Combinação de Modelos de Aprendizado de Máquina utilizando Teoria de Resposta ao Item para Avaliação de Coesão Textual em Redações no contexto do ENEM

Bruno Alexandre Rosa[1,3], **Hilário Oliveira**[2], **Rafael Ferreira Mello**[1,4], **Eduardo Araujo Oliveira**[3]

[1]Centro de Estudos e Sistemas Avançados do Recife (CESAR School)

[2]Instituto Federal do Espírito Santo (IFES) - Campus Serra

[3]University of Melbourne (UNIMELB)

[4]Universidade Federal Rural de Pernambuco (UFRPE)

babr@cesar.school, hilario.oliveira@ifes.edu.br,

rafael.mello@ufrpe.br, eduardo.oliveira@unimelb.edu.au

Abstract. *Essays are considered a valuable mechanism for evaluating learning outcomes in writing. Cohesion is a fundamental aspect of the text, as it helps establish meaningful relationships between its parts. This work aims to analyse the performance of cohesion score prediction using item response theory to ensemble scores generated by machine learning models. In this study, we selected the extended Essay-BR as a corpus comprising 6,563 essays in the National High School Exam (ENEM) style. We extracted 325 linguistic features and treated the problem as a machine learning regression task. The results indicate that the proposed approach outperforms conventional machine learning models and traditional ensemble methods in several evaluation metrics.*

Resumo. *A redação é considerada um mecanismo útil para a avaliação dos resultados da aprendizagem em escrita. A coesão é um aspecto fundamental do texto, visto que auxilia na relação de sentido entre suas diferentes partes. Este estudo teve como objetivo analisar o desempenho da previsão de pontuação de coesão usando a teoria de resposta ao item para ajustar as pontuações geradas pelos modelos de aprendizado de máquina. Para atingir esse objetivo, o corpus selecionado para o experimento é o Essay-BR estendido, que compreende 6.563 redações no estilo do Exame Nacional do Ensino Médio (ENEM). A pesquisa extraiu um total de 325 características linguísticas e tratou o problema como uma tarefa de regressão em aprendizado de máquina. Os resultados indicam que a abordagem proposta supera os modelos e os métodos de combinação convencionais de aprendizado de máquina em várias métricas de avaliação.*

1. Introdução

A escrita é uma habilidade fundamental para o desenvolvimento acadêmico, corporativo e social [Graham 2019]. A redação é um tipo de produção textual comumente praticada nas escolas para avaliar as habilidades de escrita dos alunos e deve ser organizada de acordo com uma estrutura bem definida [Graham 2019]. Dado o seu potencial de apoio

DOI: 10.5753/sbie.2024.242747

às avaliações educacionais, a redação é incluída no vestibular para a seleção de futuros alunos em muitas universidades [Klein and Fontanive 2009]. A produção de uma redação exige o correto emprego de mecanismos linguísticos essenciais para o desenvolvimento da escrita [Travaglia 2018]. Esse rigor se torna ainda mais crucial quando consideramos que a avaliação de uma redação é um processo subjetivo, envolvendo diversos critérios [Graham 2019].

A coesão textual é frequentemente um critério de avaliação em contextos educacionais em que a qualidade da escrita é analisada. Ela é um aspecto indispensável para proporcionar articulações e conexões gramaticais entre elementos do texto, como palavras, orações e frases [Halliday and Hasan 1976]. Um texto coeso apresenta ideias interligadas, o que permite ao leitor seguir o raciocínio do escritor de forma fluida [Koch 2010]. A coesão é alcançada por meio do uso adequado de mecanismos linguísticos necessários à construção textual [Halliday and Hasan 1976]. Esses mecanismos ajudam a guiar o leitor por meio do texto, tornando a mensagem clara e eficaz [Travaglia 2018]. Em um texto desprovido de coesão, as ideias fundamentais podem estar presentes, mas a falta de conexões compromete a clareza e a eficácia da comunicação, dificultando a compreensão e a interpretação do leitor [Halliday and Hasan 1976]. Portanto, espera-se que melhorar a coesão em redações beneficie a qualidade geral de outros aspectos da escrita [Crossley et al. 2019].

