Uso de técnicas de Aprendizado de Máquina e Modelos de Língua de Larga Escala para avaliação automática de textos do exame Celpe-Bras

Rafael Oleques Nunes, Bernardo Cobalchini Zietolie, Ricardo Zanini De Costa
Rodrigo Brock da Silva, João Victor Piardi Pacheco, Rafaela Dall'Agnol da Rocha
Dennis Giovani Balreira, Elisa Marchioro Stumpf, Juliana Roquele Schoffen

Universidade Federal do Rio Grande do Sul (UFRGS)

{ronunes, jvppacheco, dgbalreira}@inf.ufrgs.br

elisa.stumpf@ufrgs.br

{bernardozietolie, ricardozaninidecosta, rdrgbrck, r3ocha, julianaschoffen}@gmail.com

Abstract

O Celpe-Bras é o exame oficial brasileiro de proficiência em Português como Língua Adicional (Inep, 2020). A parte escrita do exame exige que os participantes produzam quatro textos em resposta a tarefas baseadas em vídeo, áudio e textos escritos , o que exige que a preparação para o exame seja realizada a partir de práticas de (re)escrita de textos. Por um lado, professores que trabalham na preparação de estudantes para o exame têm um alto volume de textos para corrigir, e os estudantes têm poucas opções de recursos didáticos acessíveis alinhados ao construto teórico do Celpe-Bras. Nesse contexto, e impulsionado pelos recentes avanços no Processamento de Linguagem Natural (PLN), modelos de língua de grande escala (LLMs) e Inteligência Artificial, este estudo visa mapear e comparar métodos para a avaliação automática dos textos produzidos no exame Celpe-Bras. São apresentados e testados diversos modelos, abrangendo tanto algoritmos tradicionais de aprendizado de máquina quanto modelos de linguagem pré-treinados, como BERT, BART e T5. Ao final, foi possível perceber que os melhores resultados foram obtidos pelas adaptações do modelo BERT, levemente superiores aos dos modelos restantes, mas com considerável maior custo computacional.

1 Introdução

O Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) é o certificado brasileiro de proficiência em português como língua adicional. Diferente de outros exames de proficiência que focam em habilidades isoladas, o Celpe-Bras adota uma perspectiva de uso da linguagem, em que a proficiência é demonstrada por meio do desempenho em tarefas que integram compreensão oral, leitura e produção escrita e oral de forma contextualizada (INEP, 2020). Por meio de uma única prova, o Celpe-Bras certifica quatro níveis de proficiência: Avançado Superior, Avançado, Intermediário Superior e Intermediário. A parte escrita do exame é composta por quatro tarefas integradas de compreensão e produção, que apresentam conteúdo de insumo em áudio, vídeo ou texto escrito, juntamente com um enunciado que solicita aos examinandos que produzam textos de diferentes gêneros discursivos (cartas, artigos de opinião, e-mails, entre outros) . Esse modelo de avaliação, embora pedagogicamente robusto, impõe desafios logísticos e operacionais significativos. No contexto de ensino preparatório para o exame, a avaliação dos textos dos estudantes é um processo oneroso e demorado, dificultando o acesso desses estudantes a feedbacks rápidos que podem ajudá-los a melhorar seu desempenho.

Nesse cenário, a Avaliação Automática de Redações (do inglês *Automated Essay Scoring* - AES) surge como uma solução promissora (Shermis and Wilson, 2024; Rassi and Lopes, 2024), usufruindo dos diferentes modelos e avanços em Processamento de Linguagem Natural (PLN). A transição de algoritmos clássicos de *Machine Learning*, como *Support Vector Machines* (SVM) e *Naive Bayes*, para arquiteturas baseadas em *Transformers* permitiu que as máquinas compreendessem melhor as nuances semânticas e contextuais dos textos e, consequentemente, tem gerado avanços na área de AES (Mizumoto and Eguchi, 2023; Pack et al., 2024).

Apesar da disponibilidade dessas tecnologias em línguas como o inglês, a aplicação de modelos avançados para o português brasileiro, especificamente no contexto do Celpe-Bras, ainda carece de estudos comparativos sistemáticos. O construto do exame, que valoriza a adequação ao gênero e ao propósito comunicativo acima da mera correção gramatical, apresenta um desafio adicional para os modelos de inteligência artificial.

Este trabalho tem como objetivo mapear e comparar a eficácia de diferentes abordagens de PLN

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 858–867

na atribuição automática de notas em tarefas do Celpe-Bras. Através de experimentos que utilizam desde modelos clássicos até modelos de língua de grande escala (*Large Language Models* - LLMs) pré-treinados, busca-se identificar o equilíbrio ideal entre precisão avaliativa e custo computacional, contribuindo para o desenvolvimento de ferramentas educacionais que auxiliem tanto professores quanto estudantes na preparação para o exame.

2 Trabalhos Relacionados

A avaliação automática de textos é uma área consolidada, especialmente no contexto de exames de proficiência em língua inglesa. Um dos sistemas mais conhecidos é o *e-rater*, utilizado no TOEFL, baseado em regressão linear múltipla e em um conjunto restrito de *features* linguísticas relacionadas à organização textual, complexidade lexical, vocabulário, extensão do texto e erros gramaticais (Attali and Burstein, 2004). Avaliações comparativas indicam que, entre modelos clássicos, o SVM apresenta desempenho superior a abordagens como regressão linear, *Random Forest* e *k-nearest neighbor*, alcançando valores de QWK considerados aceitáveis pelo ETS (Chen et al., 2016).

