CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Classificação ou Regressão? Avaliando Coesão Textual em Redações no contexto do ENEM

Hilário Oliveira[1], Rafael Ferreira Mello[2,3], Péricles Miranda[2], Bruno Alexandre[3], Thiago Cordeiro[4], Ig Ibert Bittencourt[4,6], Seiji Isotani[5,6]

[1] Instituto Federal do Espírito Santo - Campus Serra

[2]Universidade Federal Rural de Pernambuco

[3]CESAR School

[4]Universidade Federal de Alagoas

[5]Universidade de São Paulo

[6]Harvard Graduate School of Education

hilario.oliveira@ifes.edu.br, {pericles.miranda,rafael.mello}@ufrpe.br

babr@cesar.school, thiago@ic.ufal.br

{seiji_isotani,ig_bittencourt}@gse.harvard.edu

Abstract. *The textual production of essays is an important step in the teaching-learning process, allowing students to express their ideas. Textual cohesion is a fundamental criterion in this context. Despite the interest in automated essay assessment approaches, only a few studies focus on textual cohesion in essays written in Brazilian Portuguese. This research explores three machine learning approaches, comparing classification and regression, to estimate grades related to cohesion in essays in the ENEM context. The approaches use the TF-IDF measure, contextual embedding representations, and BERT-based models. Experiments were performed using 6,563 essays from the extended Essay-BR corpus. The BERTimbau base model for classification obtained the best performance, with a moderate Pearson correlation and a fair agreement based on the linear Kappa coefficient regarding the human evaluators' scores.*

Resumo. *A produção textual de redações é uma etapa importante no processo de ensino-aprendizagem, pois permite aos alunos expressarem suas ideias. A coesão textual é um critério fundamental nesse contexto. Apesar do interesse em abordagens automatizadas para avaliação de redações, ainda existem poucos estudos que focam na coesão textual em redações escritas em português do Brasil. Este trabalho investiga três abordagens de aprendizado de máquina, comparando o uso de classificação e regressão, para estimar notas relacionadas à coesão de redações no contexto do ENEM. As abordagens investigadas utilizam a medida TF-IDF, representações contextuais multidimensionais e o uso de modelos baseados no BERT. Experimentos foram realizados usando 6.563 redações do corpus Essay-BR estendido. O modelo BERTimbau base para classificação obteve o melhor desempenho, com uma correlação moderada de Pearson e um nível razoável de concordância, com base no coeficiente linear de Kappa, em relação às notas dos avaliadores humanos.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM)[1] é uma das principais avaliações educacionais realizadas no Brasil. Ele foi criado em 1998 pelo Ministério da Educação (MEC) e tem como objetivo principal avaliar o desempenho dos estudantes que concluem o ensino médio. O ENEM é utilizado como critério de seleção em diversas instituições de ensino superior do país, como universidades e faculdades [Marinho et al. 2021]. Sua importância consiste no fato de que o exame oferece uma medida abrangente das habilidades e competências dos estudantes, abordando diferentes áreas do conhecimento, como matemática, linguagens, ciências humanas e da natureza.

A prova de produção textual é um dos elementos centrais do ENEM, exigindo que os estudantes redijam uma redação dissertativa-argumentativa acerca de um tema específico relacionado a áreas científicas, culturais, políticas ou sociais [Klein and Fontanive 2009]. Conforme as diretrizes da prova de redação, é necessário que o participante elabore um texto coeso e coerente, aderindo às normas da escrita formal da língua portuguesa do Brasil, ao defender uma tese apoiada por argumentos fundamentados em relação ao tema proposto. A avaliação é feita por examinadores humanos que consideram as cinco competências a seguir: **(i)** Domínio da escrita formal da língua portuguesa; **(ii)** Compreensão da temática proposta; **(iii)** Seleção e organização das informações; **(iv)** Demonstração do conhecimento dos recursos linguísticos necessários para a construção da argumentação; e **(v)** Formulação de uma proposta de intervenção para o problema abordado, pautada nos princípios dos direitos humanos.