A aplicação de métodos automatizados para a correção de redações oferece várias vantagens, como a diminuição do tempo e dos custos envolvidos no processo de correção, além da redução de potenciais vieses e erros humanos [Marinho et al. 2022b]. Contudo, na avaliação de coesão em redações, os métodos ainda têm limitações. Por exemplo, eles frequentemente falham em capturar a referência e a sequência das relações semânticas, bem como a interconexão lógica entre as diferentes partes do texto. Além disso, a identificação de elementos de coesão, muitas vezes, exige um entendimento contextual que os métodos não conseguem replicar completamente. Também existe o desafio de interpretar corretamente elementos coesivos que podem ter múltiplos significados com base no contexto em que foram usados. Isso torna a tarefa de avaliação automática de coesão em redações um problema ainda em aberto [Crossley et al. 2019, Oliveira et al. 2022, Oliveira et al. 2023a].

Uma das possibilidades para lidar com essas limitações é criar um comitê de algoritmos que podem ser selecionados para realizar estimações em contextos específicos. Essa abordagem é reforçada pela prática recorrente de utilizar técnicas de avaliação do desempenho de algoritmos de Aprendizado de Máquina (AM) para selecionar os mais adequados aos variados desafios inerentes à sua aplicação. Tais técnicas de avaliação buscam compreender as vantagens e limitações dos algoritmos. Estudos recentes analisam uma perspectiva diferente de avaliação, na qual o desempenho de um algoritmo de AM é avaliado em nível de instância [Moraes et al. 2022, Uto et al. 2023]. Nesses casos específicos, o objetivo principal é identificar quais instâncias, em um conjunto de dados, foram mais ou menos difíceis para um determinado conjunto de algoritmos de AM [Moraes et al. 2022], permitindo uma análise mais precisa e detalhada da qualidade das previsões. Dessa forma, é possível selecionar conjuntos de teste mais adequados e otimizar o desempenho dos algoritmos de AM, o que pode ser útil para diferentes propósitos, tanto para o aprendizado quanto para a avaliação.

Nesse contexto, a Teoria de Resposta ao Item (TRI) é uma ferramenta estatística essencial para medir a probabilidade de um respondente responder corretamente a um item com base em sua habilidade latente [Embretson and Reise 2013]. A TRI, no processo de avaliação de algoritmos de AM, pode apoiar a modelagem das respostas de predição dos algoritmos em relação às características das instâncias [Moraes et al. 2022]. Isso possibilita identificar as instâncias que foram mais informativas para melhorar o desempenho de predição. Estudos recentes têm associado essa integração e produzido resultados promissores [Moraes et al. 2022, Uto et al. 2023]. Nesses estudos, os autores comparam o uso da TRI e fornecem uma análise teórica de seus parâmetros e habilidades nos modelos de AM. No entanto, esta pesquisa propõe ajustar a predição de saída dos algoritmos de AM usando TRI, a fim de produzir uma nova abordagem para predizer as pontuações de coesão em redações. Assim, em razão de um contexto desafiador e recém explorado no campo científico, este estudo busca responder à seguinte pergunta:

PERGUNTA DE PESQUISA (PP): *É possível estimar a coesão textual de redações de maneira mais precisa usando uma abordagem baseada em Teoria de Resposta ao Item para combinação de algoritmos de aprendizado de máquina?*