Comparações mais recentes entre modelos baseados em regras, aprendizado de máquina e modelos de linguagem mostram que abordagens baseadas em regras apresentam baixa concordância com avaliadores humanos, enquanto modelos supervisionados, como SVM e *Random Forest*, obtêm resultados substancialmente melhores, embora apresentem confusão recorrente entre níveis adjacentes de proficiência (Yeung, 2025). Nesse mesmo cenário, modelos de linguagem de grande porte alcançam os maiores valores de QWK e correlação, mas não superam o SVM em acurácia ou erro médio absoluto (Yeung, 2025).

Estudos comparativos entre modelos tradicionais e arquiteturas neurais indicam que, embora a acurácia global seja semelhante, arquiteturas baseadas em *transformers* tendem a apresentar melhores métricas de concordância ordinal. No corpus TOEFL11, o BERT supera modelos clássicos e outras redes neurais em QWK e concordância exata, apesar de modelos baseados em *features* linguísticas manterem vantagens em termos de interpretabilidade (Voss, 2025).

O impacto do *fine-tuning* em modelos de linguagem também tem sido explorado. Resultados mostram que modelos GPT ajustados superam abordagens baseadas exclusivamente em métricas de complexidade linguística, embora apresentem forte sensibilidade ao desbalanceamento dos dados, com desempenho significativamente inferior em níveis básicos de proficiência e com viéses de acordo com a L1 dos candidatos (Liu et al., 2025). Modelos baseados em *transformers* também apresentam resultados melhores do que as abordagens de ML com *features* linguísticas, como é o caso do estudo com corpus TCFLE-8, com textos de um teste de proficiência em francês (Wilkens et al., 2023).

No contexto da língua portuguesa, os estudos ainda são limitados. Apesar da existência de corpora de aprendizes[1], não há estudos sobre avaliação automática de textos produzidos em exames de proficiência. Em termos de classificação automática de textos, até o momento, temos conhecimento apenas da pesquisa de Nagasawa (2023) sobre a complexidade textual de textos de insumo do exame Celpe-Bras e de um estudo com textos de insumo de exames do Instituto Camões, com modelos pré-treinados para identificação do nível que destaca o desempenho do modelo Albertina PT-PT, mas aponta dificuldades nos níveis extremos da escala (Ribeiro et al., 2024).

3 Metodologia

3.1 Conjunto de Dados

O dataset utilizado nesta pesquisa faz parte do Corpus Celpe-Bras (CorCel) (Schoffen et al., 2025), que reúne textos escritos por examinandos e avaliados por examinadores oficiais do exame Celpe-Bras. Até o momento, o CorCel conta com 15.315 textos compilados (cerca de 3 milhões de palavras), produzidos em resposta às tarefas escritas de Celpe-Bras em quatro edições (2015-2, 2016-1, 2016-2 e 2017-1), divididos por tarefa (cada edição compreende quatro tarefas) e por nota (cada texto é avaliado com 0, 1, 2, 3, 4 ou 5). Não há metadados relativos aos examinandos que produziram os textos, mas há metadados detalhados sobre as tarefas que geraram esses textos (com base em Schoffen et al. (2018)) e a) a pontuação atribuída a cada texto; b) as notas obtidas pelo examinando nas outras tarefas da edição; c) a nota da parte oral do exame e d) o nível de certificação recebido pelo examinando.

[1] Recolha de Dados de Aprendizagem do Português como Língua Estrangeira, Corpus de Aquisição de L2 (CAL2), Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2), COPLE2, Macaws - Multilingual Academic Corpus of Assignments - Writing and Speech (Macaws)

Nota	Nº de Textos
0	25
1	211
2	628
3	715
4	477
5	237
Total	**2.293**

Table 1: Quantidade de textos do *dataset* distribuídos por nota avaliada.

Para os experimentos aqui relatados, de forma a iniciar a exploração dos dados, foi escolhido um recorte do conjunto de textos produzidos em resposta a uma tarefa, utilizado por Divino (2024) (n= 2.293), cujo critério de inclusão de textos era ter uma nota final inteira (entre 0 a 5), distribuídos como mostra a Tabela 1.

3.2 Modelos

Nesta seção, descrevemos as três abordagens investigadas para a avaliação automática do Celpe-Bras: (i) modelos clássicos de aprendizado de máquina, (ii) modelos neurais baseados exclusivamente em *encoder* e (iii) modelos neurais do tipo *sequence-to-sequence* (*encoder-decoder*). Essa diversidade de arquiteturas permite analisar o impacto do custo computacional e da capacidade de modelagem contextual no desempenho da tarefa.

Como *baselines* de baixo custo computacional, adotamos **modelos clássicos** de aprendizado de máquina, *Naive Bayes*, *Support Vector Machines* (SVM) e *Random Forest*, utilizando representações TF-IDF. A escolha desses algoritmos fundamenta-se em evidências prévias de que a distribuição e a frequência de termos são fortes preditoras de avaliações humanas em tarefas de proficiência linguística (Chen et al., 2016). Para reduzir a esparsidade inerente às representações vetoriais, aplicamos um pré-processamento composto pela remoção de *stopwords* e pelo *snowball stemming*. Essas etapas não foram empregadas nos modelos neurais subsequentes, uma vez que poderiam comprometer a integridade semântica necessária às representações contextuais aprendidas automaticamente.