A correção manual de redações é uma atividade que demanda tempo e esforço, podendo ser subjetiva em alguns casos [Costa et al. 2020, de Lima et al. 2023]. Avaliadores humanos podem apresentar tendências pessoais em relação a um determinado tópico abordado na redação, o que pode gerar inconsistências na correção. A criação de sistemas computacionais capazes de avaliar automaticamente redações com base em critérios estabelecidos pode ajudar a lidar com as demandas de tempo e os desafios de consistência na avaliação [Ferreira-Mello et al. 2019, de Lima et al. 2023]. Além disso, ao analisar os recursos linguísticos presentes nas redações, esses sistemas podem identificar informações relevantes sobre possíveis problemas, por exemplo, uso inadequado de conectivos ou repetição excessiva das palavras. Dessa forma, esses sistemas têm o potencial de auxiliar os educadores na adaptação do ensino em sala de aula, como revisar estratégias de devolutivas formativas para incluir aspectos da escrita que precisam ser aprimorados pelos estudantes [Ke and Ng 2019, Ramesh and Sanampudi 2022].

Embora estudos anteriores tenham abordado o desenvolvimento de modelos para a avaliação automática de redações com base nas competências do ENEM [Marinho et al. 2022b, Oliveira et al. 2022, Oliveira et al. 2023], ainda não foi criada uma ferramenta amplamente utilizada que seja capaz de avaliar automaticamente as diferentes competências consideradas na prova [de Lima et al. 2023]. Este trabalho busca contribuir nesse contexto, tendo como foco a competência 4 do ENEM, que é relacionada à coesão textual da redação e é considerada uma das mais desafiadoras para os candidatos [Klein and Fontanive 2009, Lima et al. 2018].

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

A coesão textual é um elemento essencial de um texto escrito de maneira formal e está relacionada ao emprego de recursos linguísticos que permitem a interligação entre os elementos do texto, como palavras, frases e parágrafos [Antunes 2005, Crossley 2020]. Dispositivos coesos (ou seja, palavras, frases e sentenças específicas) criam conexões entre os conceitos no texto. Ao usar esses dispositivos, o autor pode transmitir com maior clareza suas concepções e, assim, orientar o leitor a interpretar e conectar corretamente as ideias em um modelo mental coerente (por exemplo, um argumento que contém um conjunto de afirmações bem conectadas apoiado por evidências) [Graesser et al. 2004]. Espera-se que melhorar a coesão da redação, por sua vez, beneficie a sua qualidade geral em outros aspectos [Crossley 2020].

O objetivo deste trabalho é realizar um estudo comparativo do desempenho de algoritmos de classificação e regressão para estimar automaticamente as notas de coesão textual em redações no estilo do ENEM. Para isso, foram analisadas três abordagens que se baseiam diretamente nos textos das redações, em conjunto com algoritmos de Aprendizado de Máquina (AM) tradicionais e redes neurais profundas. As abordagens consideradas foram: a estratégia clássica de representação utilizando a medida de *Term Frequency–Inverse Document Frequency* (TF-IDF), a utilização de representações contextuais multidimensionais (*contextual embeddings*) das redações e o uso do modelo neural de linguagem *Bidirectional Encoder Representations for Transformers* (BERT). Essas abordagens foram selecionadas por apresentarem diferentes complexidades e métodos de reconhecimento de padrões, amplamente utilizados na literatura para tarefas envolvendo o uso de técnicas de Processamento de Linguagem de Natural (PLN) [Li et al. 2022]. As seguintes questões de pesquisa guiaram o desenvolvimento deste trabalho:

> **Pergunta de Pesquisa 1 (PP1):** *Qual a tarefa de aprendizado de máquina supervisionado (classificação ou regressão) é mais eficiente para modelar o problema de estimar as notas relacionadas à coesão textual (Competência 4) das redações no estilo do ENEM?*

> **Pergunta de Pesquisa 2 (PP2):** *Qual das abordagens baseadas diretamente nos textos das redações possui melhor desempenho para avaliar a coesão textual das redações no contexto do ENEM?*

Para responder às questões de pesquisa anteriores, foram realizados experimentos utilizando a base de dados do Essay-BR estendido [Marinho et al. 2022a], composta por 6.563 redações, juntamente com suas respectivas notas, seguindo os mesmos critérios de avaliação utilizados no ENEM. Os experimentos envolveram a avaliação de diferentes algoritmos de classificação e regressão, utilizando as medidas de desempenho do coeficiente linear de *Kappa*, a raiz quadrada do erro médio e o coeficiente de correlação de *Pearson*. Os resultados experimentais demonstraram que os modelos baseados no *BERTimbau* (BERT para o português do Brasil) apresentaram os melhores desempenhos em comparação com as outras abordagens consideradas, tanto para classificação quanto para regressão. Especificamente, o modelo base do *BERTimbau*, treinado por 5 épocas considerando a tarefa de classificação, obteve o melhor desempenho global, com um coeficiente linear de *Kappa* de 0,398, que representa um nível razoável de concordância, e uma correlação moderada de *Pearson* (0,668) com as notas atribuídas à coesão textual por avaliadores humanos.

