferenças observadas refletem principalmente o porte das arquiteturas e a complexidade das funções de perda.

Abordagem	**Modelo**	**Tempo (h)**	**Rodadas**
Sem *fine-tuning* (A8)	BERTimbau-large	$0{,}11 \pm 0{,}02$	7
Sem *fine-tuning* (A8)	XLM-RoBERTa-large	$0{,}18 \pm 0{,}20$	6
Sem *fine-tuning* (A8)	XLM-RoBERTa-base	$0{,}19 \pm 0{,}15$	5
Multi-saída com *fine-tuning* e WKLoss (A5)	AlBERTina-100m	1,35	1
Multi-saída com *fine-tuning* (A4)	XLM-RoBERTa-base	$9{,}76 \pm 13{,}60$	13
Multi-saída com *fine-tuning* e WKLoss (A5)	XLM-RoBERTa-base	$9{,}95 \pm 13{,}21$	13
Multi-saída com *fine-tuning* e WKLoss (A5)	XLM-RoBERTa-large	$12{,}04 \pm 8{,}21$	6
Multi-saída com *fine-tuning* (A4)	AlBERTina-100m	$12{,}13 \pm 13{,}34$	3
Com *fine-tuning* e WKLoss (A7)	XLM-RoBERTa-base	$12{,}57 \pm 9{,}95$	15
Com *fine-tuning* e WKLoss (A7)	BERTimbau-large	$13{,}15 \pm 21{,}02$	3
Com *fine-tuning* (A6)	XLM-RoBERTa-large	$13{,}47 \pm 7{,}06$	7
Multi-saída com *fine-tuning* (A4)	BERTimbau-large	$14{,}51 \pm 7{,}31$	10
Multi-saída com *fine-tuning* e WKLoss (A5)	BERTimbau-large	$14{,}90 \pm 8{,}51$	9
Multi-saída com *fine-tuning* (A4)	XLM-RoBERTa-large	$16{,}35 \pm 7{,}71$	6
Com *fine-tuning* (A6)	BERTimbau-large	$19{,}23 \pm 1{,}84$	3
Com *fine-tuning* e WKLoss (A7)	XLM-RoBERTa-large	$20{,}49 \pm 17{,}26$	8
Com *fine-tuning* (A6)	XLM-RoBERTa-base	$26{,}42 \pm 35{,}20$	4

Tabela V.3: Tempo médio total de treinamento, medido em horas de relógio de parede, para cada combinação entre abordagem e modelo, acompanhado do desvio padrão quando disponível. Os códigos das abordagens (A1-A8) correspondem às definições apresentadas no Capítulo IV. A apresentação conjunta da média e do desvio padrão permite comparar diretamente o custo computacional das diferentes arquiteturas, bem como a variabilidade observada entre as execuções.

Observa-se que as arquiteturas multilíngues de maior porte apresentam crescimento exponencial no tempo de treinamento, especialmente nas configurações multi-saída. Em alguns cenários, o tempo ultrapassa 1000 horas acumuladas de execução, tornando tais abordagens pouco viáveis em contextos institucionais com infraestrutura limitada.

Os experimentos foram conduzidos em ambiente computacional com GPU NVIDIA RTX 3090, equipamento cujo custo é significativamente inferior ao de aceleradores de última geração utilizados. Essa limitação impõe restrições práticas à escalabilidade dos experimentos, mas também posiciona este trabalho como uma investigação orientada à viabilidade econômica no contexto brasileiro.

De modo geral, os resultados evidenciam um contraste expressivo entre abordagens sem *fine-tuning* e aquelas que envolvem otimização completa ou parcial do modelo. As execuções sem ajuste fino apresentam tempos extremamente reduzidos. Embora eficientes, essas variantes servem principalmente como linha de base, dado que não incorporam adaptação para a tarefa específica de avaliação textual.

Entre as abordagens com *fine-tuning*, observa-se um crescimento substancial no custo computacional, sobretudo nos modelos multilíngues de grande porte. Nas variantes multi-saída, por exemplo, os tempos ultrapassam 1.100 horas em algumas configurações, resultado esperado em cenários que exigem otimização simultânea de múltiplas predições. Por outro lado, modelos monolíngues como o BERTimbau-large mantêm um custo muito inferior ao de suas contrapartes multilíngues, reforçando a eficiência de arquiteturas especializadas no português.

As abordagens que utilizam a função de perda *WKLoss* também apresentam, de forma recorrente, tempos superiores aos do *fine-tuning* padrão. Isso ocorre porque a função ordinal introduz uma estrutura de penalização mais sensível às discrepâncias entre classes, aumentando frequentemente o número de iterações necessárias para atingir o critério de parada. Ainda assim, essa complexidade adicional pode ser justificada quando há ganho de desempenho, uma vez que a função de perda se alinha à natureza ordinal das notas, conforme discutido na Seção II.6.

Em síntese, os resultados evidenciam que abordagens mais complexas, embora potencialmente mais precisas, implicam custos computacionais significativamente superiores. Essa diferença é particularmente relevante em cenários de larga escala, nos quais a viabilidade operacional depende do equilíbrio entre desempenho preditivo e eficiência de processamento.

V.4 Plano de Experimentos

Esta seção apresenta o plano de experimentos utilizado para avaliar o desempenho de diferentes modelos de linguagem no problema de correção automática de redações em português. O objetivo central é operacionalizar as questões de pesquisa definidas no Capítulo I, considerando os modelos, conjuntos de dados e estratégias de treinamento viáveis no escopo computacional deste trabalho.

O desenho experimental foi organizado para conectar cada questão de pesquisa a uma evidência observável. Assim, a comparação entre versões *base* e *large* do XLM-RoBERTa foi utilizada como evidência para a QP1; a análise entre modelos monolíngues e multilíngues foi usada para discutir a QP2; a comparação entre classificação categórica e formulações com WKLoss operacionalizou

a QP3; os resultados separados por competência permitiram discutir a QP4; e a avaliação dos conjuntos completo e filtrado investigou a QP5.

V.4.1 Modelos Avaliados

A escolha dos modelos considerou três critérios principais: (i) disponibilidade de versões pré-treinadas para português; (ii) compatibilidade com fine-tuning eficiente via LoRA; e (iii) viabilidade de treinamento no ambiente computacional disponível. Modelos com bilhões de parâmetros, embora potencialmente mais expressivos, foram descartados por inviabilidade computacional no contexto experimental.

Foram comparados os seguintes modelos:

- **BERTimbau-base**: versão baseada no BERT com 110 milhões de parâmetros, treinada especificamente para o português [Souza et al., 2020]. Foi utilizado como *baseline* devido ao seu custo computacional reduzido.
- **XLM-RoBERTa-base**: versão multilíngue intermediária com 270 milhões de parâmetros e treinado em 100 idiomas (incluindo o português) [Conneau et al., 2019]. Serve como referência para avaliar escalabilidade e custo computacional.
- **XLM-RoBERTa-large**: modelo multilíngue de 550 milhões de parâmetros, também treinado em 100 idiomas [Conneau et al., 2019]. Foi empregado para investigar se o aumento de escala e o pré-treinamento multilíngue contribuem para ganhos no contexto de redações.
- **AlBERTina-100m-portuguese-pt-br**: baseado na arquitetura BERT, foi pré-treinado especificamente para o português do Brasil. Com aproximadamente 100 milhões de parâmetros, apresenta menor porte em relação a outros modelos da mesma família, mas mantém alta eficiência e desempenho competitivo em tarefas de compreensão de linguagem natural [Santos, 2024].

V.4.2 Procedimentos Experimentais

O conjunto de experimentos planejados cobre tanto a comparação sistemática de abordagens tradicionais quanto a exploração de novos cenários viabilizados pela infraestrutura multi-GPU, permitindo uma análise abrangente sobre as potencialidades e limitações de cada modelo:

1. **Classificação da nota geral com *fine-tuning***

   A nota total da redação é prevista diretamente como uma classe discreta, sem modelar explicitamente a relação ordinal entre as categorias.

2. **Classificação ordinal da nota geral com *fine-tuning* e WKLoss**

   A tarefa é formulada como um problema de classificação ordinal, no qual erros mais distantes entre as categorias de nota recebem penalizações mais severas por meio da função de perda WKLoss.

3. **Classificação em cascata com *fine-tuning***

   A predição é realizada de forma sequencial em múltiplos estágios, nos quais cada modelo refina a estimativa da nota ao restringir progressivamente o intervalo de valores possíveis.

4. **Classificação multi-saída com *fine-tuning***

   Um único modelo é treinado com múltiplas saídas, cada uma responsável pela predição de uma competência, assumindo independência entre as saídas durante a otimização.

5. **Classificação multi-saída com *fine-tuning* e WKLoss**

   Preserva a arquitetura multi-saída, porém adota uma função de perda ordinal, permitindo que a otimização leve em conta a estrutura de ordenação das notas associadas a cada competência.

6. **Classificação por competência com *fine-tuning* padrão**

   São treinados modelos independentes para cada uma das cinco competências avaliativas, utilizando uma formulação de classificação padrão, sem considerar explicitamente a estrutura ordinal das notas.

7. **Classificação ordinal por competência com *fine-tuning* e WKLoss**

   Mantém a decomposição por competência, porém incorpora uma formulação ordinal explícita, explorando a relação de ordem entre as categorias por meio da função de perda WKLoss.

8. **Classificação sem *fine-tuning* (baseline)**

   As representações extraídas do modelo pré-treinado são utilizadas diretamente como características, sem qualquer ajuste adicional dos pesos durante o treinamento.

9. **Escalonamento para modelos maiores**

   Avalia o impacto do aumento do porte do modelo, comparando versões *base* e *large* do XLM-RoBERTa em termos de desempenho, custo computacional e consumo de memória.

10. **Estratégias de *fine-tuning***

    Compara o ajuste completo dos parâmetros do modelo com abordagens de adaptação eficiente, em particular o uso da técnica LoRA.

V.4.3 Questões da pesquisa

As questões de pesquisa foram apresentadas no Capítulo I. Nesta etapa, elas foram traduzidas em comparações experimentais específicas:

- **QP1**: comparação entre modelos de diferentes portes, especialmente XLM-RoBERTa-base e XLM-RoBERTa-large, para verificar se o aumento de escala se traduz em melhor desempenho no corpus analisado.
- **QP2**: comparação entre modelos monolíngues especializados em português, como BERTimbau e AlBERTina, e modelos multilíngues, considerando conjuntamente desempenho e tempo de treinamento.
- **QP3**: comparação entre formulações categóricas com Cross-Entropy Loss e formulações ordinais com WKLoss, mantendo a arquitetura de predição tão próxima quanto possível.
- **QP4**: análise dos resultados por competência, observando se dimensões distintas da matriz avaliativa apresentam padrões diferentes de erro, estabilidade e concordância.
- **QP5**: comparação entre o conjunto completo, submetido a truncamento quando necessário, e o conjunto filtrado, composto apenas por redações com até 512 *tokens*.

Essa operacionalização não esgota todas as formas possíveis de investigar as questões, mas permite obter evidências controladas dentro das restrições de dados, modelos e infraestrutura disponíveis.

V.5 Resultados e Análises

Nesta seção analisamos os resultados obtidos a partir dos experimentos realizados, tomando como base as questões de pesquisa definidas anteriormente na Seção V.4.3. Inicialmente, são apresentados os melhores desempenhos obtidos em cada uma das competências avaliadas, considerando as diferentes abordagens propostas. Em seguida, são discutidos os comportamentos observados em cada uma das competências, procurando identificar padrões de desempenho, assim como aspectos que se mostrem diferenciais na contribuição para a eficácia da abordagem. A análise também contempla a comparação entre métricas utilizadas, destacando a relação entre o QWK e outras medidas de desempenho. Por fim, são sintetizados os principais pontos, relacionando-os às questões formuladas e discutindo suas implicações para o problema abordado nesta dissertação.

V.5.1 Panorama e desempenho geral das abordagens

Para este trabalho foram desenvolvidas diferentes abordagens para lidar com a tarefa de AES para a língua portuguesa, conforme mostrado na Seção IV.3. A Tabela V.4 mostra os melhores valores de QWK obtidos para cada competência avaliada, além de trazer as principais configurações e modelos pré-treinados utilizados.

Competência	Abordagem	Modelo	Tipo de dataset	QWK
C-I	Com *fine-tuning* e *WKLoss* (A7)	XLM-RoBERTa-base	Conjunto Completo	**0,659**
C-II	Com *fine-tuning* e *WKLoss* (A7)	XLM-RoBERTa-base	Conjunto Completo	**0,657**
C-III	Com *fine-tuning* e *WKLoss* (A7)	XLM-RoBERTa-base	Conjunto Completo	**0,642**
C-IV	Com *fine-tuning* (A6)	BERTimbau-large	Conjunto Filtrado	**0,729**
C-V	Com *fine-tuning* (A6)	BERTimbau-large	Conjunto Filtrado	**0,748**

Tabela V.4: Resumo dos melhores resultados obtidos em cada competência do ENEM. A tabela apresenta, para cada dimensão avaliada, a melhor combinação de abordagem, modelo e tipo de dataset que resultou no maior valor da métrica QWK. Observa-se que as competências C-I (domínio da modalidade escrita formal), C-II (compreensão da proposta) e C-III (organização e defesa do ponto de vista) obtiveram melhor desempenho com abordagens de classificação ordinal, o que indica que explorar a estrutura de ordem entre as classes contribui para capturar nuances de progressão textual. Por outro lado, as competências C-IV (mecanismos linguísticos) e C-V (proposta de intervenção) apresentaram resultados superiores com o *fine-tuning* padrão em tarefas individuais, sugerindo que a especialização por competência permite um ajuste mais preciso a padrões discursivos específicos. Além disso, nota-se que o uso do conjunto de dados *filtrado* (redações com até 512 tokens) favoreceu o desempenho dos modelos em diversas competências, ao passo que o conjunto *completo* (sem limitação de tamanho) proporcionou melhor aproveitamento do contexto textual em tarefas com maior dependência semântica. Essa diferença reforça a influência da limitação de contexto sobre a capacidade dos modelos em representar adequadamente os textos avaliados.

É possível ver também na Tabela V.4 que para as competências C-I, C-II e C-III, a melhor abordagem foi a classificação ordinal com WKLoss e com o uso do modelo *XLM-RoBERTa-base*. Esse resultado indica que as competências associadas à coerência textual, à compreensão da proposta e à estrutura argumentativa (C-I, C-II e C-III) se beneficiam do uso de uma função de perda ordinal, capaz de capturar relações graduais de proficiência entre as classes de nota. O WKLoss impõe uma penalização proporcional à distância entre as notas previstas e as reais, favorecendo o aprendizado da ordem implícita existente nas escalas de avaliação de redações. Esse comportamento sugere que a modelagem ordinal é mais adequada para competências cuja progressão qualitativa é mais contínua e dependente de aspectos discursivos globais, como clareza, progressão e coerência.