No que diz respeito aos modelos neurais baseados em *encoder*, empregamos a arquitetura **BERT** (Devlin et al., 2018), amplamente reconhecida por sua eficácia em tarefas de compreensão textual graças ao uso de representações bidirecionais profundas (Deußer et al., 2024; Lukito et al., 2024; Barale et al., 2023; Peskine et al., 2023). Avaliamos três variantes de destaque para o português brasileiro: BERTimbau (Souza et al., 2020) nas versões *base* e *large*, Albertina (Rodrigues et al., 2023) na versão de 100m e BERTugues (Zago and dos Santos Pedotti, 2024) na versão *base* (única disponível). A inclusão dessas variantes permite investigar o impacto de diferentes estratégias de pré-treinamento e de tamanhos de corpus no desempenho da tarefa de avaliação automática.

Por fim, exploramos modelos neurais do tipo ***encoder-decoder***, que possibilitam a aprendizagem de representações densas e altamente contextuais a partir de sequências completas. Selecionamos o mBART-50 (Tang et al., 2020), em razão de sua robustez multilíngue e capacidade de processar sequências de entrada mais longas do que modelos puramente baseados em *encoder*. Além disso, utilizamos o ptt5-v2-base (Piau et al., 2024), cuja relevância decorre do pré-treinamento em um grande corpus de português brasileiro (BrWac) (Wagner Filho et al., 2018) e de sua validação prévia em tarefas de classificação e inferência. Essas características garantem maior aderência linguística e contextual ao domínio do exame Celpe-Bras.

3.3 Avaliação Experimental

A infraestrutura computacional utilizada consistiu em instâncias do *Google Colab* equipadas com GPUs NVIDIA Tesla T4. O framework *scikit-learn* foi adotado para a computação das métricas de avaliação, que incluíram acurácia, precisão, revocação, medida F1, área sob a curva ROC (AUC-ROC) e área sob a curva de precisão-revocação (AUC-PR). À exceção da acurácia, todas as métricas foram calculadas sob as óticas *macro* e *weighted*, o que permite uma análise robusta frente ao desbalanceamento entre as classes.

O protocolo de avaliação adotou uma estratégia de validação cruzada estratificada em 5 partições ($k = 5$). Em cada iteração, 10% do conjunto de treinamento foi reservado para validação, auxiliando no ajuste de parâmetros e na prevenção do sobreajuste (*overfitting*).

4 Resultados e discussão

Esta seção apresenta e discute os resultados experimentais a partir de um conjunto de perguntas orientadoras, formuladas de modo a organizar a análise de forma progressiva. Inicialmente, investiga-se

até que ponto representações lexicais e estatísticas superficiais são capazes de capturar padrões relevantes para a avaliação automática do Celpe-Bras. A partir desse diagnóstico, avalia-se se arquiteturas neurais oferecem ganhos consistentes em relação a abordagens clássicas. Em seguida, analisa-se como o desequilíbrio entre as classes afeta o desempenho preditivo dos modelos. Por fim, examina-se o impacto do pré-treinamento específico em português brasileiro no desempenho de modelos neurais aplicados à tarefa.

4.1 Até que ponto sinais lexicais e estatísticos são suficientes?

As Tabelas 2 e 3 apresentam os resultados obtidos com os modelos clássicos *Naive Bayes*, *Random Forest* e SVM, utilizando representações TF-IDF, avaliados sobre diferentes subconjuntos do *dataset*.

A análise do conjunto completo (Tabela 2) evidencia que o modelo SVM apresenta desempenho consistentemente superior ao dos demais modelos clássicos. Essa superioridade manifesta-se de forma particularmente clara no QWK, métrica mais adequada para tarefas de avaliação ordinal, o que indica que os erros do modelo tendem a se afastar menos das notas reais. Esse resultado está em consonância com achados prévios da literatura, como os de Chen et al. (2016), que apontam o SVM como uma abordagem robusta para tarefas de avaliação automática baseadas em sinais lexicais.

Além do QWK, o SVM também apresenta valores superiores de precisão e F1-macro, sugerindo maior capacidade de lidar com classes minoritárias. Ainda assim, os valores absolutos dessas métricas permanecem modestos, o que indica que representações puramente lexicais, embora informativas, são insuficientes para capturar plenamente a complexidade da tarefa de avaliação do Celpe-Bras.

A Figura 1a ilustra a matriz de confusão de uma das partições do modelo SVM sem pré-processamento. Observa-se uma concentração de predições ao longo da diagonal principal, com erros predominantemente restritos a classes adjacentes, sugerindo que o modelo é capaz de captar gradações de proficiência, embora de forma limitada.

O modelo *Random Forest* apresentou desempenho próximo, porém consistentemente inferior ao SVM, com diferenças mais pronunciadas nas métricas de precisão e QWK. Já o *Naive Bayes* apresentou desempenho substancialmente inferior. A análise das matrizes de confusão indica que esse comportamento decorre, em grande parte, da tendência do modelo a prever majoritariamente as classes mais frequentes, sobretudo em cenários desbalanceados.

No que diz respeito ao pré-processamento textual, observa-se que, na maioria dos cenários avaliados, sua aplicação resultou em reduções leves, porém consistentes, no desempenho dos modelos clássicos, com raras exceções, como no *Naive Bayes*. Esse resultado sugere que a remoção de certos padrões lexicais e morfológicos pode eliminar sinais relevantes para a predição das notas, especialmente quando representações simples como TF-IDF são empregadas. Em vez de um pré-processamento agressivo, abordagens baseadas na seleção dirigida de *features*, conforme discutido por Chen et al. (2016) e trabalhos correlatos, podem constituir um caminho mais promissor para aprimorar a interpretabilidade e o desempenho desses modelos.