2. Trabalhos Relacionados

A avaliação de redações de maneira imparcial e mantendo uma alta qualidade, além de oferecer *feedbacks* construtivos para auxiliar os estudantes, demanda muito tempo e esforço dos professores e avaliadores. Por esse motivo, diversas pesquisas têm investigado métodos baseados em Inteligência Artificial para desenvolver sistemas computacionais capazes de avaliar automaticamente redações no contexto educacional, visando agilizar o processo e melhorar a sua eficiência [Ke and Ng 2019, Ramesh and Sanampudi 2022]. Uma visão mais abrangente sobre a avaliação automática de redações, especialmente para o português do Brasil, pode ser encontrada nas revisões sistemáticas da literatura apresentadas em [Costa et al. 2020, de Lima et al. 2023]. Nesta seção, focaremos em trabalhos que analisaram a competência 4 do ENEM, que é o objeto de estudo deste trabalho.

No contexto da prova de redação do ENEM, os pesquisadores têm explorado diversas abordagens para a avaliação automática das competências consideradas no exame, utilizando as rubricas de correção predefinidas. Em um estudo conduzido por [Marinho et al. 2022b], foi apresentada uma pesquisa abrangendo diferentes estratégias e algoritmos de regressão para cada uma das cinco competências avaliadas no ENEM utilizando a base de dados do Essay-BR estendido [Marinho et al. 2022a]. Para cada competência, foram investigadas três abordagens: uma baseada em características, a utilização de representações multidimensionais (*embeddings*) das redações em conjunto com algoritmos tradicionais de AM, e o uso de redes neurais recorrentes.

O trabalho desenvolvido por [Lima et al. 2018] analisou 91 características, contemplando diversas dimensões de diversidade lexical, índices de legibilidade, contagem de conectivos e medidas de sobreposição de palavras entre sentenças e parágrafos. Foi realizada uma análise do desempenho de diversos algoritmos de classificação, utilizando uma base de dados composta por 8.584 redações similares às usadas no ENEM, com o objetivo de estimar notas relacionadas à coesão textual das redações.

No trabalho apresentado em [Oliveira et al. 2022], foi realizada uma investigação utilizando 151 atributos que englobam aspectos como o uso de conectivos, diversidade lexical, legibilidade, similaridade entre frases adjacentes, bem como várias características extraídas da ferramenta Coh-Metrix [Camelo et al. 2020]. Os autores realizaram uma comparação entre diversos algoritmos de regressão para estimar a nota da competência 4, utilizando a base de dados Essay-BR [Marinho et al. 2021]. Seguindo a mesma linha de pesquisa, o estudo desenvolvido em [Oliveira et al. 2023] explorou o uso de algoritmos de regressão por meio de uma abordagem baseada em atributos e o modelo neural de linguagem BERT para estimar as notas relacionadas à coesão textual em redações em português usando o corpus Essay-BR e em inglês usando a base de dados ASAP++ [Mathias and Bhattacharyya 2018]. Além disso, métodos de explicabilidade foram empregados com o intuito de fornecer interpretações das decisões tomadas pelos modelos para as notas estimadas.

Ao analisar os trabalhos anteriores, percebe-se que alguns deles utilizaram algoritmos de classificação, enquanto outros adotaram regressão. No entanto, nenhum estudo identificado realizou uma análise comparativa entre essas duas estratégias para modelar o problema de estimar automaticamente as notas da competência 4 do ENEM. Além disso, nota-se uma predominância no uso de abordagens baseadas em características que representam indicadores de coesão nas redações. No entanto, pesquisadores têm relatado

efeitos divergentes na utilização de recursos coesivos para a estimação de notas relacionadas à coesão textual em redações em inglês [McNamara et al. 2013, Crossley et al. 2016]. Diante disso, este trabalho busca contribuir nesse contexto, investigando a utilização de abordagens que usam diretamente o texto das redações, abrangendo desde a tradicional medida de TF-IDF até a utilização do modelo neural de linguagem BERT, considerando o problema de estimar as notas da competência 4 do ENEM como uma tarefa de classificação ou regressão.

