Quintana: Ferramenta de Avaliação de Redações*

Vanessa Soares, Danilo Freire, Lucas Bittencourt, Alexander Feitosa, Jorge Pavão, Flavio Carvalho, Kele Belloze, Gustavo Guedes, Eduardo Bezerra

[1]Programa de Pós-graduação em Ciência da Computação – CEFET/RJ
{vanessa.soares, danilo.freire, lucas.bittencourt,
alexander.feitosa, jorge.pavao, flavio.carvalho}@aluno.cefet-rj.br,
{kele.belloze, gustavo.guedes, ebezerra}@cefet-rj.br

Abstract. *This article presents Quintana, an AI-based essay assessment tool for Portuguese-language argumentative texts, based on the competencies of the Exame Nacional do Ensino Médio (ENEM). The tool uses machine learning models trained on essays extracted from the "Brasil Escola" platform and allows students to submit essays and receive predicted scores for each competency. The results are presented in a pedagogically organized way, with scores and feedback grouped by competency, supporting formative use by students and teachers without replacing teacher assessment. This demo showcases how modern Natural Language Processing techniques can be integrated into a practical and accessible educational solution with meaningful social impact.*

1. Introdução

A avaliação de textos dissertativos em larga escala é um desafio constante no contexto educacional brasileiro, especialmente no Exame Nacional do Ensino Médio (ENEM), que mobiliza anualmente milhões de redações corrigidas manualmente por professores. Esse processo demanda considerável tempo e esforço, o que torna a automação da correção de textos uma alternativa promissora para ampliar a escala do *feedback* e apoiar a padronização de práticas avaliativas [Lim et al. 2021]. Além disso, ao tornar as correções mais rápidas e assertivas, isso possibilita a melhoria do retorno para o estudante, contribuindo de forma significativa na experiência de aprendizagem [Ye and Manoharan 2021].

Este artigo apresenta a Quintana, uma plataforma aberta para avaliação automatizada de redações em português, organizada segundo as competências do ENEM: norma escrita, compreensão da proposta, argumentação, coesão e proposta de intervenção. A ferramenta integra modelos neurais ajustados para predição de notas por competência, geração de *feedback* formativo por *Large Language Models* (LLM) e visualizações pedagógicas para estudantes e professores. Diferentemente de soluções fechadas, a Quintana prioriza abertura, rastreabilidade e uso educacional em contextos públicos. Além da descrição da arquitetura e das funcionalidades, apresenta-se uma avaliação preliminar da qualidade das predições e dos *feedbacks* gerados, bem como uma discussão sobre limitações e cuidados necessários para seu uso responsável. A plataforma tem como objetivo oferecer correção automatizada por competência, *feedback* formativo imediato e visualizações pedagógicas de código aberto, acessíveis a escolas e redes públicas sem custo de licença.

*Link para o vídeo de demonstração: `https://youtu.be/oxzIV5Oce2Y`

2. Posicionamento e Diferenciais

A avaliação automatizada de redações em português segundo a matriz do ENEM já foi investigada em trabalhos baseados em atributos linguísticos, modelos estatísticos, redes neurais e transformers [Amorim and Veloso 2017, Fonseca et al. 2018]. Entretanto, muitas dessas propostas concentram-se na avaliação experimental dos modelos, sem disponibilizar uma ferramenta Web voltada ao uso pedagógico por estudantes e professores. Também existem soluções comerciais, como CoRedação e GLAU, mas em geral elas oferecem pouca informação pública sobre dados, modelos, protocolos de avaliação e possibilidades de adaptação por instituições públicas.

Um diferencial da Quintana é sua natureza aberta e acadêmica, posto que foi concebida não apenas como um serviço de correção automatizada, mas como uma plataforma educacional, com código e materiais disponibilizados publicamente, permitindo auditoria, adaptação e evolução por outras instituições. Essa característica é especialmente relevante para contextos públicos de ensino, nos quais transparência, reprodutibilidade e sustentabilidade são requisitos importantes para a adoção de tecnologias educacionais.

Outro diferencial é sua organização em torno das cinco competências do ENEM. Em vez de produzir apenas uma nota global, a plataforma apresenta resultados por competência, preservando a granularidade da matriz avaliativa e permitindo que o estudante identifique dimensões específicas de sua escrita que precisam ser aprimoradas. Essa decisão também orienta a geração dos *feedbacks* pedagógicos, que são organizados segundo os mesmos critérios utilizados na avaliação.