Por outro lado, as competências C-IV e C-V apresentaram melhores resultados com o *fine-tuning* padrão em tarefas individuais, conforme evidenciado pelos valores mais altos de QWK obtidos com o *BERTimbau-large*. Essas competências exigem análises mais específicas, o que favorece abordagens especializadas que ajustam os parâmetros do modelo exclusivamente para cada uma, ajudando a explicar por que uma versão de modelo pré-treinado focada na língua portuguesa se adaptou melhor a essa abordagem.

O Apêndice D apresenta resultados mais completos das abordagens avaliadas em cada competência, detalhando as variações de desempenho e as configurações experimentais correspondentes. Essa seção adicional complementa a análise deste capítulo, permitindo uma compreensão mais aprofundada do comportamento dos modelos sob diferentes conjuntos de dados (completo e filtrado).

Evolução do QWK ao longo do treinamento

A fim de analisar o comportamento dos modelos ao longo do treinamento, foram gerados gráficos da evolução da métrica QWK nas curvas de treino e validação para os cinco modelos que apresentaram os melhores resultados em cada competência. Esses gráficos permitem observar a estabilidade do processo de aprendizagem, o ponto de convergência e o comportamento da validação utilizado como referência para o *early stopping*.

A Figura V.1 mostra que o treinamento do modelo para a competência C-I evolui de maneira estável desde as primeiras épocas. A curva de treino cresce de forma contínua até atingir um platô por volta da vigésima época, indicando convergência consistente. Já a curva de validação oscila dentro de um intervalo estreito, permanecendo majoritariamente entre 0,55 e 0,65 sem apresentar divergências acentuadas. Esse comportamento demonstra que o modelo internaliza adequadamente os padrões relacionados ao domínio da modalidade escrita formal, mantendo boa capacidade de generalização e sem sinais relevantes de *overfitting*.

A Figura V.2 mostra que o aprendizado do modelo para a competência C-II evolui de modo consistente desde as primeiras épocas. A validação cresce de maneira rápida até cerca da quinta época e, após esse ponto, estabiliza-se na faixa entre 0,50 e 0,60, ainda que com pequenas oscilações naturais do processo de validação. Já a curva de treino continua subindo de modo gradual, atingindo valores próximos de 0,75, mas sem se afastar excessivamente da validação. Esse padrão indica capacidade de generalização adequada e justifica o ponto selecionado via *early stopping*, que captura o momento em que o desempenho de validação atinge estabilidade.

A Figura V.3 apresenta um comportamento significativamente mais instável para a competência C-III. Embora as primeiras épocas exibam crescimento consistente, a curva de validação passa a oscilar de forma intensa a partir da metade do treinamento, variando entre valores próximos de

Competência I

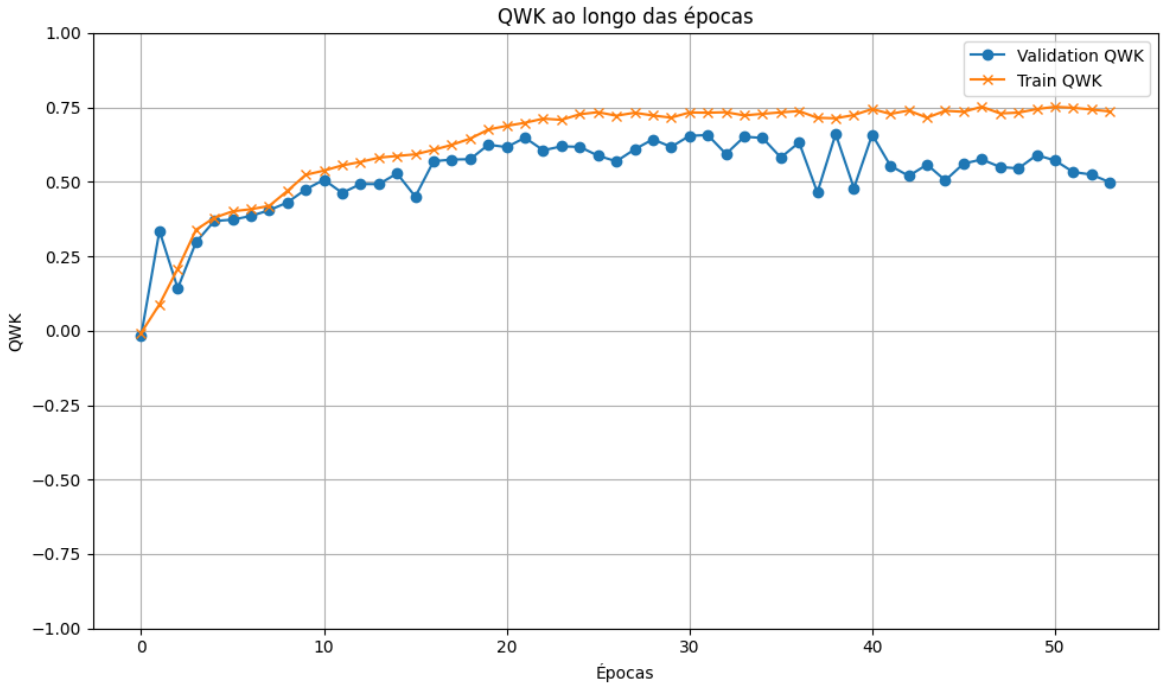


Figura V.1: Evolução da métrica QWK durante o treinamento do modelo XLM-RoBERTa-base com classificação ordinal e *WKLoss*, melhor desempenho obtido na competência C-I. A curva de treino apresenta crescimento consistente e estabilização após aproximadamente 20 épocas, enquanto a curva de validação exibe oscilações regulares em torno de valores entre 0,55 e 0,65, indicando estabilidade geral e ausência de *overfitting* pronunciado.

-0,10 e 0,55. Em contraste, a curva de treino segue trajetória muito mais suave. Essa discrepância evidencia a sensibilidade dessa competência ao ruído e às diferenças de distribuição entre treino e validação, tornando o *early stopping* fundamental para impedir que o modelo avance para regiões de *overfitting* e garantindo a escolha de um checkpoint com melhor capacidade de generalização.

A Figura V.4 mostra um processo de aprendizagem bastante estável para a competência C-IV. Treino e validação crescem rapidamente e atingem um platô elevado já nas primeiras épocas, mantendo-se praticamente constantes a partir desse ponto. A validação estabiliza-se em torno de 0,74 com flutuações mínimas, indicando que o modelo captura de forma consistente os elementos estruturais avaliados pela competência C-IV. A convergência precoce e a suavidade das curvas reforçam a adequação do *early stopping*, garantindo a seleção de um checkpoint maduro e altamente generalizável.

A Figura V.5 evidencia que o treinamento para a competência C-V apresenta um padrão de convergência rápido e altamente estável. Tanto treino quanto validação atingem valores elevados já nas primeiras épocas, e a validação permanece próxima de 0,75 com variações mínimas ao longo do restante do treinamento. Esse comportamento sugere que o BERTimbau-large captura de maneira consistente os elementos característicos da proposta de intervenção e confirma a adequação do *early*

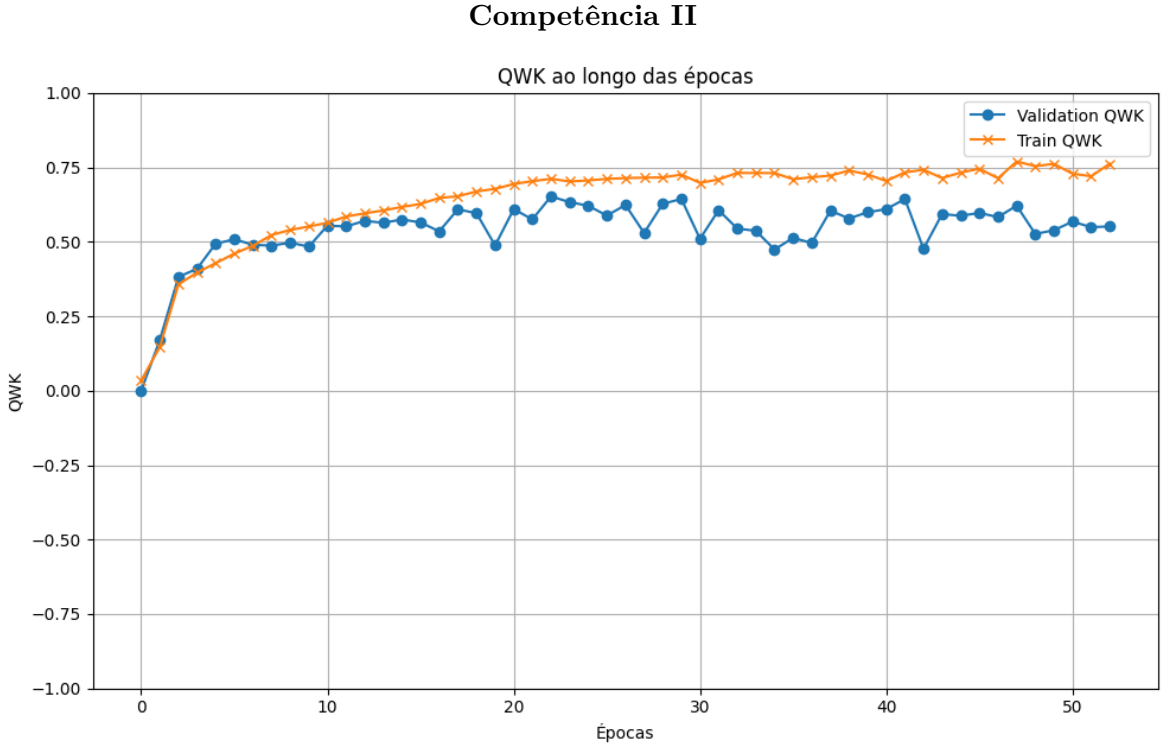


Figura V.2: Evolução da métrica QWK durante o treinamento do modelo XLM-RoBERTa-base com abordagem ordinal (WKLoss), melhor resultado obtido para a competência C-II. A curva de validação cresce rapidamente nas primeiras épocas e estabiliza-se em torno de valores entre 0,50 e 0,60, enquanto o treino segue trajetória ascendente até cerca de 0,75. A proximidade entre as curvas e a ausência de divergência acentuada indicam boa capacidade de generalização, justificando o checkpoint selecionado via *early stopping*.

stopping ao selecionar um ponto de convergência robusto.

De modo geral, as curvas de evolução da métrica QWK evidenciam comportamentos distintos entre as competências, refletindo a natureza heterogênea das dimensões avaliadas pela matriz do ENEM. Enquanto C-I, C-II e C-III apresentam maior variabilidade ao longo das épocas, as competências C-IV e C-V exibem trajetórias mais estáveis, com convergência rápida e flutuação reduzida. Esses padrões indicam que a aprendizagem das competências não ocorre de forma uniforme e que características estruturais de cada dimensão influenciam a dinâmica de treino dos modelos. Além disso, os gráficos reforçam a importância do *early stopping* como mecanismo para selecionar checkpoints estáveis e comparáveis, reduzindo o risco de *overfitting*.

V.5.2 Efeitos do tamanho do texto e limitação de contexto

Conforme descrito na Seção IV.2, os experimentos consideraram duas versões do conjunto de dados: o Completo e o Filtrado. O conjunto completo contempla todas as redações coletadas pelo *crawler*, incluindo aquelas que ultrapassam o limite de 512 *tokens* suportado pelos modelos baseados em BERT, sendo submetidas ao processo de truncamento. Já o conjunto filtrado é composto apenas por redações cujo comprimento total não excede esse limite, preservando integralmente o conteúdo

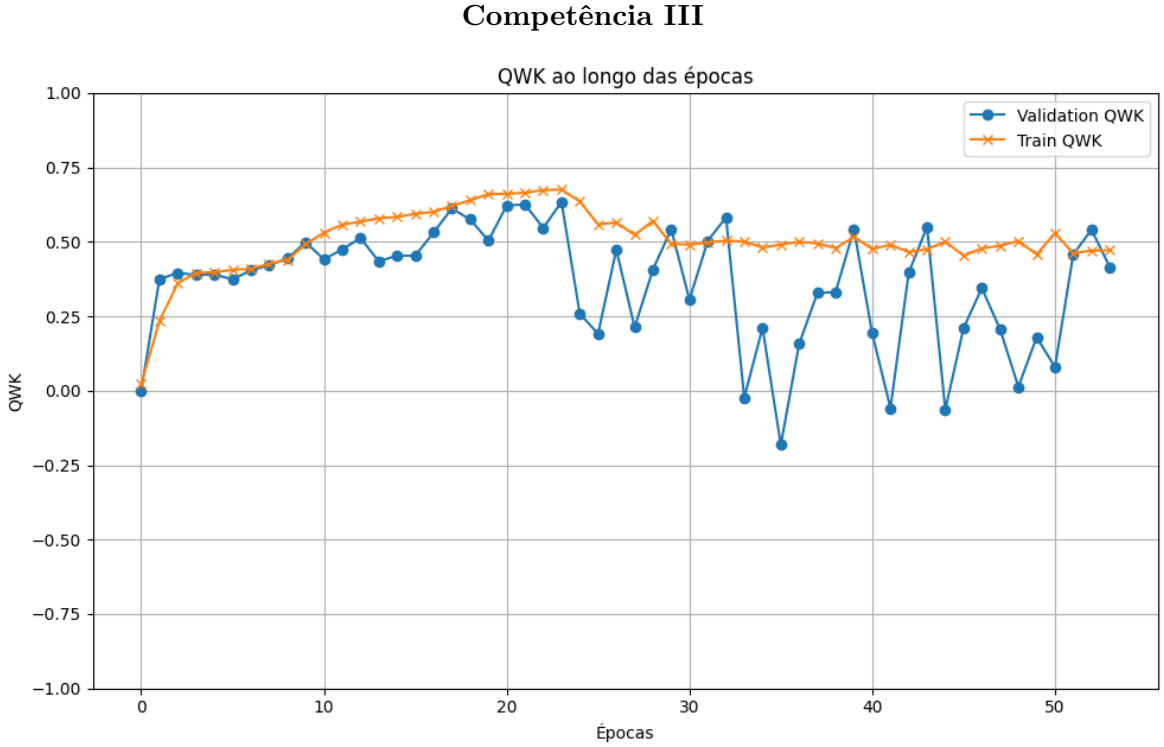


Figura V.3: Evolução da métrica QWK durante o treinamento do modelo XLM-RoBERTa-base com *WKLoss*, que obteve o melhor resultado na competência C-III. A curva de validação apresenta forte variabilidade ao longo das épocas, enquanto a curva de treino permanece mais estável, indicando maior sensibilidade dessa competência à variabilidade do conjunto de dados e reforçando a importância do *early stopping*.

textual de cada amostra. Essa distinção permitiu avaliar de forma controlada os efeitos da limitação de contexto sobre o desempenho dos modelos, analisando se a redução do tamanho máximo de entrada compromete a compreensão global do texto ou, ao contrário, favorece a estabilidade do treinamento ao eliminar ruídos decorrentes do truncamento.

