Avaliação Automática de Redações do Enem: Um Estudo Empírico sobre Representações Linguísticas e Contextuais

Gabriel Gonçalves de Matos e Valéria D. Feltrim
Departamento de Informática
Universidade Estadual de Maringá
Maringá – PR – Brasil
briellh9@gmail.com e vdfeltrim@uem.br

Resumo

A Avaliação Automática de Redações (AAR) para o português brasileiro ainda é uma tarefa desafiadora, particularmente no contexto do exame Enem, no qual a qualidade textual é avaliada por meio de múltiplas competências e as notas apresentam natureza ordinal. Neste artigo, investigamos estratégias de modelagem híbrida para AAR em nível de competência, combinando características linguísticas explícitas com representações contextuais. Utilizando o córpus Enem-AES, a avaliação de cada competência foi modelada como um problema de predição ordinal por meio do *framework* CORAL. Foi realizada uma comparação empírica controlada entre representações lexicais tradicionais, um amplo conjunto de métricas linguísticas extraídas com o sistema NILC-Metrix, características manuais orientadas à tarefa, *embeddings* contextuais e combinações dessas representações. Os resultados mostram que modelos híbridos alcançam o maior nível médio de concordância com as notas humanas, embora o desempenho varie entre competências e dependa do tipo de representação utilizada. Além disso, foi analisado o comportamento dos modelos em cenários de discordância entre avaliadores, o que evidenciou o impacto da variabilidade de anotação no desempenho dos modelos. De modo geral, os resultados fornecem evidências de que a combinação de indicadores linguísticos com *embeddings* contextuais constitui uma estratégia promissora para a tarefa de AAR no contexto do Enem.

1 Introdução

A Avaliação Automática de Redações (AAR) é uma área consolidada do Processamento de Linguagem Natural, com aplicações diretas em contextos educacionais de larga escala. No Brasil, o Exame Nacional do Ensino Médio (Enem) constitui um cenário particularmente relevante para a AAR, uma vez que milhões de redações são avaliadas anualmente segundo uma rubrica estruturada em cinco competências. Cada competência contempla diferentes dimensões da escrita, como domínio da norma padrão, desenvolvimento argumentativo, coesão textual e elaboração de proposta de intervenção, o que torna a tarefa de pontuação automática especialmente desafiadora.

Um dos aspectos centrais na AAR diz respeito à escolha da representação textual. Trabalhos iniciais para o português brasileiro exploraram predominantemente características linguísticas explícitas, como métricas lexicais, gramaticais e estruturais, que apresentam boa interpretabilidade e alinham-se diretamente a critérios avaliativos humanos (Amorim e Veloso, 2017). Estudos posteriores passaram a investigar modelos neurais, demonstrando que representações aprendidas automaticamente podem capturar informações semânticas relevantes para a tarefa (Fonseca et al., 2018; Silveira et al., 2024). Evidências empíricas mostram, entretanto, que diferentes tipos de representação apresentam comportamentos distintos conforme a competência avaliada, indicando que abordagens baseadas em uma única representação tendem a capturar apenas parte das dimensões da escrita exigidas no Enem (Silveira et al., 2024).

Nesse contexto, abordagens híbridas têm sido apontadas como uma alternativa promissora, ao combinar informações linguísticas explícitas com representações contextuais aprendidas automaticamente. A motivação central dessas abordagens reside na complementaridade entre diferentes tipos de representação: enquanto características linguísticas tendem a ser mais eficazes na modelagem de aspectos formais e estruturais do texto, *embeddings* contextuais mostram vantagens na captura de conteúdo temático, relações semânticas e organização discursiva. Apesar desse potencial, ainda são relativamente escassos estudos que avaliem, de forma sistemática e sob um mesmo protocolo experimental, o impacto dessas diferentes representações e de

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 488–497

suas combinações na avaliação por competências do Enem.

Neste trabalho, investigamos empiricamente o papel das representações textuais na AAR do Enem por meio da avaliação de modelos baseados em diferentes estratégias de representação e de suas combinações híbridas. Realizamos um estudo controlado no córpus Enem-AES (Silveira et al., 2024), comparando representações tradicionais baseadas em TF-IDF, conjuntos extensos de características linguísticas extraídas com o sistema NILC-Metrix (Leal et al., 2023), características manuais orientadas à tarefa (Neves, 2025) e *embeddings* contextuais, bem como combinações híbridas dessas representações. A avaliação foi conduzida no nível de competências, com a predição das notas formulada como um problema de regressão ordinal por meio do *framework* CORAL, respeitando a natureza discreta e ordenada das pontuações do Enem.