Por fim, a Quintana integra modelos preditivos, geração de *feedback* e visualizações pedagógicas em uma aplicação Web voltada ao fluxo real de uso por professores e estudantes, conforme detalhado nas seções seguintes. A Tabela 1 apresenta o posicionamento da Quintana em relação a trabalhos acadêmicos e ferramentas similares.

Tabela 1. Posicionamento em relação a trabalhos e ferramentas similares.

Sistema	Por comp.	Aberto	QWK reportado	Ferramenta Web
Trabalhos acadêmicos				
[Amorim and Veloso 2017]	Sim	Não	0,425[a]	Não
[Fonseca et al. 2018]	Sim	Não	0,752[b]	Não
[Silveira et al. 2024]	Sim	Sim	até 0,500	Não
Ferramentas comerciais				
CoRedação	Sim	Não	—	Sim
Glau	Sim	Não	—	Sim
Quintana	**Sim**	**Sim**	**até 0,748**	**Sim**

[a]Escala 0–2; não diretamente comparável. [b]Corpus privado, 56 644 redações.

3. Metodologia

Esta seção descreve os modelos de predição de notas e geração de *feedback* integrados à Quintana, bem como a arquitetura da plataforma e o fluxo de uso pelos perfis de estudante e professor.

3.1. Modelos de avaliação e geração de feedback

A Quintana avalia redações segundo as cinco competências do ENEM, atribuindo a cada uma delas uma nota de 0 a 200 pontos: norma escrita (C1), compreensão da proposta (C2), argumentação (C3), coesão (C4) e proposta de intervenção (C5). A nota final é obtida pela soma das cinco predições, totalizando até 1000 pontos.

Os modelos de nota foram treinados a partir de um corpus de 9.599 redações em português, coletadas da plataforma Brasil Escola[1], abrangendo 102 temas distintos e textos produzidos entre maio de 2009 e dezembro de 2024. Cada redação possui notas atribuídas às cinco competências do ENEM. Os textos não incluem identificadores pessoais e foram utilizados exclusivamente para fins de pesquisa não-comercial; o uso seguiu as diretrizes éticas do Cefet/RJ, com isenção de revisão completa por se tratar de conteúdo publicamente disponível sem dados pessoais identificáveis. Como limitação, não dispomos de informações detalhadas sobre concordância entre avaliadores ou padronização interna do processo de correção da plataforma de origem. Para evitar vazamento temático, a partição dos dados foi realizada no nível dos temas, de modo que um mesmo tema não aparecesse simultaneamente nos conjuntos de treino, validação e teste.

A versão integrada à ferramenta utiliza modelos especializados por competência, selecionados a partir de uma avaliação experimental prévia que comparou modelos por competência, modelos multi-saída e arquiteturas em cascata. Essa escolha foi motivada pelo desempenho: modelos multi-saída obtiveram Quadratic Weighted Kappa (QWK) de apenas 0,26, resultado atribuído à transferência negativa entre sinais de aprendizado heterogêneos (C1 avalia controle gramatical, enquanto C5 exige raciocínio sociopolítico de alto nível), ao passo que modelos por competência atingiram QWK entre 0,64 e 0,75. Os modelos foram baseados em *XLM-RoBERTa* [Conneau et al. 2019] e *BERTimbau*, com ajuste fino eficiente por meio de *Low-Rank Adaptation* (LoRA) [Hu et al. 2022] (rank $r = 8$, $\alpha = 16$, $\approx$0,8M parâmetros treináveis). C1–C3, que avaliam propriedades textuais globais e trans-linguísticas, foram melhor capturadas pelo XLM-RoBERTa com WKLoss; C4–C5, que dependem de marcadores coesivos e registro sociopolítico específicos do português brasileiro, obtiveram ganhos expressivos com o BERTimbau. A Tabela 2 resume o desempenho dos modelos selecionados para integração à plataforma, usando QWK como métrica de concordância ordinal com as notas humanas, reportado como média de três execuções com sementes fixas.

Tabela 2. Desempenho dos modelos integrados à Quintana (QWK no conjunto de teste, média de três execuções). O *baseline* de maioria prediz sempre a classe mais frequente no treino; QWK$= 0{,}00$ por definição para preditor constante.