De forma geral, os resultados mostram que o impacto da limitação de contexto varia conforme a natureza da competência avaliada. Conforme observado na Tabela V.4, nas competências C-I e C-II, associadas ao domínio da norma culta e compreensão da proposta, obtiveram-se os melhores desempenhos com o conjunto de dados completo, sugerindo que a possível perda de contexto por redações que ultrapassem o tamanho máximo de *tokens* não atrapalhou o desenvolvimento do modelo. Em contrapartida, as competências C-IV e C-V apresentaram ganhos expressivos com o uso do conjunto de dados filtrado, ao passo que C-III obteve melhor desempenho com o conjunto completo. Tal comportamento indica que, em tarefas mais localizadas, voltadas a mecanismos linguísticos, organização argumentativa e formulação de propostas de intervenção, a eliminação do ruído causado pelo truncamento e manutenção de sequências completas dentro do limite, favorecem consistência do aprendizado.

Além disso, observou-se que o modelo especializado em português se beneficiou mais do conjunto

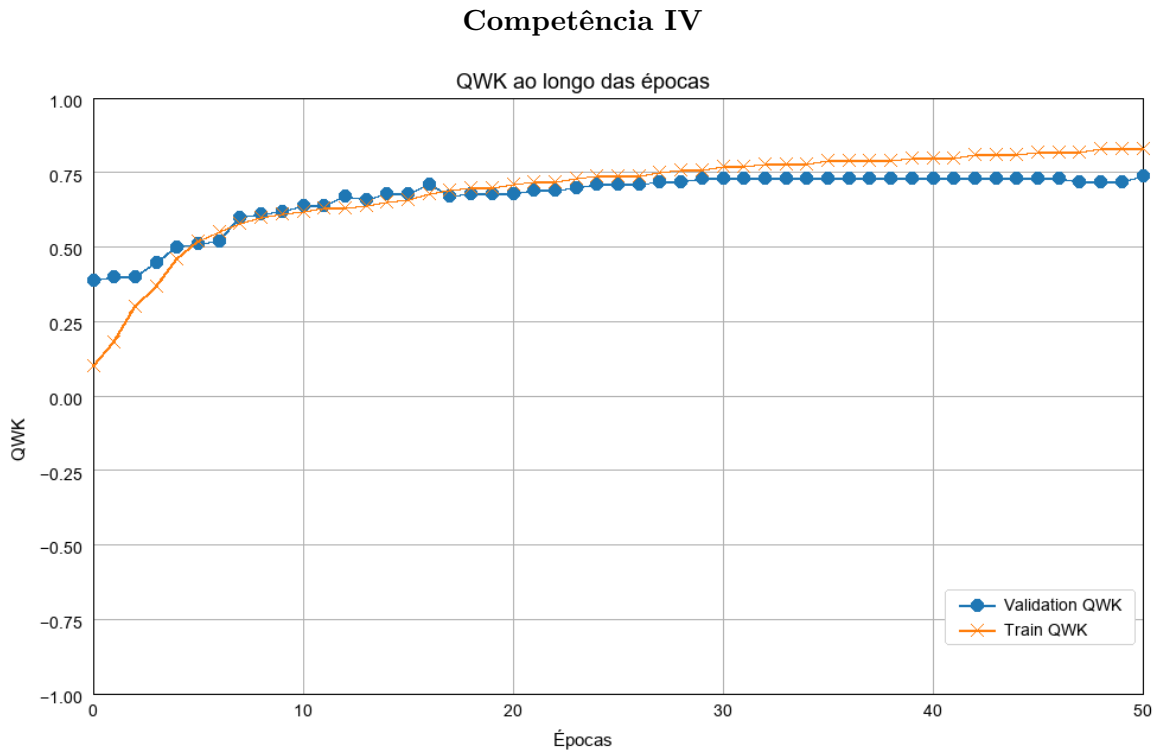


Figura V.4: Evolução da métrica QWK ao longo do treinamento do modelo BERTimbau-large com *fine-tuning* padrão, responsável pelo melhor desempenho na competência C-IV. As curvas exibem crescimento rápido e estabilização consistente, com pouca variabilidade ao longo das épocas.

de dados filtrado, enquanto o modelo multilíngue demonstrou maior robustez à perda parcial de contexto em textos longos. Em síntese, os resultados indicam que a escolha do limite de *tokens* não deve ser feita de forma arbitrária, uma vez que impacta diretamente a estabilidade do treinamento e se torna um fator crítico na modelagem de tarefas de avaliação automática de textos longos.

A avaliação das redações com nota máxima publicadas nos últimos cinco anos revelou que 46 das 47 produções apresentam comprimento superior a 512 tokens, com média de aproximadamente 672 tokens. Esse quantitativo reflete a densidade textual necessária para preencher de forma argumentativa o limite físico de 30 linhas exigido pelo exame. Consequentemente, constata-se que a janela de 512 *tokens* impõe uma restrição significativa de contexto, podendo suprimir trechos finais que são cruciais para a avaliação da argumentação e da proposta de intervenção.

Embora o conjunto filtrado tenha apresentado ganhos de desempenho em determinadas competências, é importante destacar que tal recorte não representa fielmente o cenário real do ENEM, no qual a redação é avaliada integralmente. A limitação de 512 *tokens* pode suprimir trechos relevantes, especialmente na conclusão, onde se concentram elementos decisivos para a Competência V.

Esse resultado indica que futuras investigações devem considerar arquiteturas capazes de processar textos mais longos ou estratégias como janelamento e agregação hierárquica, reduzindo o impacto do truncamento.

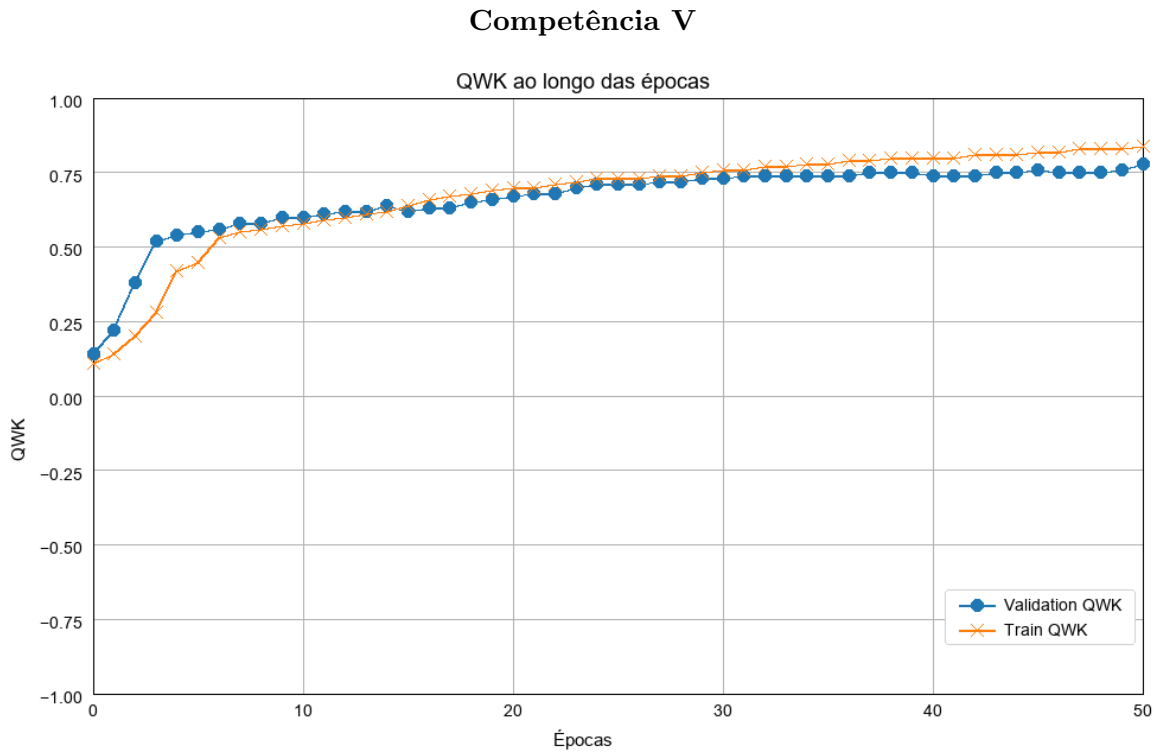


Figura V.5: Evolução da métrica QWK ao longo das épocas para o modelo BERTimbau-large com *fine-tuning* padrão, melhor resultado obtido na competência C-V. As curvas exibem convergência rápida e estabilização consistente da validação.

V.5.3 Comparação com abordagens de referência (*baseline*)

Para avaliar de forma abrangente os ganhos obtidos pelas arquiteturas baseadas em modelos de linguagem pré-treinados, os resultados são comparados com abordagens de referência clássica. Como baseline tradicional, utiliza-se uma abordagem com vetores TF-IDF combinados ao modelo Random Forest, amplamente empregada em estudos anteriores de AES.

Conforme apresentado nos resultados detalhados (Apêndice D), essa configuração apresentou valores de QWK entre 0,43 e 0,60, dependendo da competência. Embora eficiente e de baixo custo computacional, o método mostrou limitações em capturar nuances semânticas profundas, sendo superado pelas abordagens baseadas em modelos pré-treinados, que obtiveram valores superiores a 0,70 em múltiplas competências.

Além da melhora na precisão, os modelos pré-treinados mostraram maior consistência entre diferentes execuções. Ajustes pequenos na divisão dos dados ou nos hiperparâmetros não geraram variações grandes no desempenho, indicando boa capacidade de generalização. Esse comportamento mais estável ocorre porque esses modelos já partem de um conhecimento linguístico amplo, o que os torna menos sensíveis a mudanças superficiais nos dados durante o *fine-tuning*.

V.5.4 Respostas às Questões de Pesquisa

Com base nos resultados apresentados nas seções anteriores, esta subseção retoma as questões de pesquisa formuladas na Seção V.4.3, analisando as evidências empíricas obtidas para responder a cada uma delas. A discussão considera tanto o desempenho quantitativo (métricas de QWK) quanto os padrões qualitativos observados entre as competências e as estratégias de modelagem.

Questão	**Foco da Investigação**	**Evidências observadas**	**Resposta**
QP1	Impacto do aumento de escala	O modelo XLM-RoBERTa-large não figurou entre os melhores resultados, sendo frequentemente superado pelo XLM-RoBERTa-base e pelo BERTimbau-large. O aumento de parâmetros não se traduziu em ganhos diretos de desempenho nesta tarefa.	Negativa
QP2	Equilíbrio entre desempenho e custo	O BERTimbau-large obteve os melhores resultados absolutos nas competências C-IV e C-V com o conjunto filtrado (QWK até 0,748), com custo computacional inferior ao dos modelos multilíngues de maior porte nesses cenários.	Afirmativa
QP3	Efeito da formulação ordinal	As competências C-I, C-II e C-III obtiveram melhor desempenho com a WKLoss, que modela explicitamente a ordem entre as notas, superando o *fine-tuning* padrão de classificação categórica nessas dimensões.	Parcialmente afirmativa
QP4	Variação de desempenho entre competências	C-IV e C-V apresentaram os maiores valores de QWK, enquanto C-III mostrou maior instabilidade. A dificuldade de modelagem não segue apenas a divisão entre aspectos linguísticos e argumentativos, pois depende também da distribuição das notas e dos sinais disponíveis no texto.	Afirmativa
QP5	Impacto do conjunto filtrado ($\leq$ 512 tokens)	O conjunto filtrado favoreceu as competências C-III, C-IV e C-V, enquanto o conjunto completo manteve vantagem em C-I e C-II. O efeito do filtro depende da natureza da competência avaliada.	Condicional

Tabela V.5: Síntese das evidências e respostas às questões de investigação.

De modo geral, os experimentos forneceram evidências suficientes para responder às indagações centrais deste trabalho.

Em resposta à QP1, que questionava se o *XLM-RoBERTa-large* se beneficiaria do pré-treinamento

em larga escala para superar modelos menores, os resultados indicam uma resposta negativa. O modelo de menor porte (*XLM-RoBERTa-base*) apresentou melhor equilíbrio entre custo e desempenho, evidenciando que, para a tarefa de correção de redações no corpus utilizado, o aumento isolado no número de parâmetros não resulta necessariamente em ganhos de acurácia.

Quanto à QP2, a resposta é afirmativa. Os modelos especializados, particularmente o *BERTimbau*, mostraram-se altamente eficientes. O *BERTimbau-large* alcançou desempenho competitivo e, em casos como C-IV e C-V, superior aos multilíngues, com menor custo computacional. O modelo *AlBERTina*, embora consistente, obteve resultados inferiores ao *BERTimbau*, sugerindo que, apesar da vantagem da especialização no idioma, o porte do modelo e o volume de pré-treinamento continuam sendo fatores influentes.

Sobre a QP3, as evidências indicam que a formulação ordinal trouxe ganhos em parte das competências avaliadas. A abordagem com WKLoss mostrou-se superior à classificação categórica padrão especialmente nas competências C-I, C-II e C-III, o que sugere a utilidade de modelar a natureza ordenada das notas nesses cenários. Como esse ganho não ocorreu de forma uniforme em todas as competências, a resposta deve ser interpretada como parcialmente afirmativa.

Para a QP4, a resposta é afirmativa. Os resultados variaram entre as competências: C-IV e C-V apresentaram os maiores valores de QWK, enquanto C-III mostrou maior instabilidade. Isso indica que a dificuldade de modelagem não reside apenas na dicotomia entre aspectos "linguísticos"e "argumentativos", mas depende de características intrínsecas a cada competência, da distribuição das notas e dos sinais textuais disponíveis para o modelo.