3. Método

3.1. Corpus Essay-BR Estendido

O corpus Essay-BR [Marinho et al. 2021] tem sido utilizado em diversos trabalhos para avaliação automática de redações no contexto do ENEM [Marinho et al. 2022b, Oliveira et al. 2022, Oliveira et al. 2023]. Recentemente, foi apresentada uma nova versão desse corpus, chamado de Essay-BR estendido [Marinho et al. 2022a], com a inclusão de novas redações. O corpus Essay-BR estendido é composto originalmente de 6.579 redações dissertativas-argumentativas escritas seguindo o formato das redações presentes no ENEM. Essas redações abordam 151 temas, incluindo assuntos como política, *fake news*, direitos humanos, Covid-19, saúde, atividades culturais, entre outros.

O processo de coleta das redações foi realizado entre o período de dezembro de 2015 até agosto de 2021, utilizando um sistema para extrair as redações dos portais públicos do Vestibular UOL (Brasil Escola[2]) e Educação UOL[3]. Cada redação foi avaliada manualmente por examinadores profissionais, recebendo uma nota geral e notas individuais para as cinco competências, seguindo os mesmos critérios da prova de redação do ENEM [Marinho et al. 2022a]. As notas individuais das competências variam entre 0 e 200, em intervalos de 40 pontos. A nota final é obtida a partir da soma das notas individuais das competências.

Neste trabalho, foi utilizado o corpus Essay-BR estendido, focando exclusivamente nas notas atribuídas para a Competência 4 (C4) que trata da coesão textual das redações. Após uma análise manual do corpus, foram identificadas algumas redações duplicadas ou com seus textos vazios. Por isso, foram realizadas filtragens para a remoção dessas redações na base de dados original. Após esse processo, o corpus resultante adotado neste trabalho possui 6.563 redações. Na Tabela 1 são apresentadas algumas estatísticas descritivas gerais, agrupadas por nota. Para a geração das estatísticas, as redações foram pré-processadas utilizando a ferramenta spaCy[4].

Pode-se observar que a distribuição de redações por notas apresenta um alto grau de desequilíbrio. Existe uma grande quantidade de redações com nota 120 (2.455), 160 (1.821) e 200 (1.137), correspondendo a 37,41%, 27,75% e 17,32% do corpus, respectivamente. Por outro lado, as redações com nota 0 (206), 40 (65) e 80 (879) representam somente 3,14%, 0,99% e 13,39% do total de redações. Esse alto desbalanceamento, especialmente nas notas 0 e 40, impõe um desafio adicional às abordagens baseadas em algoritmos de AM investigadas.

[2]https://vestibular.brasilescola.uol.com.br/banco-de-redacoes
[3]https://educacao.uol.com.br/bancoderedacoes/
[4]https://spacy.io/

Tabela 1. Estatísticas descritivas do corpus Essay-BR estendido. Desvio padrão entre parênteses.

Nota	Total de Redações	Média de Frases	Média de Palavras
0	206	8,010 (3,699)	225,238 (67,889)
40	65	8,354 (3,743)	217,062 (87,982)
80	879	9,419 (4,681)	249,620 (85,338)
120	2.455	10,692 (4,691)	283,341 (79,265)
160	1.821	12,064 (3,777)	320,940 (72,184)
200	1.137	13,146 (3,475)	344,105 (67,945)
Geral	**6.563**	**11,220 (4,424)**	**297,304 (83,254)**

3.2. Abordagens Avaliadas

Neste trabalho, foram investigadas três abordagens baseadas diretamente no texto das redações para estimar automaticamente a coesão textual. A seguir, são apresentados brevemente mais detalhes das abordagens analisadas.

TF-IDF. Esta abordagem se baseia na representação de texto usando a tradicional medida de *Term Frequency – Inverse Document Frequency* (TF-IDF). Para isso, as redações passaram por um pré-processamento utilizando a ferramenta spaCy, que consistiu na divisão do texto em palavras e na remoção de símbolos de pontuação. Em seguida, as palavras remanescentes foram representadas por suas raízes (*lemmas*). Foi construído um vocabulário contendo as cinco mil palavras[5] (unigramas) mais frequentes a partir do corpus de treinamento. Cada redação é representada por um vetor de cinco mil posições, com o valor do TF-IDF de cada palavra do vocabulário. Por fim, essas redações foram usadas para treinar algoritmos de AM tradicionais para estimar as notas da coesão textual.