Para responder à questão de pesquisa, realizamos experimentos utilizando a base de dados estendida do Essay-BR [Marinho et al. 2022b], que inclui 6.563 redações acompanhadas de notas em diversas competências, avaliadas conforme os mesmos critérios aplicados no Exame Nacional do Ensino Médio (ENEM). Os experimentos consideraram diversos algoritmos de regressão, como *XGBoost*, *Support Vector Machine* e *Multilayer Perceptron*, bem como métodos de combinação, como *Voting* e *Stacked*. A abordagem proposta usando TRI obteve melhores resultados, apresentando um coeficiente *Kappa* linear de $0,421$ e ponderado quadrático de $0,581$, indicando um nível razoável de concordância com as notas dos examinadores. A correlação de *Pearson* foi moderada $(0,587)$ em relação às notas de coesão textual atribuídas por avaliadores humanos. Esses resultados apontam para a promissora capacidade da abordagem TRI de fornecer previsões mais precisas de coesão, o que pode ter implicações no desenvolvimento de sistemas de avaliação automática de redações mais eficazes.

2. Trabalhos Relacionados

Pesquisas recentes examinaram a relação entre características de escrita extraídas computacionalmente e avaliações humanas de coesão em redações escritas em português. Por exemplo, [Oliveira et al. 2022] realizou uma investigação usando 151 características que abrangem aspectos como o uso de conectores, diversidade lexical, legibilidade e similaridade entre frases adjacentes, juntamente com várias características extraídas da ferramenta Coh-Metrix. Os autores compararam vários algoritmos de regressão para estimar notas relacionadas à coesão usando o banco de dados Essay-BR [Marinho et al. 2022a]. Seguindo a mesma linha de pesquisa, o estudo desenvolvido em [Oliveira et al. 2023a] explorou algoritmos de regressão por meio de uma abordagem baseada em atributos e o modelo de linguagem do *Bidirectional Encoder Representations from Transformers* (BERT) para estimar as pontuações relacionadas à coesão textual em português e inglês. Além disso, métodos de explicabilidade foram empregados para fornecer interpretações das decisões tomadas pelos modelos para as pontuações estimadas.

A combinação entre TRI e AM tem como objetivo medir a habilidade e a difi-

culdade dos modelos de AM em aprender com conjuntos de dados, revelando resultados promissores em trabalhos recentes [Moraes et al. 2022, Uto et al. 2023]. O trabalho proposto por [Moraes et al. 2022] comparou o uso da TRI em diferentes regressores e forneceu uma análise teórica de seus parâmetros e habilidades em modelos AM. Este artigo modela o erro absoluto como uma função baseada na TRI, seguindo uma distribuição Gamma (Γ), para lidar com respostas positivas ilimitadas. A pesquisa se limitou a analisar os parâmetros do modelo proposto (Γ-TRI) aplicados às respostas de questões abertas em um exame de Estatística, sem explorar outros contextos de avaliação de redação ou extensões de abordagens da TRI.

Na pesquisa de [Uto et al. 2023], os autores propuseram um método que aplica a TRI para avaliar as características das pontuações atribuídas por algoritmos de Aprendizado de Máquina (AM) e as integram para gerar uma pontuação final estimada. O estudo demonstrou que o método proposto, usando o modelo TRI *Generalized Many-Facet Rasch Model* (GMFRM), obteve uma precisão média $(0,7562)$ mais alta na medida do Kappa quadrático ponderado, do inglês *Quadratic Weighted Kappa* (QWK), em comparação com os algoritmos individuais (QWK $0,7209$) e os métodos de integração convencionais (QWK $0,7395$). Esse resultado demonstra que a integração de vários modelos usando a TRI pode melhorar o desempenho, evidenciando seu potencial. No entanto, os autores se limitaram a testar a abordagem proposta em um conjunto restrito de modelos, sem análise de validação cruzada e em redações escritas em Inglês.