Por fim, em relação à QP5, observa-se uma resposta condicional. O uso do conjunto filtrado ($\leq$ 512 tokens) resultou em melhor desempenho para C-III, C-IV e C-V. No entanto, o conjunto completo mostrou-se vantajoso para C-I e C-II, sugerindo que o maior volume de dados pode compensar parcialmente a perda de contexto em alguns cenários. Adicionalmente, notou-se que o modelo *XLM-RoBERTa-base* demonstrou maior robustez ao truncamento do que os modelos monolíngues.

Em síntese, os achados reforçam que a avaliação automática de redações em português não admite uma solução única: ela exige estratégias sensíveis às particularidades do idioma e à natureza específica de cada competência, onde o equilíbrio entre a arquitetura do modelo, a função de perda e a estratégia de tratamento dos dados se mostra decisivo.

V.5.5 Análise de Erros por Matrizes de Confusão.

Além das métricas agregadas de desempenho, as matrizes de confusão apresentadas nas Figuras V.6–V.10 fornecem informações adicionais sobre o comportamento dos modelos ao longo dos níveis de pontuação. Para cada competência, é reportada a matriz de confusão da configuração

que apresentou o melhor desempenho agregado para partir de uma das melhores configurações de modelo avaliadas. Em todas as matrizes, as linhas correspondem aos rótulos verdadeiros (notas atribuídas pelos avaliadores humanos) e as colunas correspondem às previsões produzidas pelo modelo.

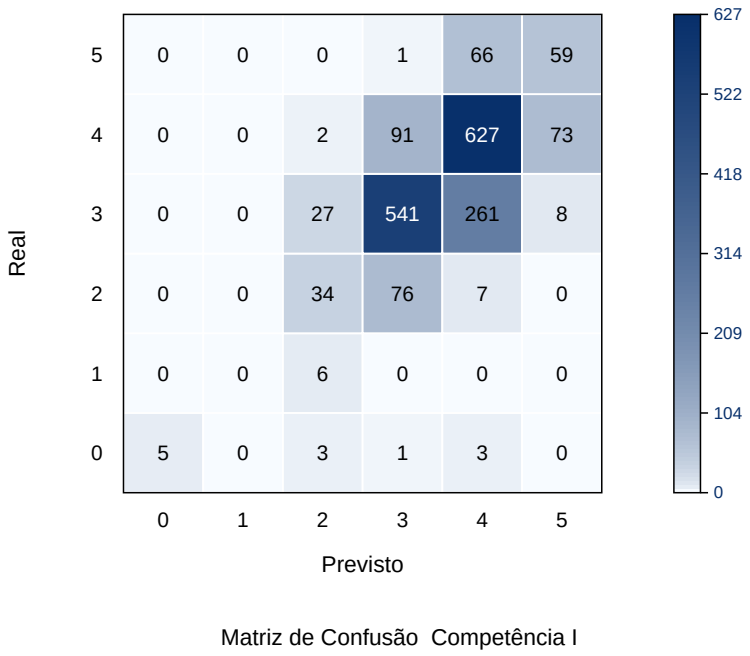


Figura V.6: Matriz de confusão representativa para a Competência I: domínio da modalidade escrita formal da língua portuguesa.

Para a Competência I (Figura V.6), a maior parte das classificações incorretas ocorre entre níveis de nota adjacentes, com concentração nas classes 3 e 4. Esse padrão indica que os modelos capturam adequadamente a qualidade gramatical global, mas apresentam dificuldade em distinguir diferenças mais sutis de precisão linguística, especialmente em redações com construções sintáticas complexas ou desvios que surgem de forma gradual ao longo do texto.

Para a Competência II (Figura V.7), as previsões concentram-se nas classes intermediárias (2–4), com a maior massa diagonal no rótulo 3 (503 amostras) e no rótulo 4 (634 amostras). Os erros são predominantemente de diferença unitária, indicando que o modelo distingue de forma confiável a adequação geral à proposta, mas apresenta dificuldade em capturar distinções mais finas quanto ao uso do repertório sociocultural.

Para a Competência III (Figura V.8), as matrizes de confusão revelam concentração das pre-

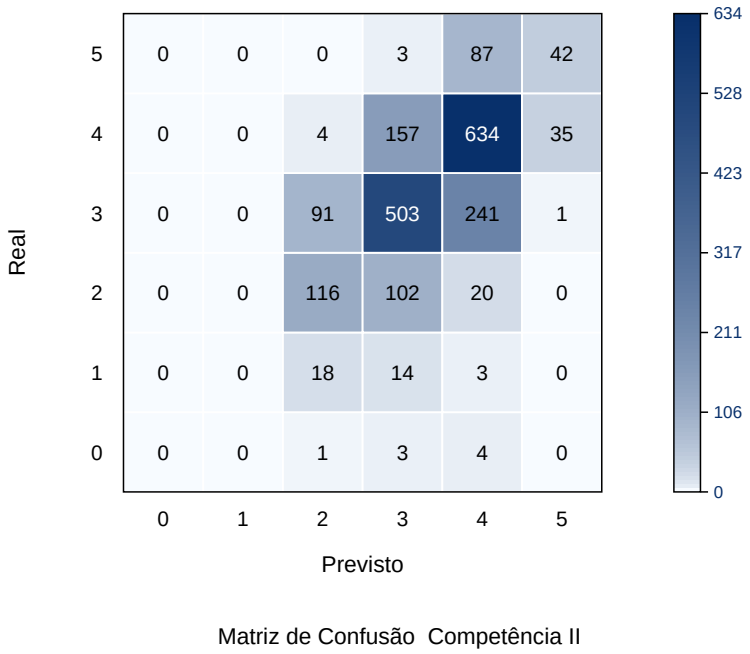


Figura V.7: Matriz de confusão representativa para a Competência II: compreensão da proposta de redação e aplicação de conceitos das áreas de conhecimento para o desenvolvimento do tema.

visões nos níveis intermediários de pontuação, com confusão frequente entre as notas 2, 3 e 4. Isso indica que os modelos detectam a presença de uma estrutura argumentativa, mas apresentam maior dificuldade em distinguir diferenças mais finas na organização e no desenvolvimento dos argumentos.

Para a Competência IV (Figura V.9), as classificações incorretas também se concentram em rótulos adjacentes nos níveis intermediários, com destaque para a confusão entre as notas 3 e 5. Como essa competência envolve o uso de mecanismos linguísticos para a construção da argumentação — dimensão que se apoia em sinais coesivos distribuídos ao longo do texto, e não em marcadores superficiais salientes —, esses erros sugerem que o modelo captura a adequação coesiva geral, mas subestima a coesão implícita, conforme ilustrado no Caso 2 apresentado a seguir.

Para a Competência V (Figura V.10), as previsões também se agrupam em torno das notas intermediárias, e os erros ocorrem com frequência entre as notas 3 e 4. Esses casos correspondem a redações que apresentam uma proposta de intervenção, mas variam quanto à completude e ao nível de detalhamento, sugerindo limitações na captura de todos os elementos estruturais exigidos pelos critérios de avaliação de pontuações mais altas.

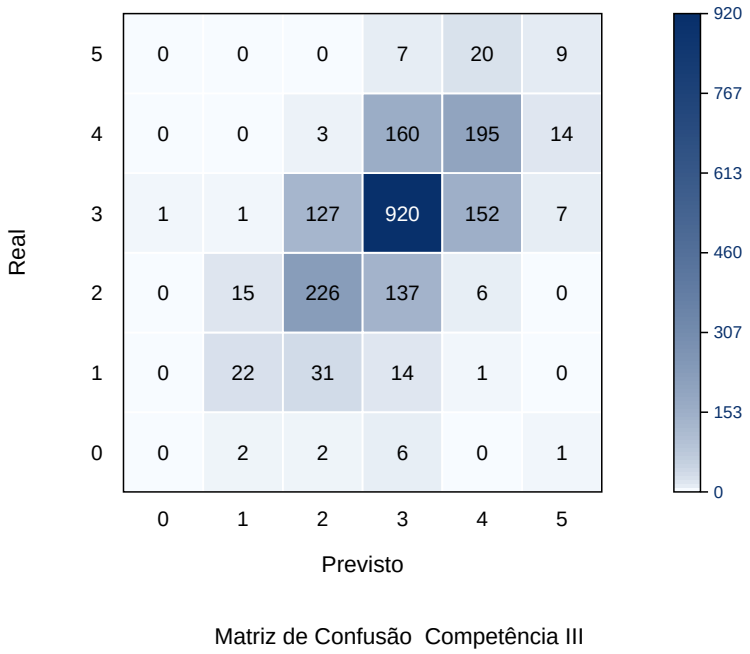


Figura V.8: Matriz de confusão representativa para a Competência III: seleção, relação, organização e interpretação de informações em defesa de um ponto de vista.

V.5.6 Análise de Casos

Para aprofundar a análise de erros, examinamos o comportamento dos modelos ao longo das faixas de pontuação. Os casos discutidos nesta seção foram selecionados como exemplos qualitativos de padrões observados nas matrizes de confusão e nas métricas agregadas; portanto, não devem ser interpretados como uma amostra estatisticamente representativa do conjunto de teste, mas como evidências ilustrativas de comportamentos recorrentes ou metodologicamente relevantes. Os resultados indicam maior concordância com os avaliadores humanos nas notas intermediárias, especialmente entre os níveis 2 e 4, ao passo que a concordância diminui nos extremos (notas 0-1 e 5). Esse padrão está alinhado com a distribuição do conjunto de dados, concentrada nos valores intermediários, o que favorece a calibração dos modelos nessas regiões. Em contraste, os casos extremos permanecem sub-representados e frequentemente envolvem características mais distintas, seja por deficiências linguísticas severas, seja por estruturas argumentativas altamente sofisticadas, tornando-os mais difíceis de modelar de forma consistente.

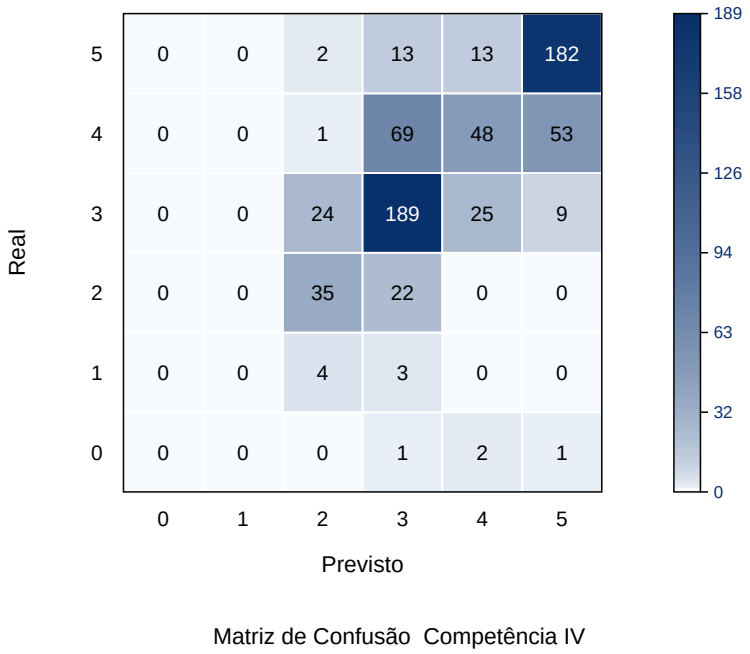


Figura V.9: Matriz de confusão representativa para a Competência IV: conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.

Caso 1 - Classificação incorreta em nível adjacente na Competência I. Neste ensaio, o modelo atribuiu nota 3 para a Competência I, enquanto o avaliador humano atribuiu nota 2, com predições corretas em todas as demais competências. Uma inspeção mais cuidadosa revela múltiplos desvios ortográficos e gramaticais, como *"questões dee nacionalidade"*, *"momento críticoo"* e *"nos últimos aanos"*. Esses erros indicam controle limitado da norma culta escrita, apesar de o texto manter coerência e legibilidade globais. Esse comportamento sugere que o modelo captura adequação estrutural e fluência global, mas falha em penalizar consistentemente erros linguísticos localizados. Tal padrão está alinhado com a análise da matriz de confusão, na qual a maioria das classificações incorretas ocorre entre níveis de nota adjacentes. O caso ilustra a sensibilidade do modelo à organização textual geral em detrimento da precisão gramatical fina.

Caso 2 - Subestimação na Competência IV. Neste caso, o modelo atribuiu nota 2 para a Competência IV, enquanto o avaliador humano atribuiu nota 3. A redação desenvolve uma narrativa causal articulando desastres naturais a fluxos migratórios e desafios sociais, como evidenciado em trechos como *"no ano de 2011 o Japão sofreu com um grande tragédia"* e *"em questão disso*

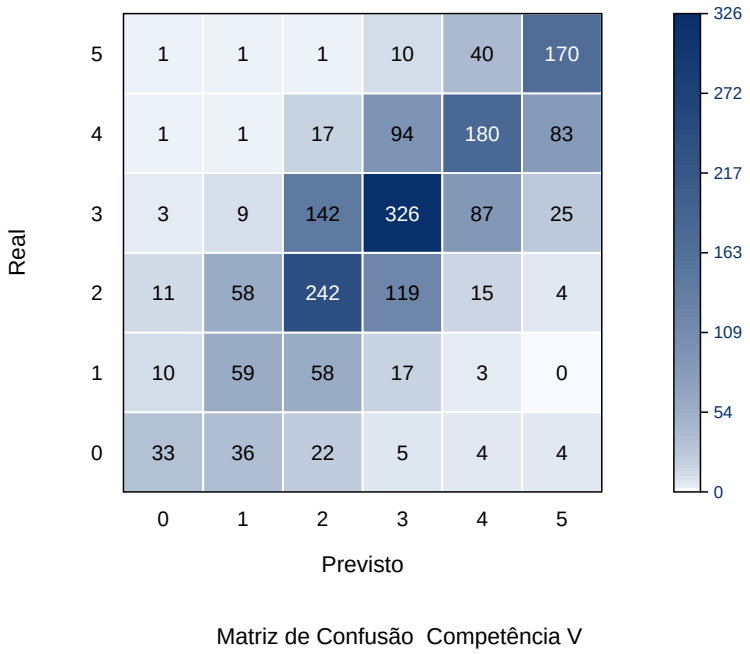


Figura V.10: Matriz de confusão representativa para a Competência V: elaboração de proposta de intervenção para o problema abordado, respeitando os direitos humanos.

muitas vezes saem para tentar uma vida melhor". No entanto, a coesão é frequentemente estabelecida de forma implícita, com uso limitado de marcadores discursivos explícitos. Adicionalmente, trechos como *"esse é um grande problema, é um pesadelo mais tragicamente a triste realidade do mundo"* revelam irregularidades sintáticas e problemas de pontuação que podem obscurecer a progressão argumentativa. Esse resultado sugere que o modelo tem dificuldade em capturar coesão distribuída globalmente quando ela não é fortemente sinalizada por conectores lexicais, o que leva à subestimação nesta competência.