Representação Contextual Multidimensional. A ideia desta abordagem é codificar as redações por meio de uma representação contextual multidimensional (*contextual embeddings*). Para isso, as representações foram extraídas utilizando o modelo base pré-treinado do BERTimbau [Souza et al. 2020] com o auxílio da ferramenta *Sentence Transformers* [Reimers and Gurevych 2019]. Cada redação é representada por um vetor de setecentos e sessenta e oito (768) valores[6], que podem ser usados para descobrir padrões sintáticos e até semânticos presentes nos textos das redações. Em seguida, assim como na abordagem de TF-IDF, as redações codificadas foram utilizadas para treinar algoritmos de AM tradicionais para predizer as notas da C4 do ENEM.

BERT. Nesta abordagem, utilizou-se o modelo neural de linguagem BERT em conjunto com uma camada densa linear adicional para estimar a coesão textual nas redações. A implementação do modelo BERT utilizado neste trabalho foi baseada na biblioteca *transformers* e os modelos pré-treinados estão disponíveis publicamente em[7]. Para as etapas de treinamento de ajuste fino, aplicamos o escalonador com taxa de aprendizado sem aquecimento (*learning rate scheduler without warm-up*), seguido de decaimento linear da taxa de aprendizado e usando uma taxa de aprendizado inicial de $5 * 10^{-5}$. Foi utilizada a implementação *BERTimbau-base cased* [Souza et al. 2020] e uma versão

[5]Esse valor foi definido empiricamente.

[6]Tamanho de representação padrão definido na biblioteca.

[7]`https://github.com/huggingface/transformers`

compactada desse modelo, chamada de DistilBERT, que está disponível em[8]. O processo de ajuste fino foi realizado com três configurações diferentes para o número de épocas de treinamento: uma, cinco e dez. Em cada época, o algoritmo executa uma etapa de avaliação usando o conjunto de validação, sendo salvo apenas o modelo com o maior valor na medida de coeficiente linear de Kappa. Por fim, esse modelo é aplicado no conjunto de testes para predizer as notas da C4.

3.3. Seleção e Avaliação dos Modelos

Para responder à Pergunta de Pesquisa 1 (**PP1**), seguimos uma análise semelhante à apresentada em [Johan Berggren et al. 2019] e investigamos a utilização de algoritmos de classificação e regressão. Como os valores estimados pelos algoritmos de regressão são contínuos, aplicamos a estratégia de conversão proposta em [Marinho et al. 2022b] para mapear os valores preditos com as notas do ENEM (0, 40, 80, 120, 160 e 200).

Para responder à Pergunta de Pesquisa 2 (**PP2**), experimentos foram realizados considerando cada uma das abordagens (TF-IDF, representação contextual multidimensional e BERT) em ambas as tarefas: classificação e regressão. Em conjunto com as abordagens TF-IDF e representação contextual multidimensional, foram analisados o desempenho de diversos algoritmos disponíveis nas bibliotecas *scikit-learn*[9], o *eXtreme Gradient Boosting* (XGBoost)[10], CatBoost[11], e *Light Gradient Boosting Machine* (LGBM)[12]. Os hiperparâmetros dos algoritmos não foram ajustados, sendo mantidos os valores padrões definidos nas bibliotecas. Devido à limitação de espaço, na Seção 4, apresentamos apenas os dois algoritmos com melhor desempenho em cada abordagem, com base no coeficiente linear de Kappa. As tabelas com os resultados de todos os algoritmos analisados podem ser encontradas em[13].

A metodologia de validação cruzada estratificada com cinco subconjuntos foi adotada nos experimentos realizados para avaliar o desempenho dos algoritmos tanto para classificação quanto para regressão. Além disso, 10% do conjunto de treinamento em cada uma das cinco execuções foi separado como conjunto de validação. As seguintes medidas de avaliação foram utilizadas: coeficiente linear de *Kappa* [Cohen 1960], a raiz quadrada do erro médio, do inglês *Root Mean Squared Error* (RMSE), e o coeficiente de correlação de *Pearson*. Essas medidas foram escolhidas por serem amplamente utilizadas na literatura para avaliar sistemas de avaliação automática de redações [Lima et al. 2018, Oliveira et al. 2022, Marinho et al. 2022b, Oliveira et al. 2023]. Embora o coeficiente de Kappa seja mais utilizado em trabalhos de classificação, ele foi usado aqui para regressão porque os valores estimados pelos regressores são mapeados para as notas do ENEM. Da mesma forma, a medida de erro RMSE foi usada para avaliar os algoritmos de classificação por conta da natureza das classes estimadas representarem notas, ou seja, existe uma ordem de significância entre elas. Por exemplo, considere uma redação com nota real de 40 e dois algoritmos, A e B, que estimaram as notas como 160 e 80, respectivamente. Mesmo que ambos os algoritmos tenham estimado as notas de

[8] https://huggingface.co/adalbertojunior/distilbert-portuguese-cased
[9] https://scikit-learn.org/
[10] https://github.com/dmlc/xgboost/
[11] https://catboost.ai/
[12] https://github.com/Microsoft/LightGBM/
[13] https://github.com/hilariooliveira/coesao_textual_enem_sbie

forma incorreta, é possível quantificar, utilizando uma medida de erro, como a RMSE, que o algoritmo B (80) apresenta uma taxa de erro menor do que o algoritmo A (160).