Assim, o uso da TRI para analisar e ajustar modelos AM permite mais pesquisas que podem melhorar a precisão e a confiabilidade das avaliações automatizadas de coesão em redações. O estudo de [Moraes et al. 2022] compara o uso da TRI em diferentes regressores e fornece uma análise teórica de seus parâmetros e habilidades em modelos de AM. Já a pesquisa de [Uto et al. 2023] concentrou-se na combinação dos resultados dos modelos de AM usando TRI. No entanto, diferentemente das propostas anteriores, esta pesquisa usa a TRI para ajustar a previsão de saída dos modelos AM a fim de propor uma nova abordagem para a previsão de pontuações de coesão em redações escritas em Português.

3. Método

O objetivo deste estudo é investigar a viabilidade de usar a TRI para integrar as pontuações de coesão de redações geradas por algoritmos de AM. Para isso, extraímos um conjunto extenso de características linguísticas propostas em trabalhos teóricos e empíricos anteriores. Os recursos foram extraídos de redações escritas em português, e utilizamos esses recursos para desenvolver dez modelos de AM. Em seguida, usamos a TRI para integrar a predição dos algoritmos de AM, a fim de produzir uma nova abordagem para a pontuação de coesão em redações.

3.1. Descrição do *Corpus*

A prova de redação é um dos componentes de avaliação de escrita mais importantes do Brasil [Klein and Fontanive 2009]. O ENEM[1] foi instituído em 1998 com o propósito de avaliar o desempenho escolar dos estudantes ao final da educação básica. Em 2009, o exame aprimorou sua metodologia e passou a ser utilizado como mecanismo de acesso

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

à educação superior. Desde 2020, os participantes podem optar por realizar o exame de forma impressa ou digital. Em 2023, foram 3.933.970[2] inscrições de candidatos para realizar o exame. Assim, milhões de futuros estudantes universitários realizam uma redação do tipo dissertativo-argumentativo sobre um tópico científico, cultural, político ou social [Klein and Fontanive 2009].

O *corpus* usado nesta pesquisa compreende redações do estilo do ENEM do *corpus* Essay-BR estendido coletadas por [Marinho et al. 2022b]. Desde versões anteriores, este *corpus* tem sido empregado em diversos estudos para a avaliação automática de redações no contexto do ENEM [Marinho et al. 2022b, Oliveira et al. 2022, Oliveira et al. 2023a, Oliveira et al. 2023b]. O *corpus* Essay-BR estendido é composto por 6.579 redações dissertativo-argumentativas, abordando 151 temas, como direitos humanos, questões políticas, saúde, atividades culturais, *fake news*, movimentos populares, Covid-19, entre outros.

A coleta das redações foi realizada no período de dezembro de 2015 até agosto de 2021. As redações foram extraídas dos portais públicos do Vestibular UOL[3] e Educação UOL[4]. Essas redações foram escritas por estudantes do ensino médio, respeitando o limite estipulado de um mínimo de 8 e um máximo de 30 linhas[5]. Seguindo os critérios de avaliação do ENEM, cada redação do *corpus* foi avaliada manualmente por especialistas, sendo atribuída uma nota geral e notas individuais para as cinco competências avaliadas, incluindo: **(i)** Escrita formal (aspectos lexicais e sintáticos); **(ii)** Entendimento do tema proposto; **(iii)** Capacidade de redação dissertativa-argumentativa; **(iv)** Coesão do texto dissertativo; e **(v)** Capacidade de propor uma intervenção para o problema descrito no ensaio. Em cada competência, os especialistas humanos atribuem uma nota que varia de 0 a 200 em intervalos iguais de 40. A nota final é obtida a partir da soma das notas individuais.