Por fim, cabe notar que os modelos clássicos utilizados não foram submetidos a uma otimização extensiva de hiperparâmetros. Estratégias como *Grid Search* e *Random Search* poderiam elevar seus resultados. Essas direções são deixadas como possibilidade de trabalho futuro, reforçando o papel dos modelos clássicos como *baselines* informativos, porém estruturalmente limitados, na avaliação automática do Celpe-Bras.

4.2 Modelos neurais oferecem ganhos reais em relação às abordagens clássicas?

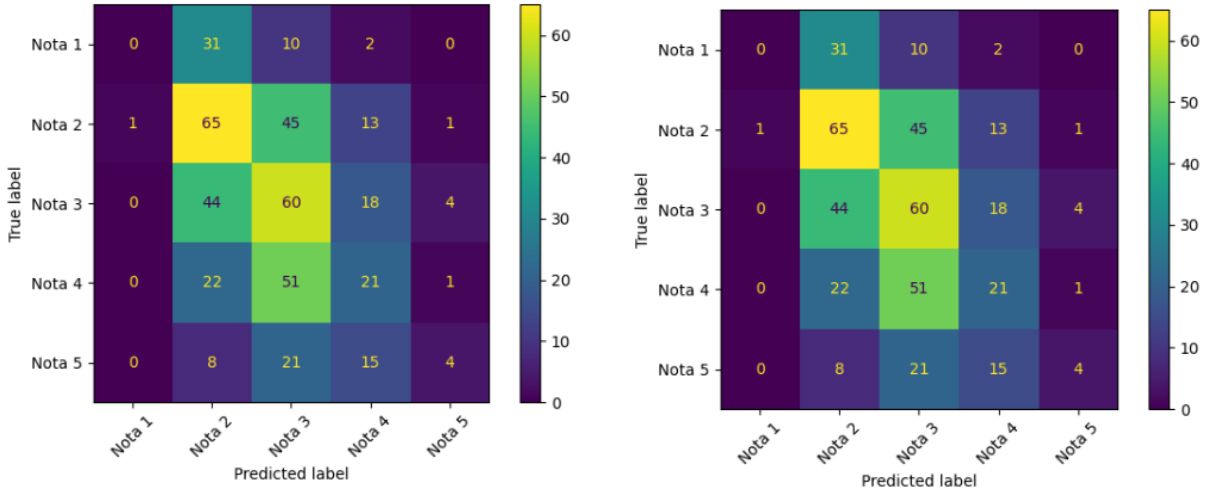
Quando comparados aos modelos clássicos, os resultados apresentados na Tabela 2 mostram ganhos consistentes nas métricas globais. Sob esse cenário, tanto as arquiteturas do tipo *encoder-decoder*, como BART e T5, quanto os modelos baseados exclusivamente em *encoder*, como BERT, superam o limite de desempenho imposto por representações lexicais e estatísticas superficiais.

Apesar desses avanços, os resultados também revelam limitações persistentes. Mesmo os modelos neurais com maior capacidade representacional apresentam erros predominantemente concentrados em classes majoritárias, tal qual as abordagens clássicas, como observado nas matrizes de confusão (Figuras 1a e 2a). Esse padrão sugere que, embora as arquiteturas neurais reduzam a influência de sinais superficiais, a distinção fina entre níveis próximos de proficiência permanece um desafio significativo.

Em conjunto, esses achados indicam que os gan-

		Macro					*Weighted*					
Modelo	**Acc**	**Prec**	**Rec**	**F1**	**AUC-ROC**	**AUC-PR**	**Prec**	**Rec**	**F1**	**AUC-ROC**	**AUC-PR**	**QWK**
Naive Bayes (NPP)	0,3127	0,0804	0,1673	0,0814	0,6454	0,2941	0,1439	0,3127	0,1517	0,6006	0,3031	0,0090
Naive Bayes (PP)	0,3157	0,1228	0,1692	0,0862	0,6407	0,2799	0,2136	0,3157	0,1597	0,5964	0,3018	0,0245
***Random Forest* (PP)**	0,3319	0,2789	0,1920	0,1595	0,6828	0,3177	0,3298	0,3319	0,2644	0,6153	0,3178	0,2151
***Random Forest* (NPP)**	0,3336	0,2303	0,1932	0,1609	0,6978	0,3251	0,3004	0,3336	0,2684	0,6315	0,3254	0,2361
SVM (PP)	0,3463	0,3071	0,2157	0,2046	0,7393	0,3600	0,3531	0,3463	0,3062	0,6626	0,3542	0,3342
SVM (NPP)	**0,3541**	**0,3481**	**0,2210**	**0,2106**	**0,7535**	**0,3917**	**0,3818**	**0,3541**	**0,3148**	**0,6759**	**0,3695**	**0,3695**
BART	0,3406	0,1552	0,1928	0,1500	0,6492	0,2624	0,2519	0,3406	0,2589	0,6089	0,3088	0,1817
T5	**0,3742**	**0,1920**	**0,2348**	**0,2051**	**0,7378**	**0,3557**	**0,3069**	**0,3742**	**0,3280**	**0,6846**	**0,3676**	**0,4397**
Albertina	0,3217	**0,3843**	0,2025	0,1853	0,7621	0,4013	**0,4039**	0,3217	0,2656	0,6787	0,3726	0,3576
Bertuguês	0,3348	0,2075	0,2106	0,1941	0,7659	0,3611	0,3104	0,3348	0,3019	0,6847	0,3776	0,4121
BERTimbau *base*	**0,4130**	0,2133	0,2487	0,2233	0,7723	0,3709	0,3372	**0,4130**	**0,3612**	0,7055	0,3999	0,4464
BERTimbau *Large*	0,3826	0,3310	**0,2767**	**0,2719**	**0,7936**	**0,4033**	0,3789	0,3826	0,3472	**0,7155**	**0,4088**	**0,5516**