Comp.	Modelo	QWK (baseline)	QWK (modelo)
C1	XLM-RoBERTa + WKLoss	0,00	0,659
C2	XLM-RoBERTa + WKLoss	0,00	0,657
C3	XLM-RoBERTa + WKLoss	0,00	0,642
C4	BERTimbau + CE	0,00	0,729
C5	BERTimbau + CE	0,00	0,748

[1] https://vestibular.brasilescola.uol.com.br/corrige-aqui

A análise dos erros de predição revela um padrão sistemático de confusão entre níveis adjacentes: em todas as competências, a maior parte dos equívocos ocorre entre classes vizinhas (e.g., prever nível 3 quando a nota humana é 4), padrão coerente com a natureza ordinal das notas e que justifica o uso do QWK, que penaliza erros proporcionalmente à distância entre classes. Os modelos apresentam melhor calibração nas faixas intermediárias (níveis 2 a 4), onde há mais exemplos de treinamento, e tendem a subestimar redações de nota máxima, cujas características de alto nível discursivo são as mais difíceis de capturar. Os valores de QWK obtidos (0,64–0,75) se aproximam do intervalo de concordância entre avaliadores humanos (0,20–0,60) reportado na literatura para corpora de estilo semelhante [Silveira et al. 2024], sugerindo que os modelos operam próximos ao teto imposto pela consistência das anotações.

Além da predição das notas, a ferramenta utiliza um LLM, especificamente o *Gemma-7B*, para gerar *feedbacks* pedagógicos personalizados. O modelo recebe como entrada a redação, o tema proposto e as notas previstas para as cinco competências, produzindo recomendações por competência, conforme o critério avaliativo descrito anteriormente. Essas devolutivas são apresentadas como apoio formativo ao estudante e ao professor, não como substitutas da avaliação docente.

3.2. Arquitetura e funcionalidades da ferramenta

A Quintana está disponível como uma aplicação Web responsiva, organizada em uma arquitetura cliente-servidor que separa a interface do usuário, os serviços de processamento e o armazenamento persistente. Como ilustrado na Figura 1, o *frontend*, desenvolvido em Next.js, oferece interfaces específicas para os perfis de *Aluno* e *Professor*. O *backend*, implementado em Python, expõe uma API RESTful documentada segundo a especificação OpenAPI e coordena o fluxo de avaliação das redações. As submissões, notas previstas, *feedbacks* e metadados associados são armazenados em um banco MongoDB, permitindo histórico de avaliações e visualizações pedagógicas.

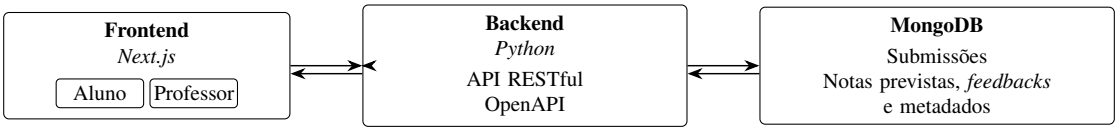
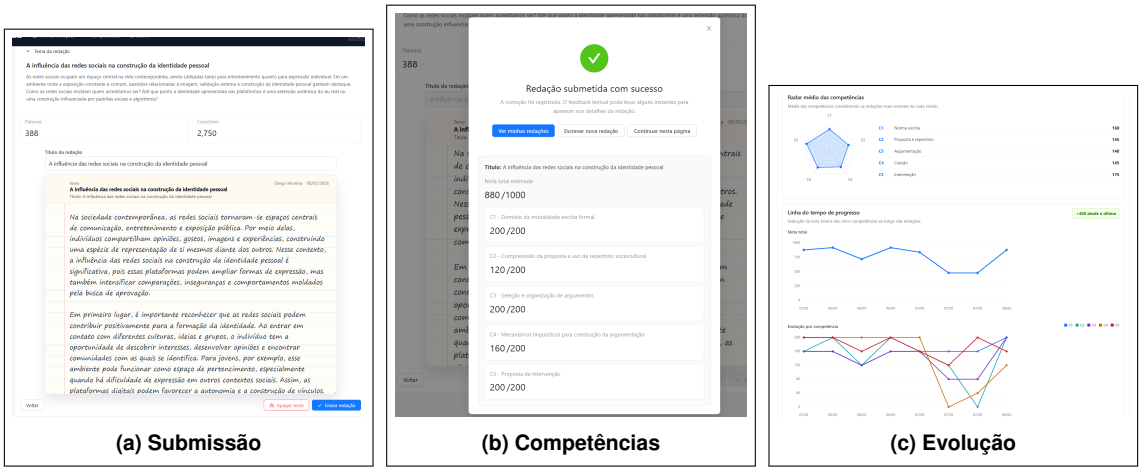


Figura 1. Arquitetura da Quintana.