Neste trabalho, o foco se dará na Competência IV, que se relaciona exclusivamente com o aspecto de coesão textual da redação, um dos critérios considerados de maior dificuldade para os candidatos [Lima et al. 2018, Grama 2022, Oliveira et al. 2022, Oliveira et al. 2023a, Oliveira et al. 2023b]. Na fase inicial de manipulação do *corpus*, uma análise manual foi realizada. Durante essa análise, algumas redações duplicadas ou com seus textos vazios foram identificadas dentro do *corpus*. Após esse processo, o *corpus* adotado passou a contar com 6.563 redações. A Tabela 1 apresenta algumas estatísticas descritivas agrupadas pelas notas do aspecto de coesão. Dentro desse conjunto de dados, foram contabilizados o total de redações, a média de frases, a média de palavras e o desvio padrão (apresentado entre parênteses). O pré-processamento das redações para a geração dessas estatísticas foi realizado por meio da linguagem *Python*, utilizando a ferramenta spaCy[6].

Ao analisar a Tabela 1, constata-se que o tamanho médio das redações é de 297,30 palavras (excluindo pontuação e espaços em branco), o que corresponde à média de 11,22 frases por redação. A análise revela também que a distribuição de redações por notas

[2] https://www.gov.br/inep/pt-br/assuntos/noticias/enem/3-9-milhoes-estao-inscritos-no-enem-2023
[3] https://vestibular.brasilescola.uol.com.br/banco-de-redacoes
[4] https://educacao.uol.com.br/bancoderedacoes
[5] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem
[6] https://spacy.io/

Tabela 1. Estatísticas do *corpus Essay-BR* estendido.

Nota Coesão	Total de Redações	Média de Frases	Média de Palavras
0	206	8,01 (3,70)	225,24 (67,89)
40	65	8,35 (3,74)	217,06 (87,98)
80	879	9,42 (4,68)	249,62 (85,34)
120	2.455	10,69 (4,69)	283,34 (79,26)
160	1.821	12,06 (3,78)	320,94 (72,18)
200	1.137	13,15 (3,47)	344,10 (67,94)
Total	**6.563**	**11,22 (4,42)**	**297,30 (83,25)**

apresenta um desequilíbrio. As redações com nota 120 representam 37,41% do corpus, totalizando 2.455; as com nota 160 constituem 27,75%, somando 1.821; e as redações com nota 200 compõem 17,32% do *corpus*, com 1.137. Em contrapartida, os textos com notas 0, 40 e 80 representam apenas 3,14% (206), 0,99% (65) e 13,39% (879) do total de redações. Esse desequilíbrio, particularmente evidente nas notas 0 e 40, apresenta um desafio para as abordagens baseadas em algoritmos de AM.

3.2. Extração de Características

O propósito dessa etapa é converter o texto em um vetor de recursos numéricos que serão processados por modelos de AM, mantendo o sentido original do texto. Segundo [Shah and Patel 2016], a extração de características desempenha um papel crucial ao transformar os elementos textuais em um formato numérico, para poderem ser utilizados por modelos de AM. A extração de características é um procedimento que pode ser realizado por meio de uma variedade de métodos, incluindo técnicas matemáticas, classificação por categorias gramaticais e relação de dependência entre as palavras [Shah and Patel 2016].

As características usadas no presente estudo são descritas na Tabela 2. Nessa etapa, computamos um conjunto abrangente de recursos independentes que normalmente são adotados em pesquisas de mineração de texto educacional e análise de aprendizagem [Ferreira-Mello et al. 2019]. Alguns desses recursos foram usados para analisar a estrutura retórica de redações [Ferreira Mello et al. 2022], ponderar redações automaticamente [Klebanov and Madnani 2022] e avaliar coesão textual [Crossley et al. 2019, Oliveira et al. 2022, Oliveira et al. 2023a, Oliveira et al. 2023b].

Embora algumas características possam abranger aspectos semelhantes, implementações diferentes podem produzir resultados distintos. Neste sentido, utilizamos todos os recursos disponíveis para garantir uma análise abrangente. A Tabela 2 resume um total de 325 medidas de características linguísticas consideradas no presente estudo. A seleção se deve ao fato de que as pesquisas demonstraram as utilidades de cada característica para avaliar redações, em especial, coesão.