Para mitigar o problema do desbalanceamento no número de exemplos de redações por nota no Essay-BR estendido, utilizamos a estratégia de ponderação das classes (*class weight*) [King and Zeng 2001] disponível na ferramenta *scikit-learn* para os algoritmos de classificação. Além disso, a mesma estratégia foi usada em conjunto com a função de perda de Entropia cruzada na abordagem do BERT para classificação.

Nossa interpretação do coeficiente linear de *Kappa* segue as diretrizes apresentadas em [Landis and Koch 1977]: **(i)** $kappa <= 0,2$ sugere um baixo nível de concordância; **(ii)** $0,21 <= kappa <= 0,4$ indica um nível razoável de concordância; **(iii)** $0,41 <= kappa <= 0,6$ representa um nível moderado de concordância; **(iv)** $0,61 <= kappa <= 0,8$ sugere um bom nível de concordância; e **(v)** $0,81 <= kappa <= 1,0$ indica um nível muito alto de concordância. Por fim, a interpretação da correlação de *Pearson* adotada é baseada no estudo de [Ratner 2009], que sugere que: **(i)** o valor 0 representa nenhuma relação linear; **(ii)** os valores $+1$ e -1 indicam uma relação perfeita positiva ou negativa, respectivamente; **(iii)** valores entre 0 e $0,3$ (0 e $-0,3$) indicam uma fraca relação positiva (ou negativa); **(iv)** valores entre $0,3$ e $0,7$ ($-0,3$ e $-0,7$) sugerem uma correlação moderadamente positiva ou negativa; e **(v)** valores entre $0,7$ e $1,0$ ($-0,7$ e $-1,0$) são considerados correlações fortes positivas ou negativas.

4. Resultados

A Tabela 2 apresenta os resultados dos experimentos realizados para responder às questões de pesquisa PP1 e PP2, considerando os algoritmos de classificação e regressão. Foram utilizadas as medidas de avaliação Kappa Linear (Kappa), RMSE e correlação de *Pearson* para cada uma das três abordagens avaliadas. Os melhores resultados em cada abordagem estão destacados em negrito e o melhor desempenho geral é indicado com o símbolo †. Na análise dos algoritmos de classificação, o modelo *BERTimbau-base* treinado por cinco épocas (*Base 5*) apresentou o melhor desempenho em todas as medidas de avaliação. Em relação à regressão, o modelo *BERTimbau-base* treinado por uma única época (*Base 1*) obteve os melhores resultados em termos do coeficiente de Kappa e da correlação de *Pearson*. O algoritmo *CatBoost*, utilizando representações multidimensionais (*Embeddings*), apresentou a menor taxa de erro na RMSE.

A abordagem usando a medida TF-IDF, embora simples, se mostrou ser um *baseline* com um bom desempenho, especialmente considerando o problema como uma tarefa de classificação e com base no coeficiente Kappa. Os algoritmos de máquina de vetores de suporte para classificação, do inglês *Support Vector Machine for Classification* (SVC), e Regressão Logística apresentaram valores competitivos com os obtidos por alguns dos modelos baseados no BERT. Considerando a abordagem usando as representações multidimensionais (*Embeddings*), os algoritmos *CatBoost* e LGBM para classificação, e CatBoost e máquina de vetores de suporte para regressão, do inglês *Support Vector Machine for Regression* (SVR), foram os que alcançaram melhor desempenho. Em geral, usando a representação multidimensional, foi possível obter menores taxas de erro na medida RMSE e valores de correlação de *Pearson* mais altos em comparação com a abordagem usando TF-IDF.