Além da comparação global de desempenho, analisamos o comportamento dos modelos sob diferentes níveis de concordância entre avaliadores humanos. Essa análise permite examinar de forma mais precisa como diferentes estratégias de representação respondem à variabilidade da anotação humana e quais limites práticos essa variabilidade impõe à AAR. Em vez de focar em ganhos absolutos de desempenho, o estudo busca contribuir para uma compreensão mais aprofundada do papel das representações textuais na avaliação automática de redações do Enem, considerando tanto o comportamento por competência quanto os limites impostos pela variabilidade da anotação humana.

2 Trabalhos Relacionados

A AAR para o português brasileiro tem sido investigada principalmente no contexto do Enem, explorando diferentes estratégias de representação textual e modelagem supervisionada. Os trabalhos existentes podem ser organizados, de forma geral, em abordagens baseadas em características linguísticas explícitas, modelos neurais e, mais recentemente, esforços voltados à construção e padronização de *benchmarks*.

Um dos primeiros estudos sistemáticos sobre AAR no Enem foi apresentado por Amorim e Veloso (2017), que propuseram um sistema multiaspecto baseado em características linguísticas manuais, incluindo métricas lexicais, gramaticais, estilísticas e específicas do domínio. Os autores conduziram uma análise detalhada da contribuição individual das características para cada competência, evidenciando que diferentes dimensões da escrita são capturadas por conjuntos distintos de atributos. Apesar da relevância metodológica, o trabalho utiliza um esquema de pontuação diferente do adotado oficialmente no Enem, o que dificulta comparações diretas com estudos posteriores.

Avançando na comparação entre representações, Fonseca et al. (2018) avaliaram, em larga escala, modelos baseados em redes neurais recorrentes e modelos baseados em um amplo conjunto de características linguísticas explícitas. Utilizando um córpus proprietário com mais de 50 mil redações, os autores observaram que abordagens baseadas em características explícitas apresentaram desempenho competitivo e, em alguns casos, superior aos modelos neurais, sobretudo em competências mais estruturais. Os resultados reforçam a ideia de que representações linguísticas cuidadosamente projetadas continuam relevantes, mesmo diante de modelos neurais mais complexos.

No mesmo eixo comparativo, Filho et al. (2021) avaliaram métodos de *feature engineering* e modelos de aprendizado profundo para a AAR do Enem. Utilizando um córpus público de aproximadamente 4.300 redações, os autores compararam modelos baseados em regressão e *Gradient Boosting Trees* com uma arquitetura neural recorrente, além de investigarem extensivamente técnicas de balanceamento de classes. Os resultados mostraram que abordagens baseadas em aprendizado profundo apresentam vantagens quando não há balanceamento, enquanto modelos baseados em características explícitas se beneficiam de técnicas de reamostragem. Embora o estudo forneça uma análise cuidadosa sobre o impacto do balanceamento e da distribuição das classes, a avaliação apresentada concentra-se majoritariamente na competência C1.

Marinho et al. (2022a,b) consolidaram esforços na criação de recursos e modelagem baseada em características manuais e *embeddings* estáticos para o português. No primeiro trabalho, os autores introduziram o córpus Essay-BR, tornando publicamente acessíveis milhares de redações avaliadas segundo os critérios oficiais do Enem. Em um estudo complementar focado na extração de características, os autores propuseram conjuntos de características manuais, abrangendo aspectos léxicos, sintáticos e semânticos para cada uma das cinco competências do Enem. A comparação de modelos

baseados em engenharia de atributos com modelos Doc2Vec e redes neurais recorrentes (LSTM) mostrou que as características manuais apresentam desempenho superior em dimensões normativas e estruturais (C1 e C2), enquanto as LSTMs são mais eficazes em capturar a subjetividade das demais competências (C3, C4 e C5). Esses achados reforçam a relevância de representações linguísticas explícitas para aspectos gramaticais, embora o forte desbalanceamento das notas continue sendo um desafio para a predição de conceitos extremos.

Mais recentemente, Silveira et al. (2024) introduziram o córpus Enem-AES, propondo um *benchmark* cuidadosamente curado, com maior controle sobre a qualidade das anotações, análise de concordância entre avaliadores e avaliação sistemática de modelos baseados em diferentes representações, incluindo *embeddings* contextuais. Os autores mostram que o desempenho dos modelos varia significativamente entre competências e que a dificuldade da tarefa está fortemente associada ao nível de concordância humana.