Caso 3 - Superestimação expressiva na Competência V. Este caso apresenta uma discrepância substancial entre a nota prevista e a humana, especialmente na Competência V: o modelo atribuiu nota 5, enquanto o avaliador humano atribuiu nota 3. A redação inclui uma proposta de intervenção com elementos como *"substituição de uso de produtos químicos por produtos menos nocivos"* e *"maior divulgação de ONGs que reutilizam essas matérias"*. Embora esses elementos indiquem consciência acerca de possíveis soluções, são apresentados de forma genérica e carecem de especificação sobre implementação, agentes responsáveis e viabilidade operacional. A estrutura

de uma intervenção está presente, porém a proposta não atende plenamente aos critérios exigidos para notas mais altas, sobretudo em termos de especificidade e completude. O caso evidencia que o modelo tende a basear sua avaliação na presença de expressões relacionadas a soluções, em vez de avaliar a profundidade e adequação da intervenção, reforçando suas limitações na captura de requisitos discursivos de nível superior.

Caso 4 - Superestimação sistemática em redações de alta qualidade. Neste exemplo, o modelo atribui notas elevadas em múltiplas competências, incluindo notas máximas nas Competências II, IV e V, enquanto o avaliador humano atribui valores ligeiramente menores. A redação demonstra elementos estruturais sólidos, incluindo referências como *"Steve Jobs - fundador da Apple"* e *"o jornalista brasileiro Caco Bacellos diz que..."*, bem como uma proposta de intervenção clara: *"cabe ao governo [...] a criação de palestras nas escolas"*. No entanto, algumas limitações estão presentes, como construções repetitivas (e.g., *"é notório que"* e *"dessa forma"*) e profundidade relativamente limitada no desenvolvimento dos argumentos. O modelo aparenta priorizar a presença de marcadores estruturais e retóricos em detrimento de uma avaliação mais criteriosa da sofisticação argumentativa, o que resulta em leve superestimação.

Caso 5 - Pequena discrepância na organização argumentativa. Neste caso, as predições do modelo se aproximam das notas humanas, com pequeno desvio na Competência III. A redação apresenta uma estrutura argumentativa clara, especialmente em trechos como *"a falta de investimento em setores como a educação"* e *"com a ausência de trabalho gera consequências graves"*. No entanto, os argumentos são desenvolvidos de forma relativamente superficial, com uso limitado de evidências ou elaborações mais aprofundadas. O modelo identifica corretamente a presença de argumentação, mas superestima levemente sua qualidade, provavelmente em razão da dependência de pistas estruturais, como consistência temática e organização dos parágrafos. O caso reforça que a maioria dos erros do modelo ocorre na fronteira entre níveis de nota adjacentes, particularmente nas competências relacionadas ao desenvolvimento argumentativo.

Esses estudos de caso fornecem evidências concretas dos padrões de erro identificados anteriormente, ilustrando como as predições do modelo são influenciadas tanto por características superficiais quanto por propriedades linguísticas mais profundas.

V.6 Discussão e síntese dos Resultados

A análise dos resultados apresentados nas seções anteriores permite sintetizar as principais observações e implicações das abordagens testadas para a tarefa de avaliação automática de redações em língua portuguesa. De modo geral, os experimentos indicam que o uso de modelos de lingua-

gem pré-treinados, especialmente aqueles ajustados por meio de *fine-tuning* com a técnica LoRA, proporciona ganhos em relação às abordagens tradicionais baseadas em engenharia de atributos no corpus analisado. Os melhores valores por competência ficaram entre 0,642 e 0,748 de QWK, com resultados superiores a 0,70 nas competências C-IV e C-V. Esses valores sugerem concordância relevante com as notas humanas, embora não eliminem a necessidade de avaliação humana nem permitam afirmar equivalência plena com avaliadores especialistas.

Ao comparar as diferentes estratégias de modelagem, observou-se que a natureza da competência influencia diretamente a eficácia de cada abordagem. As competências C-I, C-II e C-III, associadas à norma culta, à compreensão da proposta e à estrutura argumentativa, apresentaram melhor desempenho com a classificação ordinal utilizando a função de perda WKLoss. Esse resultado indica que a preservação das relações de ordem entre as classes é essencial para representar a progressão gradual de qualidade textual, evitando penalizações desproporcionais para discrepâncias pequenas de nota. Outro aspecto relevante é que, nessas competências, o modelo XLM-RoBERTa-base superou o XLM-RoBERTa-large, que não apareceu entre os melhores resultados. Esse comportamento reforça a hipótese de que, para a tarefa de avaliação automática de redações em português, o aumento no número de parâmetros não se traduz necessariamente em ganhos de desempenho, possivelmente devido à natureza específica do idioma e ao tamanho do conjunto de dados. Já as competências C-IV e C-V, mais dependentes de construções linguísticas específicas e da formulação de uma proposta de intervenção, obtiveram melhor desempenho com o *fine-tuning* padrão, em especial quando aplicadas a modelos monolíngues como o *BERTimbau-large*. Esse resultado indica que, em competências que exigem maior sensibilidade às relações sintáticas e à fluidez do texto, modelos monolíngues tendem a captar com mais precisão as nuances linguísticas próprias do português.

Outro aspecto de destaque refere-se ao impacto do tamanho do texto e da limitação de contexto. Embora o uso de textos truncados possa eliminar informações discursivas relevantes, os resultados indicam que a filtragem de redações com até 512 *tokens* contribuiu para maior estabilidade no treinamento e melhor desempenho em competências sensíveis à organização textual e à presença de elementos finais da redação, como ocorre em C-III, C-IV e C-V. Por outro lado, em C-I e C-II, o conjunto completo apresentou resultados ligeiramente superiores, possivelmente porque o maior volume de dados compensou parcialmente a perda de contexto causada pelo truncamento. Essa diferença reforça a importância de considerar a extensão textual, a posição das informações relevantes e o limite de *tokens* como fatores estratégicos no desenho experimental de tarefas de AES.

A análise comparativa com as abordagens clássicas, baseadas em TF-IDF e modelos de regressão, indica vantagem das arquiteturas baseadas em *transformers* nas configurações avaliadas. Conforme apresentado no Apêndice D, a abordagem tradicional obteve QWK entre 0,43 e 0,60, enquanto as

melhores configurações com modelos pré-treinados alcançaram valores entre 0,642 e 0,748. Essa diferença sugere que representações contextualizadas são mais adequadas para capturar padrões linguísticos e discursivos relevantes neste corpus, ainda que a magnitude do ganho varie entre competências e dependa da arquitetura, da função de perda e do recorte do conjunto de dados.

V.6.1 Comparação com Trabalhos Relacionados

A Tabela V.6 situa os resultados obtidos neste trabalho no contexto mais amplo da correção automática de redações do ENEM. Uma comparação numérica direta entre os estudos é condicionada por três fatores: (i) os *corpora* diferem substancialmente em proveniência e tamanho, variando desde as 1.840 redações do UOL utilizadas por Amorim and Veloso [2017] até o corpus privado de 56.644 redações de Fonseca et al. [2018]; (ii) a escala de notas adotada por Amorim and Veloso [2017] compreende cinco níveis no intervalo 0,0-2,0, não sendo diretamente comparável à escala padrão do ENEM de 0-200 utilizada pelos demais estudos; e (iii) a qualidade das anotações varia entre os conjuntos, como evidenciado pelo QWK entre anotadores de 0,2-0,6 reportado por Silveira et al. [2024] para *corpora* no estilo Essay-BR. Considerando essas ressalvas, os modelos por competência desenvolvidos neste trabalho alcançaram valores de QWK entre 0,642 e 0,748, consistentemente superiores aos resultados por competência reportados por Silveira et al. [2024] em um corpus de proveniência comparável, e equivalentes ou superiores aos de **?** na Competência III.

O escore agregado obtido por Fonseca et al. [2018] permanece o maior QWK geral reportado na literatura (0,752); no entanto, sua dependência de um corpus privado com uma ordem de magnitude maior do que o utilizado neste trabalho limita a reprodutibilidade e a validade externa dessa comparação. Nesse contexto, os resultados aqui apresentados oferecem uma referência adicional para estudos alinhados às competências na avaliação automática de redações em português brasileiro com base em *transformers*, especialmente por reportarem resultados separados por competência e por abordagem.

De forma sintética, os resultados corroboram as hipóteses formuladas nesta pesquisa: (i) modelos multilíngues de grande porte, como o *XLM-RoBERTa*, não necessariamente apresentam desempenho superior aos modelos menores, indicando que a escala de parâmetros não é o principal fator determinante para a tarefa; (ii) modelos especializados em português, como o *BERTimbau* e o *AlBERTina*, mostraram-se mais eficientes e equilibrados quanto ao custo computacional e desempenho; (iii) abordagens avançadas, como a classificação ordinal e a modelagem multi-saída, trouxeram ganhos sobre os métodos convencionais; e (iv) o uso de conjuntos de dados filtrados pode favorecer a estabilidade e a coerência das predições, especialmente em competências que demandam análise estrutural mais detalhada.

Em síntese, os achados deste estudo indicam que a combinação entre *fine-tuning* eficiente (via

Trabalho	Corpus (Tamanho)	Melhor modelo	C1	C2	C3	C4	C5	Total
Amorim and Veloso [2017][a]	UOL (1.840)	RegLin + atributos	0,335	0,182	0,273	0,267	0,154	0,425
Fonseca et al. [2018][b]	Privado (56.644)	GBM	**0,676**	0,511	0,508	0,619	0,577	**0,752**
Silveira et al. [2024]	Essay-BR + UOL (3.586)	BERTimbau ordinal	0,290	0,370	0,470	0,420	0,500	–
Carvalho et al. [2024][c]	Essay-BR (6.565)	LGBM + atributos + BERT	–	–	0,588	–	–	–
Este trabalho	Brasil Escola (9.599)	XLM-R LoRA BERTimbau-L	0,659	**0,657**	**0,642**	**0,729**	**0,748**	0,690

Tabela V.6: Comparação dos valores de QWK com trabalhos anteriores sobre avaliação automática de redações do ENEM. São reportados os escores por competência (C1-C5) e o escore geral (Total) sob a melhor configuração de cada estudo. Os corpora, escalas de notas e protocolos de avaliação diferem entre os estudos; ver notas.

[a] Utiliza uma escala por competência de cinco níveis (0,0-2,0), e não a escala padrão do ENEM (0-200); os valores de QWK não são estritamente comparáveis aos dos demais estudos.

[b] Reporta resultados em corpus privado não disponível publicamente; o corpus é aproximadamente uma ordem de magnitude maior do que os utilizados nos estudos subsequentes.

[c] Avalia apenas a Competência III; as demais competências não são reportadas.

LoRA), modelagem por competência e funções de perda que respeitam a natureza ordinal das notas constitui uma estratégia promissora para a avaliação automática de redações em português. Os resultados obtidos ampliam as evidências disponíveis para pesquisas futuras, evidenciando o potencial das arquiteturas baseadas em *transformers* para apoiar práticas educacionais automatizadas, desde que consideradas suas limitações de contexto, dependência dos rótulos humanos e sensibilidade às características do corpus.

Capítulo VI Conclusão

Este capítulo apresenta uma síntese dos resultados obtidos no desenvolvimento e na avaliação de modelos para a correção automática de redações segundo competências inspiradas na matriz do ENEM. O objetivo central foi comparar diferentes abordagens de modelagem, incluindo modelos específicos por competência, arquiteturas multi-saída, estratégias em cascata, formulações ordinais e *fine-tuning* eficiente via LoRA, avaliando sua concordância com as notas humanas, seu custo computacional e suas limitações em cenários práticos.

Os principais avanços obtidos podem ser destacados da seguinte forma:

- **Coleta e organização de um corpus extenso e bem estruturado**, formado por 9.599 redações referentes a 102 temas. Após a raspagem, os textos passaram por etapas de limpeza, normalização de caracteres e padronização das entradas, garantindo consistência no conteúdo disponibilizado para os experimentos. O corpus resultante foi então organizado de maneira sistemática para ser utilizado diretamente nas análises e experimentações conduzidas nesta pesquisa;

- **Implementação e comparação de múltiplas abordagens de modelagem**, abrangendo desde técnicas tradicionais baseadas em TF-IDF até modelos de linguagem pré-treinados de grande porte. Os experimentos revelaram que modelos baseados em PLMs, como o XLM-RoBERTa com LoRA, apresentaram desempenho substancialmente superior, confirmando a hipótese de que representações contextualizadas capturam melhor nuances linguísticas;

- **Análise detalhada das estratégias específicas por competência**, evidenciando que modelos especializados tendem a capturar fenômenos discursivos com maior precisão em competências como C-III e C-IV. Esse comportamento confirma parcialmente a hipótese de que a especialização por competência melhora o desempenho, embora modelos multi-saída se mostrem competitivos em C-I e C-II;

- **Avaliação da técnica LoRA**, que demonstrou ser eficaz para reduzir significativamente o custo computacional do fine-tuning, preservando desempenho equivalente ao ajuste completo dos modelos. Essa evidência destaca a relevância da adaptação eficiente para viabilizar modelagens mais complexas mesmo em ambientes com recursos limitados.

Apesar dos avanços significativos, algumas limitações e pontos de aprimoramento foram identificados:

- **Dependência da diversidade do corpus**, uma vez que o conjunto utilizado não representa plenamente a heterogeneidade sociolinguística das redações reais do ENEM, o que limita a generalização dos modelos;
- **Restrição imposta pelos limites de contexto dos Transformers**, que exige truncamento dos textos e pode comprometer competências cuja avaliação depende de organização global;
- **Ausência do tema e dos textos motivadores como entrada dos modelos**, uma vez que os experimentos utilizaram apenas o texto da redação. Essa restrição limita a avaliação automática de aspectos que dependem da relação entre a produção textual e a proposta original, especialmente a aderência ao tema e a adequação da proposta de intervenção ao problema abordado;
- **Dependência de rótulos produzidos por avaliação humana**, pois as notas utilizadas refletem decisões de correção orientadas por critérios institucionais e pedagógicos inspirados na matriz do ENEM. Assim, os modelos aprendem a reproduzir padrões presentes no corpus, mas não eliminam possíveis variações de julgamento, vieses avaliativos ou diferenças entre corretores;
- **Desempenho desigual entre competências**, refletindo a maior complexidade de competências discursivas e pragmáticas, que continuam desafiadoras mesmo para PLMs;

Portanto, as conclusões parciais indicam que os resultados obtidos são promissores, evidenciando que modelos de linguagem pré-treinados, combinados com técnicas eficientes como LoRA, são capazes de aproximar o desempenho computacional daquele observado em avaliadores humanos, especialmente em competências com mais estrutura linguística e argumentativa. Ao mesmo tempo, os resultados das hipóteses mostram que modelos específicos por competência e abordagens ordinais oferecem ganhos importantes, embora não universais, e que arquiteturas em cascata não se mostraram adequadas para este tipo de tarefa.