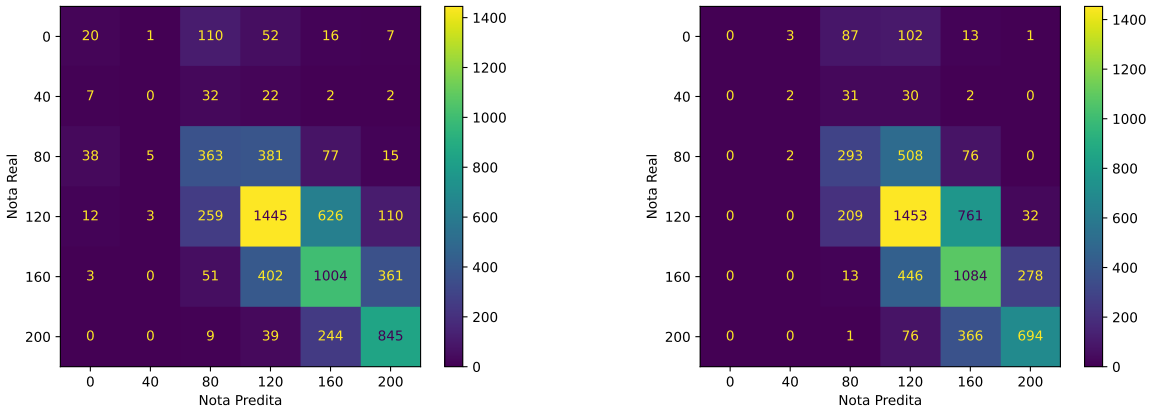
Na Figura 1, são apresentadas as matrizes de confusão do modelo com melhor

Tabela 2. Resultados dos experimentos considerando as tarefas de classificação e regressão.

Classificação				
Abordagens		**Kappa**	**RMSE**	**Pearson**
TF-IDF	SVC	0,350 (0,008)	39,279 (0,607)	0,593 (0,014)
	Reg. Logística	0,323 (0,014)	47,314 (1,254)	0,542 (0,024)
Embeddings	CatBoost	0,331 (0,015)	37,181 (0,615)	0,600 (0,016)
	LGBM	0,325 (0,011)	37,906 (0,503)	0,594 (0,013)
BERTimbau	Distil 1	0,299 (0,021)	39,396 (0,371)	0,606 (0,006)
	Distil 5	0,344 (0,025)	38,116 (0,824)	0,625 (0,012)
	Distil 10	0,330 (0,015)	38,347 (0,539)	0,609 (0,009)
	Base 1	0,335 (0,012)	37,546 (0,326)	0,647 (0,010)
	Base 5	**0,398† (0,012)**	**36,355 (2,068)**	**0,668† (0,012)**
	Base 10	0,381 (0,014)	36,467 (2,577)	0,660 (0,020)
Regressão				
TF-IDF	Bayesian Ridge	0,292 (0,008)	36,688 (0,429)	0,597 (0,013)
	CatBoost	0,279 (0,023)	37,286 (0,746)	0,580 (0,021)
Embeddings	SVR	0,315 (0,021)	35,137 (0,652)	0,638 (0,016)
	CatBoost	0,305 (0,016)	**34,986**† (0,578)	0,641 (0,015)
BERTimbau	Distil 1	0,315 (0,014)	36,071 (0,523)	0,616 (0,013)
	Distil 5	0,337 (0,022)	35,652 (0,734)	0,636 (0,019)
	Distil 10	0,326 (0,025)	35,964 (1,267)	0,630 (0,021)
	Base 1	**0,352 (0,020)**	35,037 (0,785)	**0,663 (0,017)**
	Base 5	0,337 (0,028)	35,353 (0,593)	0,653 (0,019)
	Base 10	0,347 (0,030)	35,398 (0,755)	0,660 (0,016)

desempenho (*BERTimbau-base*) considerando o coeficiente linear de Kappa para as tarefas de classificação (Figura 1a) e regressão (Figura 1b). Observa-se que ambos os modelos são mais precisos ao estimar as redações com notas 120 e 160, que são as mais frequentes no corpus Essay-BR estendido. Para as redações com nota 200, os modelos também apresentaram mais estimativas corretas do que incorretas, especialmente na tarefa de classificação. Em relação às redações com nota 80, ambos os modelos produziram predições corretas, porém com uma taxa de erro um pouco maior. É interessante notar que, mesmo nos casos de predições erradas, esses erros ocorrem geralmente em notas adjacentes, por exemplo, uma redação com nota real 120 sendo estimada como 80 ou 160. Essas divergências são observadas até mesmo entre avaliadores humanos, considerando a complexidade da tarefa de avaliação da coesão textual.