Table 2: Resultados sobre o *dataset* completo ordenados pelo QWK.



(a) Conjunto de textos completo.

(b) *Dataset* otimizado.

Figure 1: Matrizes de confusão do modelo SVM sem pré-processamento em diferentes configurações do *dataset*.

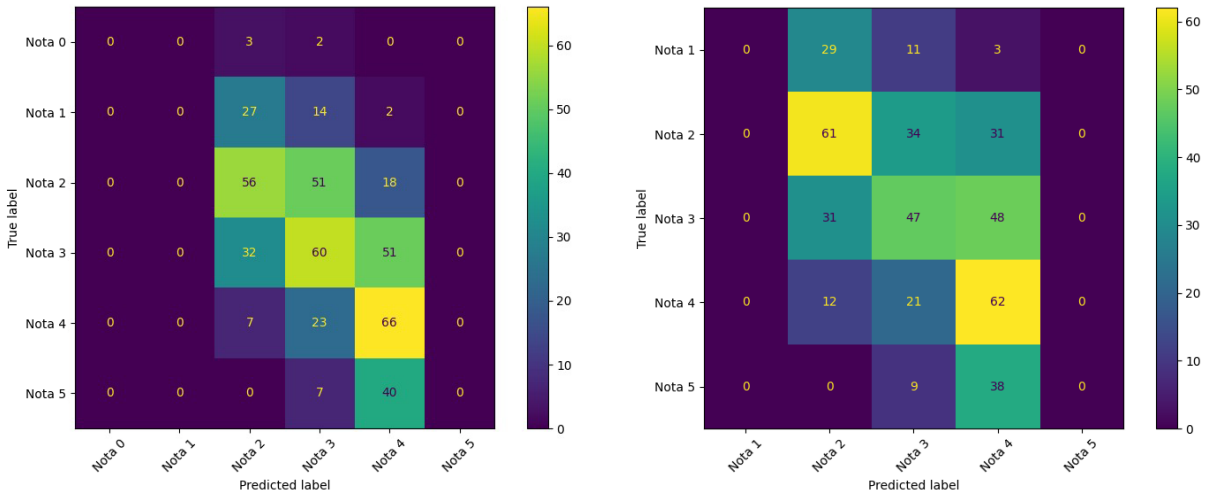
hos proporcionados por modelos neurais são reais e mensuráveis, mas condicionados à qualidade e ao equilíbrio dos dados. A avaliação automática de proficiência escrita continua sendo uma tarefa complexa, na qual o aumento da capacidade do modelo, embora necessário, não é suficiente para eliminar ambiguidades inerentes à natureza gradual e subjetiva da avaliação humana.

4.3 Como o desbalanceamento entre classes afeta o comportamento dos modelos?

Independentemente da arquitetura considerada, os resultados evidenciam um padrão consistente de viés preditivo em direção às classes intermediárias, especialmente as notas 2 e 3, que concentram a maior parte das amostras no *dataset* original. Esse comportamento é observável de forma transversal nas matrizes de confusão dos modelos clássicos (Figura 1a) e dos modelos neurais (Figuras 2a e 3a), indicando que o efeito não está restrito a uma arquitetura específica.

No conjunto completo, esse viés manifesta-se como um efeito de atração das classes centrais, no qual textos avaliados com notas extremas tendem a ser sistematicamente reclassificados como pertencentes a classes adjacentes mais frequentes. Nas matrizes de confusão dos modelos neurais (Figuras 2a e 3b), observa-se, por exemplo, a ausência ou escassez de acertos para as notas 0, 1 e 5, enquanto as classes 2, 3 e 4 concentram praticamente todas as predições corretas. Esse padrão também se repete, ainda que de forma menos acentuada, nos modelos clássicos (Figura 1a).

Esse comportamento tem impacto direto nas métricas globais. Embora a acurácia permaneça relativamente estável, métricas sensíveis ao equilíbrio entre classes e à ordenação, como F1-macro e QWK, são significativamente penalizadas, refletindo a incapacidade dos modelos de distinguir adequadamente níveis extremos de proficiência. Assim, o desbalanceamento atua como um fator estruturante do erro, e não apenas como uma limitação quantitativa do *dataset*.



(a) Conjunto de textos completo.

(b) *Dataset* otimizado.

Figure 2: Matrizes de confusão do modelo PTT5-v2 em diferentes configurações do *dataset*.