Esses achados reforçam tanto o potencial quanto as limitações do uso de modelos de linguagem para a avaliação automática de redações no contexto brasileiro. Para avançar, as próximas etapas desta pesquisa deverão incluir:

- Ampliação curada do conjunto de dados, com a captura de novas redações da fonte utilizada anteriormente e com mecanismos adicionais de controle de qualidade, procedência e possível

contaminação por textos gerados ou auxiliados por inteligência artificial generativa. Essa ampliação deve considerar que mais dados não implicam necessariamente melhor desempenho caso introduzam vieses, duplicações ou padrões textuais pouco representativos da escrita humana avaliada no contexto do ENEM.

- Investigação de estratégias alternativas para lidar com redações que ultrapassam o limite de contexto dos modelos, como janelamento, agregação hierárquica ou mecanismos de seleção e resumo de trechos. Essas abordagens podem permitir o uso de informações distribuídas ao longo de todo o texto sem depender exclusivamente do truncamento ao final da redação.

- Integração dos modelos em ferramentas educacionais, por meio da evolução da plataforma Lumen (ver Apêndice E), já contemplada no edital *CAPES/InovaEducação*[1]. Essa iniciativa permitirá testar os modelos em contextos reais de ensino, envolvendo docentes e estudantes, e avaliar aspectos como aderência pedagógica, clareza dos feedbacks gerados e potencial de uso em larga escala.

As limitações observadas neste estudo decorrem predominantemente de restrições computacionais e não de impedimentos teóricos das abordagens analisadas. Trabalhos futuros podem explorar modelos de maior escala e arquiteturas mais complexas, avaliando o custo-benefício sob a perspectiva de aplicação institucional em larga escala.

Em suma, os avanços alcançados até aqui demonstram a viabilidade e o potencial da utilização de modelos de linguagem pré-treinados para a avaliação automática de redações segundo as competências do ENEM, ao passo em que evidenciam desafios metodológicos e práticos que precisam ser enfrentados.

[1] `https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/educacao-a-distancia/editais-uab/edital-no-3-2025-inovaeducacao`

Referências Bibliográficas

Matthew Kearney, Reuben Binns, and Yarin Gal. Language models change facts based on the way you talk, 2025. URL `https://arxiv.org/abs/2507.14238`. 2

Vanessa Bates Ramirez. Chatgpt is changing the words we use in conversation, 2025. URL `https://www.scientificamerican.com/article/chatgpt-is-changing-the-words-we-use-in-conversation`. 2

Mark D. Shermis and Jill Burstein, editors. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, 2013. 3, 9

Ellis Batten Page. The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5):238–243, 1966. 3, 7, 14

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284, 1998. doi: 10.1080/01638539809545028. 3

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. Automated scoring using a hybrid feature identification technique. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210, 1998. 3, 7

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring. `https://kaggle.com/competitions/asap-aes`, 2012. Kaggle. 3, 7

Evelin Amorim and 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*, pages 94–102, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/E17-4010. 3, 7, 8, 9, 19, 25, 28, 68, 69

Elijah Mayfield and Alan W Black. Should you fine-tune bert for automated essay scoring? In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 151–162, 2020. 3, 7, 17

Austin Pack, Alex Barrett, and Juan Escalante. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234, 2024. 4

Peter Foltz, Darrell Laham, and Thomas Landauer. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), 1999. 7, 14

Semire Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006. 7

C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee. A comprehensive review of automated essay scoring (aes) research and development. *Pertanika Journal of Science and Technology*, 29(3): 1875–1899, 2021. URL `www.scopus.com`. 7

INEP. A redação do enem: cartilha do(a) participante, 2025. URL `https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2025_cartilha_do_participante.pdf`. Acesso em: 12 mar. 2025. 8, 9

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 10, 12, 15

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 10, 12, 15

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017. 10, 11, 13, 15

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237, 2018. 10, 11, 13

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 11, 13, 16

Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 11, 12

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. 12

Suzana Ilić, Edison Marrese-Taylor, Jorge A. Balazs, and Yutaka Matsuo. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, Brussels, Belgium, 2018. 13

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 13, 17, 49

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 13, 16, 19

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 180–189, 2011. 14

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. A flexible approach to automated essay scoring using bayesian optimization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, 2015a. 14, 19

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 14, 19

Masaki Uto. Neural ordinal regression for automated essay scoring. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2374–2384, 2020. 14, 18, 19

Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. Technical report, Educational Testing Service, 2006. 14

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 15

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 15

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society Series B*, 42(2):109–142, 1980. 15

Peter Phandi, Ngoc Pham, and Hwee Tou Ng. Flexible hybrid models for ordinal regression. In *Proceedings of ACL*, pages 153–158, 2015b. 15

Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. 15, 16, 19

Kaggle asap automated essay scoring competition. `https://www.kaggle.com/competitions/asap-aes`, 2012. 15

Domenic V. Cicchetti and Alvan R. Feinstein. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558, 1990. 15

Courtney Napoles, Aoife Cahill, and Nitin Madnani. Rethinking evaluation metrics for aes: A study on human–machine agreement. *Proceedings of BEA Workshop*, 2016a. 16

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 17

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 17

Jordi de La Torre, Domenec Puig, and Aida Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018. 18

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 19

Fábio Souza, Rodrigo Nogueira, and 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. 19, 25, 26, 49

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. Correcting the correctors: Quantifying human errors in grammatical error annotation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–135, 2016b. 19

Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of EMNLP*, 2016. 19

Rogério F de Sousa, Jeziel C Marinho, Francisco AR Neto, Rafael T Anchiêta, and Raimundo S Moura. Piln at propor: A bert-based strategy for grading narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 10–13, 2024. 21, 24, 25, 28

Pei Yee Liew and Ian KT Tan. On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, pages 204–211, 2024. 21, 24, 25

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL `https://www.bmj.com/content/372/bmj.n71`. 22, 23

Aluizio Haendchen Filho, Hércules A do Prado, Edilson Ferneda, and Jonathan Nau. An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, 126:788–797, 2018. 25, 28

Jeziel C Marinho, Fábio Cordeiro, Rafael T Anchiêta, and Raimundo S Moura. Automated essay scoring: An approach based on enem competencies. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 49–60. SBC, 2022. 25

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17, 2024. 25, 28

Rodrigo Santos, João Rodrigues, Luís Gomes, João Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family, 2024. 25

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.141. URL `https://aclanthology.org/2020.findings-emnlp.141`. 25, 28, 33

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024*, pages 228–237. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.propor-1.23. 25, 28, 68, 69

Aluizio Filho, Fernando Concatto, Hércules Antonio Do Prado, and Edilson Ferneda. Comparing Feature Engineering and Deep Learning Methods for Automated Essay Scoring of Brazilian National High School Examination:. In *Proceedings of the 23rd International Conference on Enterprise Information Systems*, pages 575–583. SCITEPRESS - Science and Technology Publications, 2021. ISBN 978-989-758-509-8. doi: 10.5220/0010377505750583. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010377505750583. 25, 26, 28

Joyce M Silva, Rafael T Anchiêta, Rogério F de Sousa, and Raimundo S Moura. Investigating methods to detect off-topic essays. In *Brazilian Conference on Intelligent Systems*, pages 346–357. Springer, 2024. 25, 26, 28

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. Automatically Grading Brazilian Student Essays. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, *Computational Processing of the Portuguese Language*, volume 11122, pages 170–179. Springer International Publishing, 2018. ISBN 978-3-319-99721-6 978-3-319-99722-3. doi: 10.1007/978-3-319-99722-3_18. URL https://link.springer.com/10.1007/978-3-319-99722-3_18. 25, 26, 28, 68, 69

Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. Essay-BR: A Brazilian Corpus of Essays. In *Anais Do III Dataset Showcase Workshop (DSW 2021)*, pages 53–64, Brasil, October 2021. Sociedade Brasileira de Computação. doi: 10.5753/dsw.2021.17414. 26

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018. 32

Jordi de la Torre, Domenec Puig, and Aida Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2017.05.018. URL https://www.sciencedirect.com/science/article/pii/S0167865517301666. Machine Learning and Applications in Artificial Intelligence. 38, 39, 40

Eibe Frank and Mark Hall. A simple approach to ordinal classification. volume 2167, pages 145–156, 08 2001. ISBN 978-3-540-42536-6. doi: 10.1007/3-540-44795-4_13. 39

Minjun Chen, Yue Wu, Byron Wingerd, Zhichao Liu, Joshua Xu, Shraddha Thakkar, Thomas J Pedersen, Tom Donnelly, Nicholas Mann, Weida Tong, et al. Automatic text classification of drug-induced liver injury using document-term matrix and xgboost. *Frontiers in Artificial Intelligence*, 7:1401810, 2024. 40

Alberto et al. Santos. Albertina: A lightweight portuguese language model. *arXiv preprint arXiv:2401.00000*, 2024. 49

Ruan Carvalho, Lucas Lins, Luiz Rodrigues, Péricles Miranda, Hilario Oliveira, Thiago Cordeiro, Ig Bittencourt, Seiji Isotani, and Rafael Mello. *Exploring NLP and Embedding for Automatic Essay Scoring in the Portuguese*, pages 228–233. 07 2024. ISBN 978-3-031-64314-9. doi: 10.1007/978-3-031-64315-6_18. 69

Apêndice A Estrutura dos Dados

Este apêndice apresenta a estrutura do dado obtido na etapa de coleta de dados, como resultado da estruturação inicial dos dados obtidos pelo script de crawler desenvolvido para extrair informações do portal *Brasil Escola*. O *script* percorre as páginas do site, identifica os textos das redações e coleta metadados associados, como título, tema, notas atribuídas, comentários dos avaliadores e justificativas por competência.

O resultado da coleta é estruturado em formato `JSON`, de forma que cada redação corresponde a um objeto contendo os campos listados a seguir.

- **tema:** Dado textual que representa o título do tema da redação proposto pelo portal *Brasil Escola*.
- **texto:** Dado textual que armazena o conteúdo integral da redação escrita pelo estudante, sem comentários ou anotações.
- **texto_comentado:** Dado textual contendo o mesmo texto da redação acrescido dos comentários e observações do corretor, inseridos entre parênteses ao longo do texto.
- **comentarios:** Dado textual que resume o feedback final do corretor sobre o desempenho do estudante, com sugestões gerais de melhoria.
- **nota:** Dado numérico que representa a nota total atribuída à redação, variando de 0 a 1000 pontos.
- **titulo:** Dado textual que corresponde ao título atribuído pelo próprio autor à redação.
- **id:** Dado numérico que identifica unicamente a redação dentro do portal *Brasil Escola*.
- **link:** Dado textual que contém a URL da redação no portal, permitindo acesso direto ao conteúdo original.
- **competencias:** Estrutura do tipo lista (`array`) composta por cinco objetos, cada um representando uma das competências avaliadas no modelo do ENEM. Cada elemento da lista contém:

  - **competencia:** Dado textual que identifica a competência avaliada (Conforme as competências exibidas na Tabela IV.3).
  - **nota:** Dado numérico entre 0 e 200 que representa a pontuação atribuída a essa competência.
  - **motivo:** Dado textual que descreve o nível de desempenho do estudante e a justificativa da nota atribuída.

- **contexto_tema:** Dado textual que armazena o enunciado completo do tema proposto pelo portal, incluindo a contextualização e a instrução para produção da redação.

Apêndice B Exemplo de Registro

Este apêndice apresenta um exemplo completo de registro recuperado pelo *crawler*, no formato `JSON`. Cada objeto representa uma redação coletada no portal *Brasil Escola*, contendo o texto integral, comentários do avaliador, notas atribuídas, justificativas e informações contextuais do tema.

Apêndice C Exemplo de Registro

Este apêndice apresenta um exemplo completo de registro recuperado pelo *crawler*, no formato `JSON`. Cada objeto representa uma redação coletada no portal *Brasil Escola*, contendo o texto integral, comentários do avaliador, notas atribuídas, justificativas e informações contextuais do tema.


```
{
    "tema": "Insegurança alimentar no Brasil",
    "texto": "Na Idade Média, o crescimento populacional e as adversidades
climáticas acarretaram um período de intensa instabilidade alimentar na
   Europa,
onde milhares de pessoas morreram de inanição. Esse evento catastrófico
   ficou
conhecido como a Grande Fome. Atualmente, por mais que o campo agropecuá
   rio tenha
encontrado maior equilíibrio e consiga oferecer uma quantidade de
   alimentos
suficientes ao bem-estar da população, a preocupação com a fome e a
desnutrição ainda compõem a realidade de milhares de brasileiros.",
    "comentarios": "Otimas discussões. Mantenha os aspectos positivos.
Não deixe de exercitar a sua escrita.",
    "nota": "880.0",
    "titulo": "A Grande Fome",
    "id": "17919",
    "link": "https://vestibular.brasilescola.uol.com.br/banco-de-redacoes/
       17919",
    "competencias": [
        {
            "competencia": "Domínio da modalidade escrita formal",
            "nota": "160",
            "motivo": "Nível 4 - Demonstra bom domínio da modalidade
               escrita
formal da língua portuguesa e de escolha de registro, com poucos desvios
```

```
gramaticais e de convenções da escrita."
        },
        {
            "competencia": "Compreender a proposta e aplicar conceitos das
                várias áreas de conhecimento para desenvolver o texto
                dissertativo-argumentativo em prosa",
            "nota": "200",
            "motivo": "Nível 5 - Desenvolve o tema por meio de argumentaçã
                o
consistente, a partir de um repertório sociocultural produtivo e apresenta
excelente domínio do texto dissertativo-argumentativo."
        },
        {
            "competencia": "Selecionar, relacionar, organizar e
                interpretar
informações em defesa de um ponto de vista",
            "nota": "160",
            "motivo": "Nível 4 - Apresenta informações, fatos e opiniões
relacionados ao tema, de forma organizada, com indícios de autoria, em
    defesa
de um ponto de vista."
        },
        {
            "competencia": "Conhecimento dos mecanismos linguísticos
                necessários
para a construção da argumentação",
            "nota": "200",
            "motivo": "Nível 5 - Articula bem as partes do texto e
                apresenta
repertório diversificado de recursos coesivos."
        },
        {
            "competencia": "Proposta de intervenção com respeito aos
                direitos
humanos",
            "nota": "160",
            "motivo": "Nível 4 - Elabora bem proposta de intervenção
relacionada ao tema e articulada à discussão desenvolvida no texto."
        }
    ],
```

```
    "contexto_tema": "Insegurança alimentar é a condição de fome ou a
ausência de acesso a uma alimentação rica em nutrientes."
}
```

Apêndice D Resultados detalhados por abordagem e competência

Este apêndice apresenta, de forma complementar aos resultados discutidos no Capítulo V, as tabelas completas de desempenho obtidas para cada uma das cinco competências avaliadas no ENEM. Os experimentos aqui descritos visam oferecer uma visão mais abrangente das diferentes abordagens de modelagem testadas, bem como das variações de configuração adotadas em cada cenário.