O maior desafio encontrado pelos modelos avaliados reside nas redações com notas 0 e 40, que são pouco representadas no corpus. Para essas redações, ambos os modelos tiveram dificuldades em realizar estimações corretas, e os erros ocorreram principalmente para as notas 80 e 120. Acreditamos que a quantidade limitada de redações com notas 0 e 40 não foi suficiente para que os modelos baseados no BERT pudessem generalizar adequadamente os padrões aprendidos. Por outro lado, observamos que alguns algoritmos tradicionais de AM treinados com as abordagens de TF-IDF e de representações multidimensionais, conseguiram apresentar estimativas corretas para as redações com notas



(a) BERTimbau-base 5 Classificação **(b)** BERTimbau-base 1 Regressão

Figura 1. Matrizes de confusão dos algoritmos com melhor desempenho usando classificação e regressão.

0 e 40, mas apresentaram muitos erros nas demais notas. A mitigação do problema do desbalanceamento entre as notas é um desafio que exigirá mais investigação no futuro.

5. Considerações Finais e Trabalhos Futuros

Neste estudo, realizamos uma análise comparativa de três abordagens de Aprendizado de Máquina, considerando as tarefas classificação e regressão, com o objetivo de estimar as notas relacionadas à coesão textual (competência 4) em redações no estilo do ENEM. As abordagens avaliadas utilizam diretamente os textos das redações e incluíram a tradicional medida TF-IDF, representações multidimensionais e modelos neurais baseados no *BERT*. Os resultados experimentais demonstraram que, de maneira geral, a modelagem do problema como uma tarefa de classificação levou a um desempenho superior em comparação com regressão no corpus Essay-BR estendido, especialmente considerando o coeficiente linear de Kappa. Entre as abordagens testadas, os modelos baseados no BERT obtiveram os melhores resultados, destacando-se o modelo *BERTimbau-base* treinado por cinco épocas usando classificação. Esses resultados sugerem que o uso de abordagens que consideram diretamente o texto das redações pode ser uma alternativa viável em relação aos métodos que se baseiam em características para a avaliação da coesão textual.

Apesar dos resultados encorajadores deste estudo, há algumas limitações a serem consideradas. Em primeiro lugar, a base de dados utilizada é relativamente pequena, o que restringe o potencial dos modelos neurais, como o BERT, que costumam obter melhores resultados com conjuntos de treinamento maiores. Além disso, o corpus Essay-BR estendido apresenta um desequilíbrio significativo nas notas da competência 4, especialmente nas notas 0 e 40, o que impactou diretamente o desempenho dos modelos, como evidenciado pelas matrizes de confusão e pelos valores relativamente baixos obtidos no coeficiente linear de Kappa. Por isso, para futuros estudos, planejamos: **(i)** explorar abordagens específicas para lidar com bases de dados textuais desbalanceadas no contexto educacional; **(ii)** investigar a aplicação de uma abordagem híbrida que combine características extraídas da redação com abordagens que utilizam diretamente o texto, como as investigadas neste trabalho; e **(iii)** expandir a análise para outras competências do ENEM.

Referências

Antunes, I. (2005). *Lutar com palavras: coesão e coerência*. Parábola Editorial, São Paulo.

Camelo, R., Justino, S., and Mello, R. (2020). Coh-metrix pt-br: Uma api web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, pages 179–186, Porto Alegre, RS, Brasil. SBC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Costa, L., Oliveira, E., and Júnior, A. C. (2020). Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412, Porto Alegre, RS, Brasil. SBC.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

de Lima, T. B., da Silva, I. L. A., Freitas, E. L. S. X., and Mello, R. F. (2023). Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Johan Berggren, S., Rama, T., and Øvrelid, L. (2019). Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy. Association for Computational Linguistics.

Ke, Z. and Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Klein, R. and Fontanive, N. (2009). Uma nova maneira de avaliar as competências escritoras na redação do enem. *Ensaio: Avaliação e Políticas Públicas em Educação*, 17(65):585–598.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Lima, F., Haendchen Filho, A., Prado, H., and Ferneda, E. (2018). Automatic evaluation of textual cohesion in essays. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Marinho, J., Anchiêta, R., and Moura, R. (2021). Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64, Online. Sociedade Brasileira de Computação.

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

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022b). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Mathias, S. and Bhattacharyya, P. (2018). Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1169–1173.

McNamara, D. S., Crossley, S. A., and Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, 45(2):499–515.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Oliveira, H., Miranda, P., Isotani, S., Santos, J., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Estimando coesão textual em redações no contexto do enem utilizando modelos de aprendizado de máquina. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 883–894. SBC.

Ramesh, D. and Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, 17(2):139–142.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.