		Média *Macro*					Média *Weighted*					
Modelo	**Acc**	**Prec**	**Rec**	**F1**	**AUC-ROC**	**AUC-PR**	**Prec**	**Rec**	**F1**	**AUC-ROC**	**AUC-PR**	**QWK**
Naive Bayes (NPP)	0,3158	0,1550	0,2191	0,1411	0,6190	0,2867	0,2233	0,3158	0,2033	0,6085	0,3137	0,1335
Naive Bayes (PP)	0,3313	0,1916	0,2301	0,1575	0,6176	0,2841	0,2625	0,3313	0,2266	0,6065	0,3126	0,1534
***Random Forest* (PP)**	0,3295	0,2341	0,2356	0,1989	0,6525	0,3015	0,2783	0,3295	0,2714	0,6224	0,3159	0,2281
***Random Forest* (NPP)**	0,3410	0,2799	0,2460	0,2129	0,6690	0,3236	0,3082	0,3410	0,2867	0,6362	0,3333	0,2700
SVM (PP)	**0,3712**	**0,4563**	**0,2879**	**0,2816**	0,7046	0,3629	**0,4113**	**0,3712**	**0,3392**	0,6693	0,3630	0,3736
SVM (NPP)	0,3694	0,4193	0,2807	0,2667	**0,7176**	**0,3819**	0,3927	0,3694	0,3327	**0,6812**	**0,3794**	**0,3940**
BART	0,3306	0,1816	0,2344	0,1915	0,6303	0,2917	0,2473	0,3306	0,2668	0,6114	0,3168	0,2238
T5	**0,3682**	**0,2249**	**0,2884**	**0,2401**	**0,7078**	**0,3647**	**0,3011**	**0,3682**	**0,3152**	**0,6770**	**0,3771**	**0,4571**
BERTimbau (*base*)	0,3973	0,2389	0,3000	0,2639	0,7612	0,4105	0,3205	0,3973	0,3519	0,7216	0,4119	0,5299
Bertuguês	**0,4201**	0,3786	0,3563	0,3559	**0,7636**	**0,4587**	0,3961	**0,4201**	0,3999	**0,7292**	**0,4585**	0,5692
Albertina	0,3653	0,3348	0,3135	0,3133	0,7370	0,3816	0,3493	0,3653	0,3497	0,6971	0,3935	0,5747
BERTimbau (*large*)	0,4155	**0,4652**	**0,4127**	**0,4180**	0,7611	0,4484	**0,4290**	0,4155	**0,4079**	0,7158	0,4308	**0,6305**

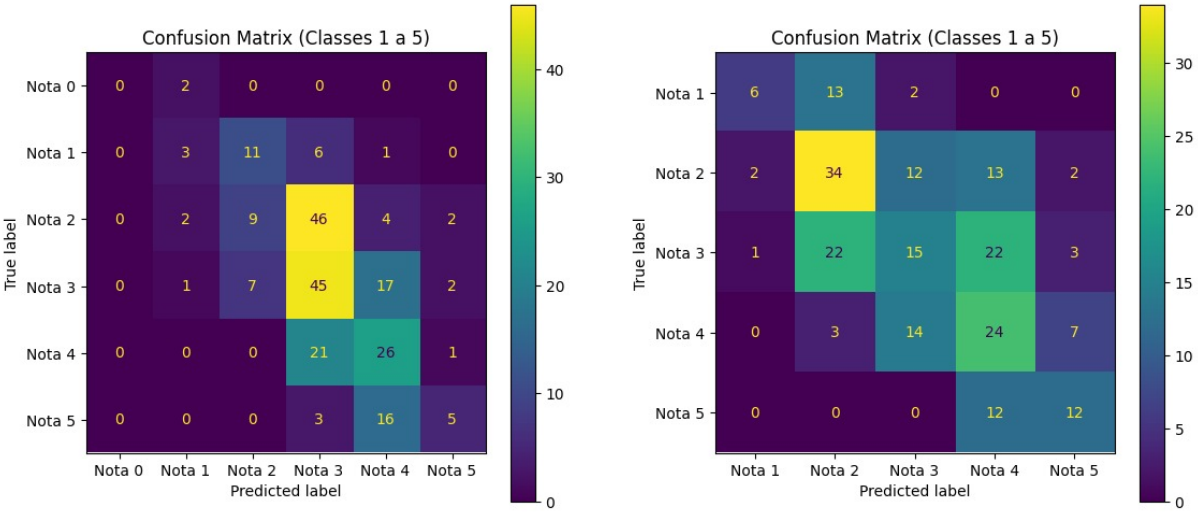
Table 3: Resultados sobre o *dataset* "otimizado" ordenados pelo QWK.

Com o objetivo de isolar e investigar esse efeito, foi construído um *dataset* otimizado. Nesse cenário, textos de nota 0 foram removidos, tanto por sua baixa representatividade quanto por sua natureza atípica no contexto da avaliação de proficiência linguística, uma vez que essa nota pode refletir fatores extralinguísticos, como a fuga do tema ou do gênero discursivo. A motivação para essa decisão é sustentada empiricamente pelas matrizes de confusão do conjunto completo (Figuras 2a e 3a), nas quais a classe 0 é sistematicamente ignorada pelos modelos, mesmo quando presente no treinamento.

Além disso, as classes 2 e 3, majoritárias no *dataset* original, foram balanceadas por subamostragem aleatória. As matrizes de confusão dos modelos neurais no cenário não otimizado evidenciam um viés recorrente em direção à classe 3, que atua como classe de convergência para predições incorretas, caracterizando um efeito de atração das classes centrais. Ao reduzir artificialmente esse desequilíbrio, observa-se uma redistribuição das predições, com menor concentração em uma única classe intermediária.