Em um trabalho complementar, Silveira et al. (2025) investigaram a robustez de sistemas de AAR baseados em *Transformers* frente a ataques adversariais simples, evidenciando que modelos fortemente dependentes de representações aprendidas podem ser sensíveis a pistas superficiais do texto. Esses resultados reforçam a importância de compreender quais aspectos da escrita são efetivamente capturados por diferentes representações, especialmente em contextos educacionais.

Em síntese, a literatura indica que abordagens baseadas exclusivamente em características linguísticas explícitas ou em representações neurais apresentam limitações complementares. Apesar de evidências pontuais sobre essa complementaridade, ainda são escassos estudos empíricos que avaliem, sob um mesmo protocolo experimental e em nível de competência, o impacto de diferentes representações e de suas combinações híbridas na AAR do Enem. Este trabalho se insere nesse contexto ao realizar uma análise comparativa focada no papel das representações textuais.

3 Metodologia

Esta seção descreve o conjunto de dados utilizado, as estratégias de representação textual avaliadas, a modelagem adotada e o protocolo experimental empregado na avaliação dos modelos. Para fins de reprodutibilidade, o código-fonte, os hiperparâmetros e os scripts de treinamento utilizados estão disponíveis publicamente no repositório do projeto[1].

3.1 Conjunto de Dados

Os experimentos foram conduzidos utilizando o córpus Enem-AES (Silveira et al., 2024), um *benchmark* público composto por 1.165 redações do Enem avaliadas segundo as cinco competências oficiais do exame. Cada redação possui, para cada competência, uma nota discreta pertencente ao conjunto $\{0, 40, 80, 120, 160, 200\}$, com distribuição desbalanceada das notas.

Cada redação possui três avaliações, sendo duas realizadas por anotadores treinados e uma extraída da plataforma *online* de onde as redações foram coletadas. Segundo os autores, a análise de concordância entre anotações, via métrica *Quadratic Weighted Kappa* (QWK), mostra que o acordo entre os anotadores e as notas da plataforma situou-se na faixa de razoável a moderado ($0,2 \leq \text{QWK} \leq 0,6$). Já a concordância entre os anotadores participantes do estudo foi superior, atingindo níveis moderados ($0,4 \leq \text{QWK} \leq 0,8$). Essa diferença evidencia que as notas podem apresentar inconsistências ou ruídos que dificultam o aprendizado dos modelos.

Utilizamos as partições de treino (65%), validação (17%) e teste (18%) disponibilizadas pelos autores. O conjunto de teste disponibilizado inclui anotações de divergência entre avaliadores humanos, permitindo distinguir um subconjunto não-divergente, definido por diferenças inferiores a 80 pontos entre avaliações, o qual foi explorado em análises adicionais.

3.2 Representações Textuais

Investigamos diferentes estratégias de representação textual, abrangendo abordagens tradicionais, linguísticas e contextuais, bem como combinações híbridas dessas representações.

TF-IDF: Como representação lexical de referência, utilizamos vetores TF-IDF com redução de dimensionalidade para 1.000 características, servindo como *baseline* tradicional amplamente empregado em tarefas de PLN.

Características linguísticas: Foram avaliados dois conjuntos distintos de características linguísticas explícitas, que diferem em escopo e propósito

[1] https://github.com/GGMmattos/AES-ENEM-HybridModel.git

analítico. O primeiro é composto por 200 métricas extraídas automaticamente pelo sistema NILC-Metrix (Leal et al., 2023). Este conjunto oferece uma caracterização multinível da complexidade textual do português brasileiro, abrangendo indicadores lexicais, sintáticos (como profundidade da árvore de dependência), semânticos e discursivos (Leal et al., 2023). Por ser uma ferramenta de propósito geral, o NILC-Metrix captura a sofisticação da escrita sem estar atrelado a um gênero específico (Leal et al., 2023).

Em contraste, o segundo conjunto, proposto por Neves (2025), corresponde a 27 características orientadas à tarefa, projetadas especificamente para capturar sinais linguísticos mapeados nas competências da matriz de referência do Enem. Esse conjunto agrega características oriundas dos trabalhos de Amorim e Veloso (2017), Li et al. (2023), entre outras, e inclui indicadores de conformidade com a norma culta, presença de elementos da estrutura dissertativo-argumentativa, conectivos de coesão e aderência à temática proposta, priorizando a interpretabilidade pedagógica em detrimento da exaustividade linguística. Enquanto o NILC-Metrix descreve a complexidade geral do texto, o conjunto de Neves (2025) foca em quão bem o texto atende aos requisitos do exame.