Para cada competência, são comparados os principais modelos e estratégias de ajuste aplicados, considerando métricas de concordância ponderada (QWK) e configurações experimentais como taxa de aprendizado, tamanho do *batch* e número de épocas de treinamento.

As tabelas estão organizadas segundo o tipo de conjunto de dados utilizado durante o treinamento, distinguindo-se entre duas versões principais:

- **Conjunto de dados completo:** inclui todas as redações disponíveis na base original, independentemente do comprimento textual. Essa versão reflete um cenário mais amplo e realista, permitindo que o modelo processe textos de diferentes tamanhos, mas também introduz maior variabilidade e risco de truncamento em sequências longas.

- **Conjunto de dados filtrado ($\leq$ 512 tokens):** considera apenas as redações cujo comprimento total não ultrapassa o limite máximo suportado pelos modelos de linguagem pré-treinados. Essa filtragem busca garantir que cada redação seja processada integralmente, preservando o contexto sem cortes de sequência e favorecendo a consistência semântica das representações aprendidas.

A seguir as comparações dos melhores resultados de todas as abordagens implementadas para cada competência avaliada.

D.1 Competência 1 - Domínio da modalidade escrita formal

Abordagem	QWK	Modelo	Tipo de dataset	Principais configurações
Classificação ordinal com *WKLoss* (A7)	**0,659**	XLM-RoBERTa-base	Completo	70 épocas; *batch* = 22; *lr* = 1e-5
Fine-tuning padrão (A6)	0,649	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída (A4)	0,633	XLM-RoBERTa-large	Completo	150 épocas; *batch* = 12; *lr* = 2e-5
Classificação multi-saída com *WKLoss* (A5)	0,510	BERTimbau-large	Filtrado	25 épocas; *batch* = 10; *lr* = 2e-5
Regressão tradicional (*TF-IDF + XGBoost*) (A9)	0,434	XLM-RoBERTa-base	Completo	15 épocas; *batch* = 14; *lr* = 1e-5
Classificação sem *fine-tuning* (A8)	0,112	BERTimbau-large	Completo	70 épocas; *batch* = 15; *lr* = 2e-5
Classificação ordinal em cascata (A3)	0,046	XLM-RoBERTa-base	Completo	50 épocas; *batch* = 15; *lr* = 1e-5

Tabela D.1: Desempenho das principais abordagens testadas para a competência C1 (Domínio da modalidade escrita formal da língua portuguesa). A tabela apresenta os valores da métrica QWK obtidos por diferentes modelos e tipos de dataset, destacando suas principais configurações experimentais. A abordagem de classificação ordinal com *WKLoss*, utilizando o modelo XLM-RoBERTa-base, obteve o melhor resultado (QWK = 0.659), evidenciando que a modelagem ordinal favorece a consistência na atribuição das notas e a captura de padrões gramaticais e ortográficos característicos dessa competência.

D.2 Competência 2 - Compreender a proposta e aplicar conceitos das várias áreas de conhecimento para desenvolver o texto dissertativo-argumentativo em prosa

Abordagem	QWK	Modelo	Tipo de dataset	Principais configurações
Classificação ordinal com *WKLoss* (A7)	**0.657**	XLM-RoBERTa-base	Completo	70 épocas; *batch* = 22; *lr* = 1e-5
Classificação multi-saída (A4)	0.628	XLM-Roberta-large	Completo	150 épocas; *batch* = 10; *lr* = 2e-5
Fine-tuning padrão (A6)	0.617	XLM-RoBERTa-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída com *WKLoss* (A5)	0.598	XLM-RoBERTa-large	Completo	150 épocas; *batch* = 12; *lr* = 2e-5
Regressão tradicional (*TF-IDF + XGBoost*) (A9)	0.498	XLM-RoBERTa-base	Completo	15 épocas; *batch* = 14; *lr* = 1e-5
Classificação sem *fine-tuning* (A8)	0.154	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação ordinal em cascata (A3)	0.048	XLM-RoBERTa-base	Completo	50 épocas; *batch* = 15; *lr* = 1e-5

Tabela D.2: Desempenho das principais abordagens testadas para a competência C2 (Compreensão da proposta e aplicação de conceitos das várias áreas do conhecimento para desenvolver o texto dissertativo-argumentativo em prosa). A tabela apresenta os valores da métrica QWK obtidos por diferentes modelos e tipos de dataset, destacando suas principais configurações experimentais. A abordagem de classificação ordinal com *WKLoss*, utilizando o modelo XLM-RoBERTa-base, apresentou o melhor desempenho (QWK = 0.657), indicando que o uso de uma função de perda ordinal contribui para previsões mais consistentes em relação à escala de notas dessa competência.

D.3 Competência 3 - Selecionar, relacionar, organizar e interpretar informações em defesa de um ponto de vista

Abordagem	QWK	Modelo	Tipo de dataset	Principais configurações
Classificação ordinal com *WKLoss* (A7)	**0.642**	XLM-RoBERTa-base	Completo	70 épocas; *batch* = 22; *lr* = 1e-6
Classificação multi-saída (A4)	0.610	XLM-RoBERTa-large	Completo	150 épocas; *batch* = 12; *lr* = 2e-5
Fine-tuning padrão (A6)	0.606	XLM-RoBERTa-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída com *WKLoss* (A5)	0.501	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Regressão tradicional (*TF-IDF + XGBoost*) (A9)	0.467	XLM-RoBERTa-base	Completo	15 épocas; *batch* = 14; *lr* = 1e-5
Classificação sem *fine-tuning* (A8)	0.187	XLM-RoBERTa-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação ordinal em cascata (A3)	0.071	XLM-RoBERTa-base	Completo	50 épocas; *batch* = 15; *lr* = 1e-5

Tabela D.3: Desempenho das principais abordagens testadas para a competência C3 (Seleção, organização e interpretação de informações em defesa de um ponto de vista). A tabela apresenta os valores da métrica QWK obtidos por diferentes modelos e tipos de dataset, destacando suas principais configurações experimentais. Observa-se que a abordagem de classificação ordinal com *WKLoss*, utilizando o modelo XLM-RoBERTa-base, obteve o melhor desempenho (QWK = 0.642).

D.4 Competência 4 - Conhecimento dos mecanismos linguísticos necessários para a construção da argumentação

Abordagem	QWK	Modelo	Tipo de dataset	Principais configurações
Fine-tuning padrão (A6)	**0.729**	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída (A4)	0.726	BERTimbau-large	Completo	150 épocas; *batch* = 12; *lr* = 2e-5
Regressão tradicional (*TF-IDF + XGBoost*) (A9)	0.609	XLM-RoBERTa-base	Completo	15 épocas; *batch* = 14; *lr* = 1e-5
Classificação ordinal com *WKLoss* (A7)	0.606	XLM-RoBERTa-large	Completo	70 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída com *WKLoss* (A5)	0.595	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação sem *fine-tuning* (A8)	0.096	XLM-RoBERTa-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação ordinal em cascata	0.048	XLM-RoBERTa-base	Completo	50 épocas; *batch* = 15; *lr* = 1e-5

Tabela D.4: Desempenho das principais abordagens testadas para a competência C4 (Conhecimento dos mecanismos linguísticos necessários para a construção da argumentação). A tabela apresenta os valores da métrica QWK obtidos por diferentes modelos e tipos de dataset, destacando suas principais configurações experimentais. Observa-se que o *fine-tuning* padrão com o modelo BERTimbau-large, utilizando dados filtrados ($\leq$ 512 tokens), obteve o melhor resultado geral (QWK = 0.729), superando as demais abordagens. Esse resultado sugere que o recorte de contexto favoreceu a captura de aspectos linguísticos e coesivos relevantes para a avaliação dessa competência, evidenciando a vantagem de modelos especializados em português.

D.5 Competência 5 - Proposta de intervenção com respeito aos direitos humanos

Abordagem	QWK	Modelo	Tipo de dataset	Principais configurações
Fine-tuning padrão (A6)	**0.748**	BERTimbau-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação multi-saída (A4)	0.715	XLM-RoBERTa-large	Completo	150 épocas; *batch* = 10; *lr* = 2e-5
Classificação ordinal com *WKLoss* (A7)	0.703	XLM-RoBERTa-large	Filtrado	150 épocas; *batch* = 10; *lr* = 2e-5
Regressão tradicional (*TF-IDF + XGBoost*) (A9)	0.601	XGBoost Regressor	Completo	*batch* = 14; *lr* = 1e-5
Classificação multi-saída com *WKLoss* (A5)	0.600	BERTimbau-large	Completo	150 épocas; *batch* = 12; *lr* = 2e-5
Classificação sem *fine-tuning* (A8)	0.185	BERTimbau-large	Completo	100 épocas; *batch* = 10; *lr* = 2e-5
Classificação ordinal em cascata (A3)	0.071	BERTimbau-large	Completo	150 épocas; *batch* = 10; *lr* = 2e-5

Tabela D.5: Desempenho das principais abordagens testadas para a competência C5 (Proposta de intervenção com respeito aos direitos humanos). A tabela apresenta os valores da métrica QWK obtidos por diferentes modelos e tipos de dataset, destacando suas principais configurações experimentais. O *fine-tuning* padrão com o modelo BERTimbau-large e dados filtrados ($\leq$ 512 tokens) alcançou o melhor resultado geral (QWK = 0.748), superando as demais estratégias. Observa-se ainda que os dois melhores resultados para essa competência foram obtidos com o uso de redações limitadas a 512 tokens, indicando que o recorte de contexto favoreceu a consistência das previsões.

Apêndice E Ferramenta Quintana

Este apêndice descreve a **Quintana**, ferramenta Web desenvolvida com dois objetivos principais: (i) viabilizar o uso prático dos modelos de predição de notas por competência em contextos educacionais reais, e (ii) oferecer uma infraestrutura flexível para incorporar novos modelos de avaliação desenvolvidos a partir desta pesquisa ou de trabalhos futuros.

E.1 Visão geral da ferramenta

A Quintana é uma aplicação Web responsiva que permite a submissão, avaliação e visualização de redações alinhadas à matriz de competências do ENEM. A ferramenta foi concebida para atender dois perfis de usuário:

- **Aluno**: pode selecionar um tema de redação, digitar ou colar seu texto na interface Web e solicitar a avaliação automática. O sistema retorna as notas previstas para cada uma das cinco competências do ENEM, a nota total (soma das competências) e um conjunto de comentários pedagógicos organizados por competência.
- **Professor**: pode cadastrar e gerenciar temas de redação, acompanhar as produções associadas a cada tema e consultar, para cada redação, as notas previstas pelos modelos, além do texto completo e dos feedbacks gerados.

A Figura E.1 ilustra a interface de gerenciamento de temas, em que os professores podem cadastrar novos tópicos de redação e visualizar os temas já existentes.

Ao selecionar um tema, o aluno tem acesso à proposta completa e ao campo para escrever a redação diretamente na plataforma, conforme ilustrado na Figura E.2. Após finalizar o texto, o estudante pode apagar o conteúdo ou enviá-lo para correção automática.

E.2 Arquitetura e integração com os modelos

A arquitetura da Quintana segue o padrão cliente-servidor, com separação clara entre a camada de apresentação e os serviços de predição:

- **Frontend**: desenvolvido em Next.js, responsável pelas interfaces de aluno e professor, submissão de textos, visualização de notas e feedbacks.

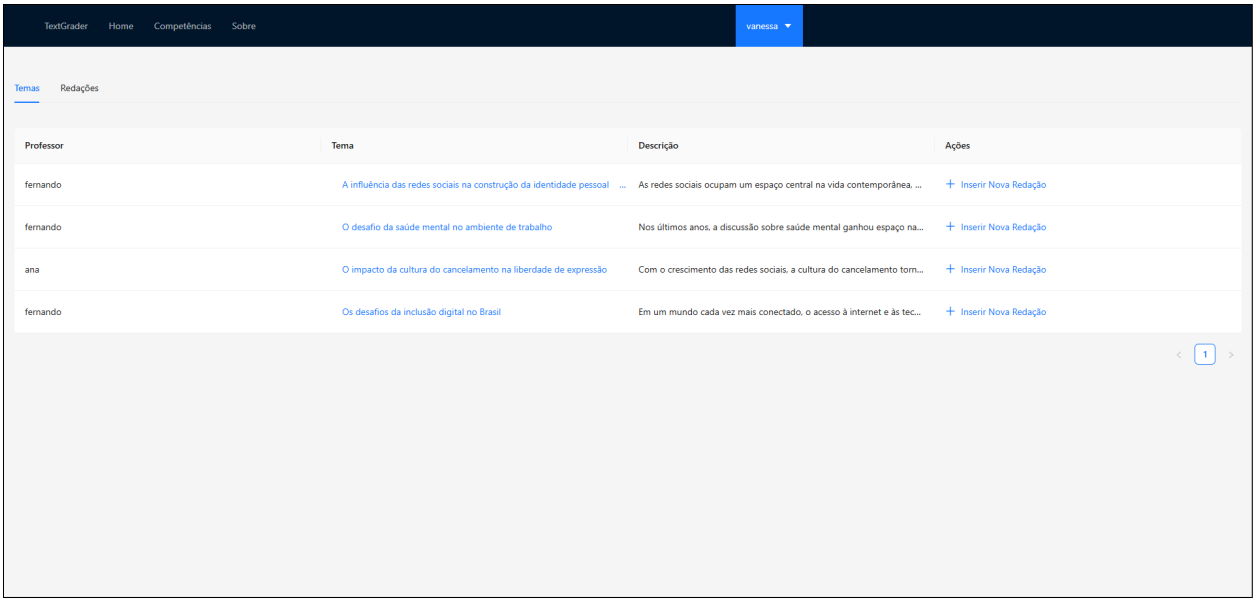


Figura E.1: Tela de listagem de temas na Quintana. Cada linha corresponde a um tema cadastrado pelo professor, incluindo título, descrição e ações para inserção de novas redações associadas.