Os efeitos dessa intervenção são refletidos nos resultados apresentados na Tabela 3, que mostram aumentos consistentes nas métricas F1-macro e QWK para os modelos neurais. Esses ganhos indicam não apenas uma melhora quantitativa, mas também uma mudança qualitativa no padrão de erro, corroborada pelas matrizes de confusão do *dataset* otimizado (Figuras 2b e 3b), nas quais a dispersão das predições ao longo da diagonal principal torna-se mais equilibrada.

A avaliação do *dataset* otimizado foi restrita aos modelos neurais, uma vez que os resultados dos modelos clássicos no cenário original (Tabela 2) indicam que essas abordagens atingem rapidamente um platô de desempenho, com ganhos marginais mesmo sob condições mais favoráveis de balanceamento. Assim, a análise do impacto do desbalanceamento é direcionada às arquite-



(a) Conjunto de textos completo.

(b) *Dataset* otimizado.

Figure 3: Matrizes de confusão do modelo BERTimbau (*large*) em diferentes configurações do *dataset*.

turas com maior capacidade representacional, que demonstraram efetivamente se beneficiarem do controle da distribuição das classes.

4.4 Qual é o impacto do pré-treinamento em português brasileiro?

A comparação entre modelos neurais com diferentes estratégias de pré-treinamento revela que o pré-treinamento específico em português brasileiro exerce impacto relevante no desempenho da tarefa, sobretudo em cenários desbalanceados.

No conjunto completo, modelos monolíngues, como o PTT5-v2 e o BERTimbau, apresentaram desempenho superior às alternativas multilíngues, sugerindo maior robustez inicial decorrente da adaptação linguística. Em particular, o PTT5-v2 superou o mBART-50 em todas as métricas, enquanto o BERTimbau apresentou vantagem em relação a todos os outros modelos em grande parte das métricas.

Após a otimização do *dataset*, observou-se que modelos multilíngues, como o mBART-50, apresentaram ganhos mais expressivos, reduzindo significativamente a diferença em relação aos modelos monolíngues. Ainda assim, o BERTimbau (*large*) destacou-se como o melhor modelo global, alcançando os maiores valores de F1-macro e QWK entre todas as abordagens avaliadas.

Esses resultados indicam que o pré-treinamento em português brasileiro confere vantagens iniciais importantes, especialmente em cenários realistas e desbalanceados, mas que estratégias adequadas de seleção e balanceamento de dados podem mitigar parte dessa vantagem. Em conjunto, os achados reforçam que o desempenho na avaliação automática do Celpe-Bras depende tanto da qualidade do pré-treinamento quanto do cuidado no desenho experimental.

5 Conclusão

A pesquisa demonstra que a avaliação automática de textos produzidos no exame Celpe-Bras é viável, embora envolva alguns desafios. Em primeiro lugar, a predominância de estudos baseados em dados em inglês dificulta a realização de comparações diretas. Além disso, ainda há escassez de recursos para o processamento automático do português, o que limita a obtenção de métricas relacionadas a *features* linguísticas que poderiam contribuir para aprimorar os resultados. Soma-se a isso o fato de que os textos do corpus apresentam diversas formas não padrão de uso da língua, o que demanda estratégias adicionais para sua interpretação. Nesse contexto, o presente estudo configura-se como um ponto de partida, capaz de indicar caminhos para o desenvolvimento de métodos de avaliação automática desses textos.

Os dados experimentais demonstram que, embora algoritmos tradicionais de *Machine Learning* ofereçam um ponto de partida para a classificação, são os modelos de língua pré-treinados, como o BERTimbau e o BERTuguês, que atingem os melhores índices de desempenho. O destaque para o BERTimbau (*large*) Opt, que alcançou os maiores valores de acurácia (0.4128) e F1-Score (0.4049) nos dados otimizados, reforça a superioridade das

arquiteturas baseadas em *Transformers* na captura de nuances semânticas e contextuais.

As adaptações do modelo BERT, BERTuguês e BERTimbau apresentaram resultados superiores aos demais modelos em praticamente todas as métricas, com valores satisfatórios de QWK (por volta de 0,59 para o *dataset* completo e 0,57 para o *dataset* otimizado), o que indica certa concordância entre notas preditas e as reais, de forma condizente com os resultados obtidos por Voss (2025). Nesse sentido, a adaptação Albertina apresentou resultados inferiores nas métricas de Recall Macro, F1 Macro e QWK, devido à maior quantidade de previsões para as notas 2 e 3, indicando que o modelo teve dificuldades na previsão de classes minoritárias, com valores próximos aos retornados por modelos clássicos e pelos modelos BERT e T5.

O modelo clássico SVM e o modelo T5 apresentaram resultados semelhantes entre si, com o modelo T5 apresentando recall ligeiramente melhor e precisão consideravelmente inferior, o que sinaliza uma possível dificuldade do modelo para a previsão de classes minoritárias. Essa hipótese pode ser fortalecida pela presença de outros valores inferiores para as métricas macro no modelo T5, como recall macro, F1 macro e AUC PR macro.