***Embeddings*:** Como representação contextual, utilizamos *embeddings* obtidos a partir do modelo BERTimbau Base (Souza et al., 2020), com dimensionalidade de 768. A escolha do modelo *Base*, em vez de variantes *Large* ou modelos *Decoder-only*, fundamentou-se no equilíbrio entre desempenho preditivo e viabilidade computacional. Conforme demonstrado por Barbosa et al. (2025), embora modelos de maior escala possam oferecer ganhos incrementais em competências subjetivas, como Argumentação (C3), modelos *Encoder-only* de tamanho base continuam altamente competitivos para capturar aspectos estruturais e normativos, como fluência (C1) e coesão (C4). Além disso, essa configuração garante a escalabilidade do sistema em cenários reais de aplicação, permitindo o processamento de grandes volumes de redações em ambientes com recursos de *hardware* limitados sem prejuízo significativo à qualidade da avaliação. Para redações com tamanho superior a 512 *tokens*, aplicamos segmentação com *stride* de 256 *tokens*, seguida da agregação por média dos *embeddings* associados ao token [CLS], de modo a preservar o contexto global do texto.

Representações híbridas: As representações híbridas foram obtidas por meio da concatenação vetorial entre *embeddings* contextuais e conjuntos de características linguísticas explícitas. Os vetores resultantes foram normalizados por *z-score*, garantindo média zero e variância unitária antes do treinamento dos modelos.

3.3 Modelagem, Arquitetura e Hiperparâmetros

A predição das notas foi realizada de forma independente para cada competência. Para respeitar a natureza discreta e ordenada das pontuações do Enem, formulamos o problema como regressão ordinal, adotando o *framework* CORAL (Cao et al., 2020). Em todos os experimentos, mantivemos fixa a arquitetura do modelo, variando apenas a dimensionalidade da camada de entrada conforme a representação utilizada, de modo a isolar o impacto das representações textuais. A Tabela 1 apresentada a dimensionalidade da camada de entrada para cada representação textual investigada neste estudo. No caso das representações híbridas, a camada de entrada foi ajustada à dimensionalidade do vetor produzido pela concatenação das representações em uso em cada experimento. Por exemplo, nos experimentos com a representação híbrida *embedding*+ características tarefa-específicas, a camada de entrada teve dimensão 795.

Representação	**Dimensão** (d)
TF-IDF	1.000
NILC-Metrix	200
Tarefa-específicas	27
Embeddings	768

Tabela 1: Dimensionalidade da camada de entrada para cada representação textual isolada.

A arquitetura empregada consistiu em um classificador neural do tipo *multilayer perceptron* (MLP), composto por três camadas ocultas totalmente conectadas, com 256, 128 e 64 neurônios, respectivamente, e funções de ativação ReLU. A camada de saída foi implementada por meio de uma `CoralLayer` com cinco unidades, correspondentes aos limiares ordinais entre os seis níveis de pontuação do Enem.

O treinamento foi realizado utilizando o otimizador Adam, com taxa de aprendizado inicial de $0,01$ e *batch size* de 128. Empregou-se *early stopping* com base no desempenho no conjunto de validação, monitorando a métrica QWK. Não foi

realizado ajuste extensivo de hiperparâmetros, de modo a preservar a comparabilidade experimental e manter o foco analítico sobre o papel das representações textuais. A Tabela 2 sumariza a arquitetura e hiperâmetros utilizados nos experimentos.

Componente	Configuração
Arquitetura	MLP
Camadas ocultas	$256 \rightarrow 128 \rightarrow 64$
Função de ativação	ReLU
Camada de saída	`CoralLayer` (6 unid.)
Otimizador	Adam
Taxa de aprendizado	0,01
Batch size	128
Critério de parada	*Early stopping* (QWK)
Normalização	*Z-score*

Tabela 2: Arquitetura e hiperparâmetros utilizados nos experimentos.

3.4 Protocolo Experimental

Todos os modelos foram treinados utilizando o conjunto de treino e selecionados com base no desempenho no conjunto de validação. A avaliação final foi realizada no conjunto de teste disponibilizado por Silveira et al. (2024). Para cada estratégia de representação, treinamos um modelo separado por competência, mantendo constantes os demais componentes do *pipeline*.

Além da avaliação no conjunto de teste completo, realizamos análises adicionais considerando o subconjunto não-divergente, permitindo investigar o impacto da variabilidade da anotação humana no desempenho dos modelos. A Figura 1 apresenta uma visão geral do *pipeline* experimental adotado neste estudo.

Como métrica principal de avaliação, utilizamos o QWK, amplamente empregado em tarefas de AAR por medir o grau de concordância entre predições automáticas e avaliações humanas, penalizando discrepâncias mais severas entre níveis distantes de pontuação (Attali e Burstein, 2006; Amorim e Veloso, 2017).