- **Backend**: implementado em Python, expõe uma API REST que recebe as redações, executa o pré-processamento, aciona os modelos de avaliação e retorna as notas previstas.
- **Camada de modelos**: composta pelos modelos de aprendizado de máquina investigados nesta dissertação (Capítulos IV e V), incluindo abordagens por competência, multi-saída e arquiteturas ordinais.
- **Persistência**: banco de dados relacional (PostgreSQL) para armazenamento de usuários, temas, redações, notas previstas e metadados.

A integração entre a aplicação Web e os modelos de avaliação ocorre de forma transparente para o usuário. A Quintana foi projetada com uma camada de abstração que separa a lógica de interface das estratégias de predição, permitindo que diferentes modelos, incluindo aqueles desenvolvidos ao longo desta dissertação, possam ser incorporados ao sistema sem necessidade de alterações estruturais.

Essa arquitetura modular possibilita três capacidades fundamentais:

- **Flexibilidade na troca de modelos**: a ferramenta pode utilizar diferentes arquiteturas de predição (por competência, multi-saída ou cascata) sem que a interface Web precise ser modificada.
- **Incorporação de novos modelos**: extensões e experimentos produzidos nesta pesquisa podem ser adicionados ao sistema de forma direta, permitindo que a Quintana acompanhe avanços metodológicos futuros.

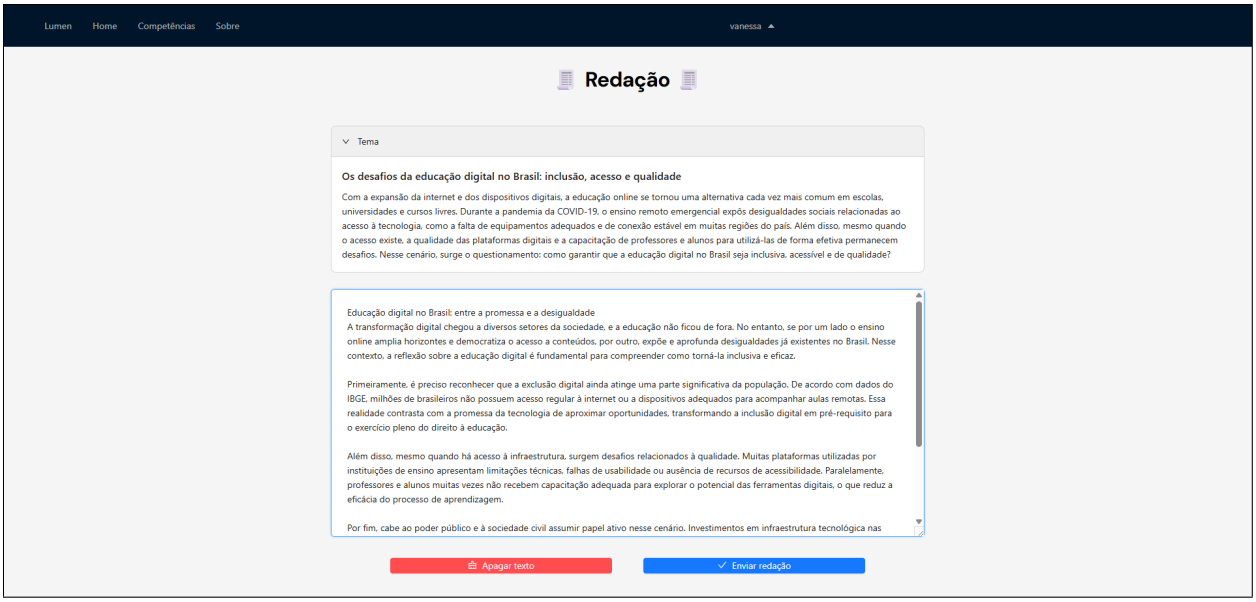


Figura E.2: Tela de submissão de redações na Quintana. O aluno seleciona o tema, escreve ou cola o texto na área principal e solicita a avaliação automática.

- **Configuração dinâmica**: administradores ou pesquisadores podem definir, no ambiente de execução, qual modelo será utilizado, favorecendo testes controlados, comparações experimentais e evolução contínua da ferramenta.

Em síntese, a Quintana atua como um ambiente unificado que conecta os modelos investigados nesta dissertação ao uso prático em contexto educacional, preservando tanto a escalabilidade quanto a facilidade de manutenção.

E.3 Fluxo de uso e visualização das avaliações

Após a submissão de uma redação, o backend aciona o modelo configurado, que retorna as notas previstas para as cinco competências do ENEM. A interface de aluno exibe essas notas em uma visão consolidada, permitindo ao estudante compreender seu desempenho de forma detalhada.

A Figura E.3 mostra um exemplo de visualização de redação já avaliada, incluindo o texto completo e o acesso às notas por competência.

As notas previstas são apresentadas de forma organizada, competência a competência, como ilustrado na Figura E.4. Esse tipo de visualização reforça a aderência à matriz oficial do ENEM e facilita a interpretação pedagógica dos resultados.

Além das notas, a Quintana integra um modelo de linguagem de grande porte para geração de *feedbacks* pedagógicos personalizados. A Figura E.5 apresenta um exemplo de tela com os comentários organizados por competência, destacando pontos fortes e sugerindo melhorias específicas.

Do ponto de vista docente, a ferramenta também disponibiliza visualizações agregadas por tema, com listagem das redações associadas, permitindo ao professor acompanhar o conjunto de

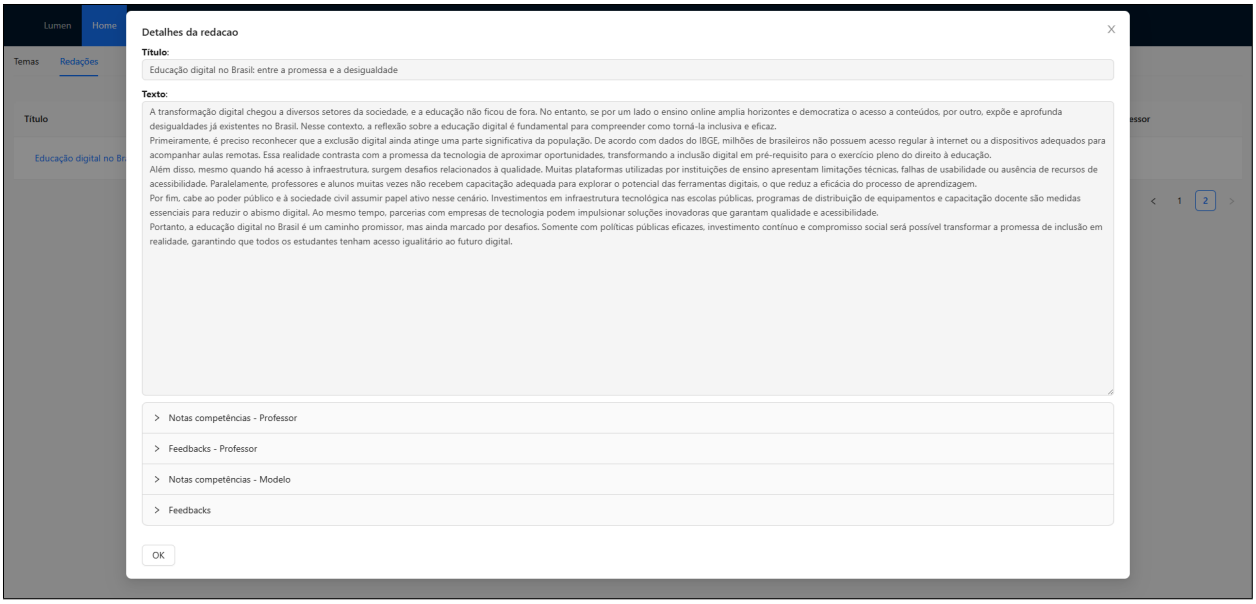


Figura E.3: Detalhes de uma redação avaliada na Quintana, com exibição do texto completo e acesso às notas atribuídas em cada competência.

produções da turma. A Figura E.6 mostra um exemplo de detalhamento de tema com sua descrição e acesso às redações vinculadas.

E.4 Disponibilização e reprodutibilidade

A Quintana foi disponibilizada em código aberto, juntamente com a documentação de instalação e uso, de forma a favorecer a reprodutibilidade dos resultados obtidos nesta dissertação e a adoção da ferramenta por outras instituições. O código-fonte encontra-se no repositório GitHub[1], e um vídeo de demonstração do sistema está disponível em plataforma pública de vídeo[2].

Ao integrar os modelos desenvolvidos neste trabalho em uma aplicação Web acessível, a Quintana cumpre o objetivo de aproximar os resultados de pesquisa de seu uso efetivo em contextos educacionais, oferecendo uma infraestrutura modular que pode ser estendida com novos modelos, bases de dados e funcionalidades em trabalhos futuros.

[1] `https://github.com/AILAB-CEFET-RJ/quintana_tool`
[2] `https://youtu.be/zY_CWbmlyOY`



Figura E.4: Visualização das notas por competência na Quintana. Cada competência recebe uma pontuação específica, permitindo uma análise detalhada do desempenho do aluno.

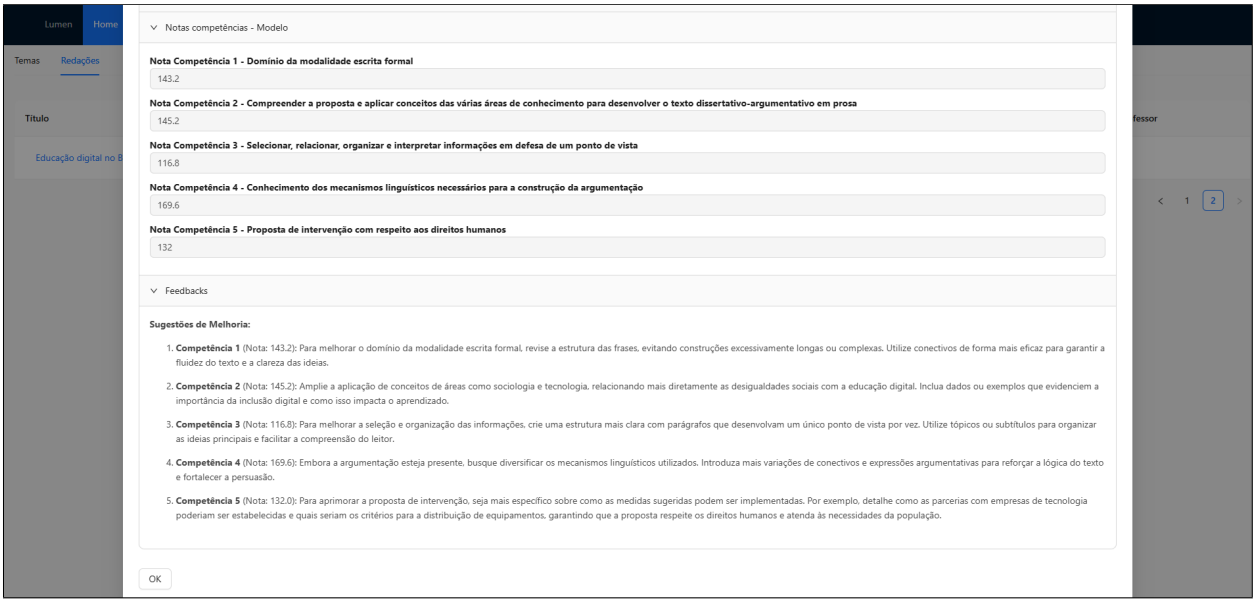


Figura E.5: Tela de feedback pedagógico na Quintana. Os comentários são organizados por competência e auxiliam o estudante a compreender as razões da nota e como aprimorar sua escrita.

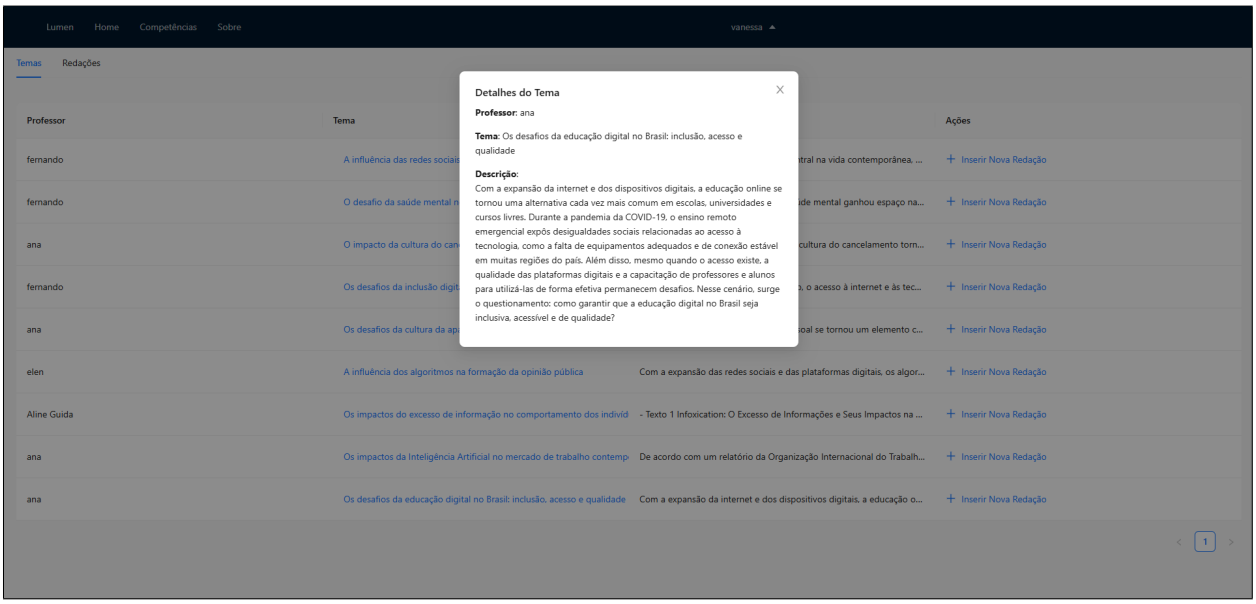


Figura E.6: Tela de detalhes de um tema de redação. A partir dessa interface, o professor pode consultar as redações associadas e acompanhar o desempenho da turma.