Os modelos clássicos *Naive Bayes* e RF, juntamente com o modelo BART, apresentaram resultados insatisfatórios em comparação com os demais modelos analisados, com destaque especial para o algoritmo *Naive Bayes*, incapaz de prever corretamente classes minoritárias em um *dataset* desbalanceado. Em trabalhos futuros, o aumento do *dataset* pode fortemente aprimorar as métricas obtidas com o modelo BART. Por fim, torna-se claro que as variações do modelo BERT apresentaram melhores resultados, o que já era esperado diante da clara vantagem dessas arquiteturas na percepção de padrões contextuais mais complexos. Contudo, modelos mais simples, e de consequente treinamento mais rápido, como o SVM, apesar de não apresentarem resultados equiparáveis aos modelos de arquitetura *Transformer*, podem fornecer uma solução de qualidade ligeiramente inferior, por um custo computacional mínimo, principalmente quando em conjunto com métricas baseadas em *features*.

6 Próximos Passos

Os próximos passos para esta pesquisa envolvem a adoção de estratégias para melhorar os resultados dos modelos testados, bem como a possível adoção de novos modelos. Para os modelos clássicos, a seleção e utilização de *features* linguísticas podem ser extremamente eficientes, permitindo maior número de acertos em previsões para notas extremas.Além disso, pretende-se também utilizar mais textos do CorCel, aumentando a diversidade de gêneros discursivos representados. Os modelos de grande escala podem se beneficiar do aumento do *dataset* utilizado, além de seu balanceamento, o que não foi possível para o estudo atual. Por fim, a otimização de hiperparâmetros pode ser aplicada a todos os modelos testados, o que pode aprimorar os resultados obtidos até certo ponto.

Agradecimentos

Este artigo é resultado parcial do projeto de pesquisa intitulado “Desenvolvimento de ferramentas de inteligência artificial a partir de modelos de linguagem de grande escala para a descrição de proficiência linguística e elaboração de recursos pedagógicos em português como língua adicional”, financiado pelo CNPq (Chamada CNPq/MCTI/FNDCT No 22/2024). Os autores agradecem ainda o financiamento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, da Universidade Federal do Rio Grande do Sul (UFRGS), da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) e o apoio da Petrobras.

References

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2):i–21.

Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Do language models learn about legal entity types during pretraining? In *Proceedings of the Natural Legal Language Processing Workshop*, volume 2023, pages 25–37.

Jing Chen, James H Fife, Isaac I Bejar, and André A Rupp. 2016. Building e-rater® scoring models using machine learning methods. *ETS Research Report Series*, 2016(1):1–12.

Tobias Deußer, Cong Zhao, Lorenz Sparrenberg, Daniel Uedelhoven, Armin Berger, Maren Pielka, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. 2024. A comparative study of large language models for named entity recognition in the legal domain. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4737–4742. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Luiza Sarmento Divino. 2024. Contribuições da linguística de corpus para a definição de níveis de proficiência escrita no exame celpe-bras. Master's thesis, Universidade Federal do Rio Grande do Sul.

INEP. 2020. *Documento base do exame Celpe-Bras*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Yingying Liu, Huilei Qi, and Xiaofei Lu. 2025. Enhancing gpt-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures. *Computer Assisted Language Learning*, pages 1–20.

Josephine Lukito, Bin Chen, Gina M. Masullo, and Natalie Jomini Stroud. 2024. Comparing a BERT classifier and a GPT classifier for detecting connective language across multiple social media. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19140–19153, Miami, Florida, USA. Association for Computational Linguistics.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Ellen Yurika Nagasawa. 2023. O conteúdo de insumo em tarefas que integram leitura e escrita no celpe-bras: uma abordagem informada por corpus. Doctoral dissertation.

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

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding gpt for multi-label classification. In *EMNLP 2023, Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marcos Piau, Roberto Lotufo, and Rodrigo Nogueira. 2024. ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language.

Amanda Pontes Rassi and Priscilla de Abreu Lopes. 2024. Correção automática de redação. In Helena de Medeiros Caseli and Maria das Graças Volpe Nunes, editors, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português*, 3 edition, pages 610–633. BPLN, São Carlos.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Avaliação automática do nível de complexidade de textos em português europeu. *Linguamática*, 16(2):115–139.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

Juliana Schoffen, Margarete Schlatter, Simone Paula Kunrath, Ellen Yurika Nagasawa, Gabrielle Rodrigues Sirianni, Kaiane Mendel, Luana Ramos Truyllio, and Luiza Sarmento Divino. 2018. Estudo descritivo das tarefas da parte escrita do exame celpe-bras: Edições de 1998 a 2017. Technical report, Porto Alegre.

Juliana Roquele Schoffen, Elisa Marchioro Stumpf, Luiza Sarmento Divino, Isadora Dahmer Hanauer, Deise Amaral, Amanda Michel Raupp, and Brenda de Souza Xavier. 2025. Corcel: A brazilian portuguese corpus of celpe-bras exam written texts. *Revista Brasileira de Linguística Aplicada*, 25(1):e50034.

Mark D Shermis and Joshua Wilson. 2024. Introduction to automated essay evaluation. In *The Routledge international handbook of automated essay evaluation*, pages 3–22. Routledge.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Erik Voss. 2025. Comparison of traditional machine learning and neural network approaches for automated scoring of second language english essays. *Language Testing*, 42(4):369–396.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. Tcfle-8: a corpus of learner written productions for french as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465.

Steven Yeung. 2025. A comparative study of rule-based, machine learning and large language model approaches in automated writing evaluation (awe). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 984–991.

Ricardo Mazza Zago and Luciane Agnoletti dos Santos Pedotti. 2024. Bertugues: A novel bert transformer model pre-trained for brazilian portuguese. *Semina: Ciências Exatas e Tecnológicas*, 45:e50630–e50630.