4 Resultados e Discussão

Nesta seção, apresentamos e discutimos os resultados obtidos para as diferentes estratégias de representação textual avaliadas. A análise é estruturada em três partes: (i) desempenho global médio dos modelos, (ii) desempenho por competência no conjunto de teste completo e no subconjunto não-divergente, e (iii) implicações da divergência entre avaliadores humanos na interpretação dos resultados.

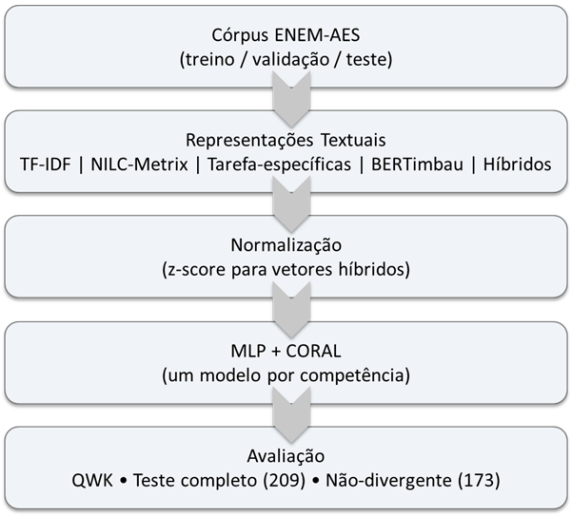

Figura 1: Visão geral do *pipeline* experimental.

4.1 Desempenho Global dos Modelos

A Tabela 3 apresenta o desempenho médio dos modelos, calculado como a média do QWK ao longo das cinco competências, considerando o conjunto de teste completo (209 redações). Observa-se que modelos baseados exclusivamente em representações lexicais tradicionais, como TF-IDF, apresentam o pior desempenho médio, enquanto abordagens baseadas em *embeddings* contextuais obtêm resultados substancialmente superiores.

Modelo	QWK
Embedding+Tarefa-específicas	**0,42**
Embedding	0,40
Tarefa-específicas	0,39
NILC-Metrix	0,34
TF-IDF	0,27

Tabela 3: Desempenho médio dos modelos.

Entre os modelos avaliados, a combinação de *embeddings* contextuais com características linguísticas orientadas à tarefa apresentou o melhor desempenho médio (QWK = 0,42), superando tanto o uso isolado de *embeddings* (QWK = 0,40) quanto o uso exclusivo de características manuais (QWK = 0,39). Esse resultado indica que informações linguísticas explicitamente projetadas para o domínio do Enem permanecem relevantes mesmo na presença de representações contextuais, reforçando a hipótese de complementaridade entre diferentes

tipos de representação textual.

4.2 Análise dos Modelos Híbridos

Para compreender melhor o impacto da combinação de representações, realizamos uma análise de ablação comparando diferentes configurações híbridas envolvendo *embeddings* e conjuntos de características linguísticas explícitas. Em particular, investigamos se a concatenação de múltiplas fontes de informação resulta necessariamente em ganhos adicionais de desempenho médio.

A Tabela 4 apresenta o QWK médio obtido por diferentes configurações híbridas, considerando o conjunto de teste completo. Observa-se que a combinação de *embeddings* com características tarefa-específicas apresenta o melhor desempenho médio, enquanto combinações mais extensas, envolvendo múltiplos conjuntos de características simultaneamente, não produzem ganhos adicionais consistentes. Esse comportamento sugere que a simples concatenação de representações pode introduzir redundância ou ruído, especialmente em cenários com dimensionalidade elevada e tamanho de conjunto de dados limitado.

Representação	**QWK**
Embedding	0,40
Embedding+Tarefa-específicas	**0,42**
TF-IDF+*Embedding*	0,39
NILC-Metrix+*Embedding*	0,37
TF-IDF+NILC-Metrix+*Embedding*	0,36

Tabela 4: Análise de ablação dos modelos híbridos.

4.3 Desempenho por Competência

A Tabela 5 detalha o desempenho dos cinco melhores modelos para cada competência, reportando os valores de QWK tanto no conjunto de teste completo quanto no subconjunto não-divergente (173 redações). Os resultados evidenciam que o desempenho varia significativamente entre competências, refletindo as diferentes dimensões linguísticas e semânticas avaliadas no Enem. A Tabela 6 resume o melhor desempenho observado em cada competência no conjunto de teste completo.

De modo geral, competências mais associadas a aspectos formais e estruturais da escrita tendem a apresentar melhores resultados na presença de representações linguísticas explícitas, enquanto competências relacionadas a conteúdo, argumentação e proposta de intervenção favorecem representações contextuais.