Para submeter uma redação, o aluno seleciona um tema previamente cadastrado pelo professor e escreve o texto argumentativo pela interface Web (Figura 2a). A submissão é encaminhada à API, que aciona os modelos de predição das cinco competências do ENEM e, em seguida, o módulo de geração de *feedback* pedagógico. Ao final do processamento, o sistema apresenta as notas por competência e recomendações de melhoria organizadas de acordo com a matriz avaliativa do ENEM. Esse fluxo permite que a ferramenta seja usada tanto pelo estudante, em atividades de reescrita e autoavaliação, quanto pelo professor, no acompanhamento do desempenho da turma. A Figura 2b exibe a distribuição das notas previstas por competência, permitindo identificar quais dimensões da escrita apresentam melhor desempenho e quais demandam maior atenção pedagógica. A Figura 2c apresenta os *feedbacks* textuais em linguagem acessível, organizados por competência, destacando aspectos positivos e possibilidades de aprimoramento como apoio ao processo formativo.



(a) Submissão (b) Competências (c) Evolução

Figura 2. Telas do perfil do estudante: submissão de redações, visualização do resultado da avaliação por IA, gráficos de evolução do aprendizado.

No perfil do professor (Figura 3), a plataforma agrega as avaliações automáticas para apoiar o acompanhamento pedagógico da turma. A interface permite identificar dificuldades coletivas, observar padrões de desempenho por aluno e competência, e organizar intervenções didáticas a partir de *rankings*, alertas e agrupamentos pedagógicos. A ferramenta também oferece funcionalidades de gerenciamento de temas de redação, permitindo criar, editar e remover propostas alinhadas aos objetivos pedagógicos de cada atividade.

(a) Alertas (b) Heatmap (c) Competências (d) Distribuição

Figura 3. Telas do perfil do professor: alertas, heatmap aluno/competência, distribuição por competência e agrupamentos para intervenção.

4. Considerações Finais

A Quintana apresenta uma proposta aberta e acessível para apoiar a avaliação formativa de redações em português, combinando predição de notas por competência, geração de

feedback textual e visualizações voltadas ao acompanhamento pedagógico. Seu principal potencial está em ampliar as oportunidades de retorno ao estudante e oferecer ao professor indicadores organizados sobre o desempenho textual da turma. Distribuída sob licença Creative Commons, a Quintana está disponível para acesso, uso e adaptação por meio do repositório eduCAPES[2]. Além disso, o código da ferramenta está disponível em um repositório no Github[3].

A Quintana é uma ferramenta de apoio à avaliação formativa, e não um substituto da avaliação docente. As notas previstas podem apresentar erros, especialmente em casos de redações muito curtas, muito longas ou com características pouco representadas no corpus de treinamento. A qualidade dos *feedbacks* gerados pelo Gemma-7B não foi avaliada formalmente nesta versão; avaliações automáticas (e.g., BERTScore) e por julgamento humano estão previstas como trabalho futuro; até lá, recomenda-se que professores os utilizem como ponto de partida, não como avaliação definitiva. A plataforma também permite o registro posterior de notas e comentários pelo professor, mantidos como uma camada separada de avaliação humana, de modo que os resultados automáticos funcionam como subsídios pedagógicos e não como decisão avaliativa final.

Como perspectivas futuras, planeja-se expandir sua aplicação para outros gêneros discursivos e incorporar modelos generativos mais poderosos e consequentemente mais capazes de oferecer *feedback* textual automatizado, potencializando seu papel como ferramenta de apoio ao ensino e à aprendizagem.

Referências

Amorim, E. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for brazilian portuguese. In *Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *International Conference on Computational Processing of the Portuguese Language*, pages 170–179. Springer.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *PROPOR 2024*, pages 228–237. Association for Computational Linguistics.

Ye, X. and Manoharan, S. (2021). Performance comparison of automated essay graders based on various language models. In *ICOCO 2021*, pages 152–157.

[2]`https://educapes.capes.gov.br/`

[3]`https://github.com/AILAB-CEFET-RJ/quintana_tool`