Na competência C1, relacionada ao domínio da norma padrão, observa-se desempenho equivalente entre diferentes representações, com destaque para o modelo híbrido *Embedding*+Tarefa-específicas (QWK = 0,47). Já para a competência C2, associada à compreensão do tema e ao desenvolvimento do texto dissertativo, o melhor desempenho foi obtido pelo modelo baseado exclusivamente em *embeddings* (QWK = 0,41). Esse comportamento indica que a modelagem de conteúdo semântico e coerência global do texto é favorecida por representações distribuídas aprendidas automaticamente. Curiosamente, na competência C3, relacionada à organização argumentativa, o melhor resultado foi alcançado pela combinação TF-IDF+*Embedding* (QWK = 0,45), sugerindo que sinais lexicais globais podem complementar os *embeddings* na captura da progressão argumentativa do texto. Na C4, associada a mecanismos linguísticos de coesão, os melhores desempenhos foram observados em modelos que incorporam características linguísticas explícitas, isoladamente ou em combinação com *embeddings*, com QWK máximo de 0,40. Esse padrão destaca a importância de atributos estruturais e discursivos para essa competência. Por fim, na competência C5, que avalia a elaboração de uma proposta de intervenção, o modelo baseado exclusivamente em *embeddings* apresenta o melhor desempenho (QWK = 0,51), indicando que essa competência demanda maior capacidade de modelagem semântica e pragmática, menos diretamente capturada por características linguísticas tradicionais.

4.4 Análise sob Divergência entre Avaliadores

Para investigar o impacto da variabilidade da anotação humana, comparamos o desempenho dos modelos no conjunto de teste completo e no subconjunto não-divergente, conforme apresentado na Tabela 5.

De forma geral, os resultados indicam que a redução da divergência entre avaliadores não implica, necessariamente, um aumento sistemático do desempenho dos modelos. Embora alguns modelos apresentem ganhos de QWK no subconjunto não-divergente, outros exibem quedas ou comportamento estável, dependendo da competência e da estratégia de representação adotada. Esse padrão evidencia que o subconjunto não-divergente não deve ser interpretado como um conjunto intrinsecamente mais fácil, mas como um conjunto com maior consistência de anotação e menor variabili-

Comp.	Modelo	Teste completo	Teste não-divergente
C1	*Embedding*+Tarefa-específicas	**0,47**	0,45
	Embedding	**0,47**	0,40
	Tarefa-específicas	**0,47**	0,52
	TF-IDF+*Embedding*	0,45	0,38
	NILC-Metrix+*Embedding*	0,43	0,44
C2	*Embedding*	**0,41**	0,23
	TF-IDF+NILC-Metrix+*Embedding*	0,38	0,32
	Embedding+Tarefa-específicas	0,35	0,31
	NILC-Metrix	0,34	0,28
	TF-IDF	0,31	0,28
C3	TF-IDF+*Embedding*	**0,45**	0,38
	Tarefa-específicas	0,41	0,44
	TF-IDF+Tarefa-específicas	0,39	0,35
	TF-IDF+*Embedding*+Tarefa-específicas	0,37	0,40
	Embedding	0,34	0,38
C4	*Embedding*+Tarefa-específicas	**0,40**	0,33
	TF-IDF+*Embedding*	**0,40**	0,33
	Embedding	0,39	0,40
	Tarefa-específicas	0,39	0,42
	NILC-Metrix+*Embedding*	0,36	0,31
C5	*Embedding*	**0,51**	0,53
	TF-IDF+*Embedding*	0,48	0,49
	Embedding+Tarefa-específicas	0,48	0,40
	TF-IDF+NILC-Metrix+*Embedding*	0,40	0,44
	TF-IDF	0,38	0,31

Tabela 5: QWK dos cinco melhores modelos por competência (ordenados por desempenho no teste completo).

Comp.	Melhor modelo	QWK
C1	*Embedding*+T.-específicas	0,47
C2	*Embedding*	0,41
C3	TF-IDF+*Embedding*	0,45
C4	*Embedding*+T.-específicas	0,40
C5	*Embedding*	0,51

Tabela 6: Melhor desempenho por competência

dade subjetiva.

Observa-se que modelos baseados em características linguísticas orientadas à tarefa tendem a se beneficiar de forma mais clara da redução da divergência, apresentando ganhos no subconjunto não-divergente em competências como C1, C3 e C4. Em contraste, modelos baseados exclusivamente em *embeddings* mostram comportamento menos estável, com melhorias pontuais em competências como C5, mas quedas em outras, como C2. Esse resultado sugere que representações linguísticas explícitas estão mais alinhadas a critérios avaliativos compartilhados entre avaliadores humanos, enquanto *embeddings* contextuais podem ser mais sensíveis à distribuição dos dados e à variabilidade das anotações.

Em conjunto, a análise evidencia que a divergência entre avaliadores impõe limites práticos ao desempenho da Avaliação Automática de Redações e afeta de maneira distinta diferentes estratégias de representação textual. Esses resultados reforçam a importância de considerar explicitamente a qualidade e a consistência da anotação humana na interpretação dos resultados e na comparação entre modelos.

4.5 Discussão

Os resultados obtidos fornecem evidências empíricas de que a escolha da representação textual exerce impacto significativo no desempenho da AAR no contexto do Enem. Diferentes competências demandam sinais linguísticos distintos, e nenhuma representação isolada é suficiente para capturar todas as dimensões avaliadas.

As abordagens híbridas exploradas neste tra-

balho mostraram-se particularmente eficazes ao integrar informações complementares, conciliando interpretabilidade parcial das características linguísticas com a capacidade de generalização dos *embeddings*. Ao mesmo tempo, a análise sob divergência humana evidencia que ganhos adicionais podem ser limitados pela própria inconsistência da anotação, ressaltando a necessidade de avaliações cuidadosas e protocolos controlados.

Em conjunto, essas observações indicam que os avanços em AAR para o Enem podem depender menos de arquiteturas mais complexas e mais de uma compreensão aprofundada do papel das representações textuais, da adequação às diferentes competências avaliadas e da qualidade da anotação dos córpus.

Adicionalmente, a análise de ablação indica que ganhos obtidos por modelos híbridos não decorrem simplesmente do aumento da dimensionalidade da representação, mas da combinação criteriosa de sinais complementares. Esse resultado sugere que estratégias mais sofisticadas de integração ou seleção de características podem ser mais promissoras do que concatenações extensivas, especialmente em cenários com conjuntos de dados de tamanho moderado.

5 Conclusão

Neste trabalho, investigamos o impacto de diferentes estratégias de representação textual na avaliação automática de redações do Enem, com foco na comparação entre características linguísticas explícitas, *embeddings* contextuais e suas combinações híbridas. A partir de um estudo empírico controlado no córpus Enem-AES, avaliamos os modelos no nível de competências, adotando um protocolo de regressão ordinal e métrica de concordância amplamente utilizadas na literatura.

Os resultados mostram que a escolha da representação textual exerce influência no desempenho dos modelos, tanto em termos de desempenho médio quanto no comportamento específico por competência. Representações baseadas exclusivamente em *embeddings* mostraram-se particularmente eficazes em competências associadas a conteúdo, argumentação e proposta de intervenção, enquanto características linguísticas orientadas à tarefa apresentaram maior estabilidade e melhor alinhamento com competências relacionadas a aspectos formais e estruturais da escrita. As abordagens híbridas, ao integrar informações complementares, alcançaram os melhores resultados médios, reforçando a hipótese de que nenhuma representação isolada é suficiente para capturar plenamente todas as dimensões avaliadas no Enem.

A análise sob diferentes níveis de divergência entre avaliadores humanos indica que a variabilidade da anotação impõe limites práticos ao desempenho da AAR e afeta de maneira distinta as estratégias de representação. Em particular, modelos baseados em características linguísticas explícitas tendem a se beneficiar mais de subconjuntos com uma anotação mais consistente, enquanto modelos dependentes de *embeddings* mostraram maior sensibilidade à distribuição dos dados. Essas observações reforçam a importância de interpretar os resultados à luz da qualidade da anotação humana e de evitar comparações baseadas apenas em métricas agregadas.

A partir deste estudo, diversas direções podem ser exploradas em trabalhos futuros. Uma primeira linha consiste em investigar estratégias mais sofisticadas de integração entre representações, que possam explorar de forma mais profunda as relações entre sinais linguísticos explícitos e representações contextuais. Adicionalmente, propõe-se examinar o impacto de modelos *Transformer* de maior escala e com capacidades nativas de raciocínio lógico (*reasoning*), como o DeepSeek-R1 e o Sabiá-3. Conforme discutido por Barbosa et al. (2025), esses modelos têm demonstrado avanços significativos na avaliação de competências que exigem maior profundidade semântica, como a argumentação (C3) e a compreensão temática (C2).

Outra direção futura relevante envolve a análise de interpretabilidade dos modelos, buscando compreender quais características textuais contribuem para a predição em cada competência e como essas contribuições se alinham aos critérios de correção do Enem. Para subsidiar essa análise, está prevista uma análise qualitativa dos erros e acertos cometidos pelos diferentes modelos e representações. Essa análise permitirá compreender melhor como o modelo híbrido integra sinais linguísticos explícitos e representações neurais para mitigar as inconsistências na avaliação de competências específicas.

Estudos futuros também podem considerar variações no protocolo experimental utilizado, incluindo a avaliação em outros conjuntos de dados ou exames de redação, bem como a incorporação de anotações com maior controle de qualidade ou múltiplos níveis de consenso entre avaliadores.

Limitações

Embora os resultados apresentados forneçam evidências empíricas relevantes sobre o papel das representações textuais na AAR do Enem, este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados.

Uma primeira limitação diz respeito ao conjunto de dados utilizado. Todos os experimentos foram conduzidos exclusivamente sobre o córpus Enem-AES, o que restringe a generalização dos achados para outros conjuntos de redações, rubricas ou contextos avaliativos. Embora o Enem represente um cenário de grande relevância prática, diferenças em critérios de correção ou em perfis de avaliadores podem influenciar o comportamento dos modelos.

Outra ameaça à validade está relacionada à qualidade e à consistência das anotações humanas. Conforme discutido ao longo do artigo, a divergência entre avaliadores impõe limites práticos ao desempenho dos modelos e afeta de maneira distinta as estratégias de representação textual. Embora tenhamos explorado subconjuntos não-divergentes para mitigar esse efeito, parte da variabilidade observada pode estar associada a ruídos inerentes às notas extraídas da plataforma *online*.

Do ponto de vista metodológico, o estudo adota um protocolo de modelagem fixo, com regressão ordinal e arquitetura neural constante para todas as representações avaliadas. Essa escolha visou isolar o impacto das representações textuais, mas limita a análise de interações entre representação e arquitetura, bem como a exploração de estratégias alternativas de modelagem.

Além disso, as abordagens híbridas investigadas baseiam-se em concatenação vetorial simples, o que pode não explorar plenamente as relações entre diferentes tipos de representação. Estratégias mais sofisticadas de combinação podem resultar em ganhos adicionais, mas não foram consideradas neste estudo para manter o foco empírico e a comparabilidade experimental.

Outra limitação refere-se à ausência de testes formais de significância estatística para a comparação entre os desempenhos dos modelos. Embora as análises apresentadas revelem tendências consistentes no comportamento das diferentes representações textuais, diferenças numéricas pequenas em métricas como QWK podem não ser estatisticamente significativas, especialmente em conjuntos de teste de tamanho moderado e sujeitos a ruído de anotação. Assim, os resultados devem ser interpretados de forma comparativa e descritiva, e não como evidência conclusiva de superioridade estatística entre modelos.

Dessa forma, os resultados reportados devem ser compreendidos como evidências empíricas obtidas em um escopo controlado, contribuindo para a compreensão do papel das representações textuais na AAR, mas sem pretensão de estabelecer limites definitivos para o desempenho alcançável por sistemas automáticos.

Referências

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

Yigal Attali e Jill Burstein. 2006. Automated essay scoring with e-rater [r] v. 2. *Journal of Technology, Learning, and Assessment*, 4(3).

André Barbosa, Igor Cataneo Silveira, e Denis Deratani Mauá. 2025. An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Wenzhi Cao, Vahid Mirjalili, e Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

Aluizio Haendchen Filho, Fernando Concatto, Hércules Antonio do Prado, e Edilson Ferneda. 2021. Comparing feature engineering and deep learning methods for automated essay scoring of brazilian national high school examination. Em *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, páginas 575–583. INSTICC, SciTePress.

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, e Alessandro Bokan. 2018. Automatically grading brazilian student essays. Em *International Conference on Computational Processing of the Portuguese Language*, páginas 170–179. Springer.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, e Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources & Evaluation*, 58:73—-110.

Feng Li, Xuefeng Xi, Zhiming Cui, Dongyang Li, e Wanting Zeng. 2023. Automatic essay scoring method based on multi-scale features. *Applied Sciences*, 13(11):6775.

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

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

Murilo Luis Calvo Neves. 2025. Extração e avaliação de indicadores lingísticos no contexto de correção automatizada de redações aplicadas ao exame nacional do ensino médio (enem). Trabalho de Conclusão de Curso (Graduação em Ciência da Computação), Universidade Estadual de Maringá, Maringá, PR, Brasil.

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

Igor Cataneo Silveira, André Barbosa, e Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in portuguese. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, páginas 228–237.

Fábio Souza, Rodrigo Nogueira, e Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. Em *Intelligent Systems*, páginas 403–417, Cham. Springer International Publishing.