3.3. Processamento de modelos de aprendizado de máquina

Nessa etapa, o objetivo é selecionar, treinar, validar e testar diferentes modelos de AM para estimar as notas relacionadas à coesão em redações. Embora os escores de coesão nos conjuntos de dados adotados para o estudo sejam discretos (ver Tabela 1), decidiu-se explorar modelos de regressão em vez de classificadores. Tomou-se essa decisão com base

Tabela 2. Quantidade de Características por Grupo.

Grupo	Referência	#Característica	%
Coh-Metrix	[Camelo et al. 2020]	87	26,77%
LIWC	[Balage Filho et al. 2013]	64	19,69%
Informações Morfossintáticas de Palavras	[Leal et al. 2021]	39	12,00%
Conectivos	[Grama 2022]	33	10,15%
Coesão Semântica	[Leal et al. 2021]	21	6,46%
Diversidade Lexical	[Leal et al. 2021]	18	5,54%
Complexidade Sintática	[Leal et al. 2021]	17	5,23%
Medidas Descritivas	[Leal et al. 2021]	11	3,38%
Coesão Referencial	[Leal et al. 2021]	9	2,77%
Simplicidade Textual	[Leal et al. 2021]	8	2,46%
Coesão Sequencial	[Leal et al. 2021]	7	2,15%
Legibilidade	[Palma and Atkinson 2018]	7	2,15%
Densidade de Padrões Sintáticos	[Leal et al. 2021]	4	1,23%
Total		**325**	**100,00**

na literatura que trata da análise de coesão de texto e até sistemas de avaliação automática de redações, como problemas de regressão [Oliveira et al. 2022, Oliveira et al. 2023a].

Para a execução dos experimentos de previsão da nota de coesão, selecionamos algoritmos de regressão baseados na literatura [Ferreira-Mello et al. 2019, Oliveira et al. 2023a] que usam diferentes abordagens. Os algoritmos de AM aplicados nos experimentos foram: *Bayesian Ridge*[7], *CatBoost Regressor*[8], *Decision Tree Regressor*[7], *Extra Trees Regressor*[7], *LGBM Regressor*[9], *Linear Regression*[7], *MLP Regressor*[7], *Random Forest*[7], *SVR*[7] e *XGB Regressor*[10]. Os hiperparâmetros dos algoritmos permaneceram inalterados, mantendo-se os valores padrões estabelecidos em suas respectivas bibliotecas.

Além dos algoritmos individuais, estratégias para combinar as previsões individuais foram adotadas para melhorar o desempenho preditivo de um único modelo [Sagi and Rokach 2018]. No estudo atual, avaliamos os seguintes algoritmos de conjunto (*ensemble*) mais populares: *Voting Regressor*, que calcula a média das previsões individuais dos regressores para formar uma previsão final, e a abordagem do conjunto *Stacked Regressor*, que combina regressores com um regressor final. Usamos todos os modelos individuais como entrada na abordagem de conjunto. Para os resultados na abordagem de *Stacked Regressor*, usamos os modelos *Linear Regression* e *SVR*. Utilizamos os parâmetros de ajuste padrão definidos nas bibliotecas para todos os algoritmos e conjuntos.

3.4. Processamento da abordagem baseada na teoria de resposta ao item

Nesta etapa, o objetivo é combinar e ajustar as notas de coesão previstas pelos modelos de AM, gerando uma previsão final. A TRI fornece uma estrutura para modelar e entender a interação entre indivíduos e itens de teste [Embretson and Reise 2013]. A TRI pode ser

[7] https://scikit-learn.org/
[8] https://catboost.ai/
[9] https://github.com/Microsoft/LightGBM/
[10] https://github.com/dmlc/xgboost/

abordada no contexto de AM, considerando que as instâncias do conjunto de dados se equiparam aos itens de teste, ao passo que os algoritmos se assemelham aos respondentes [Moraes et al. 2022, Uto et al. 2023].

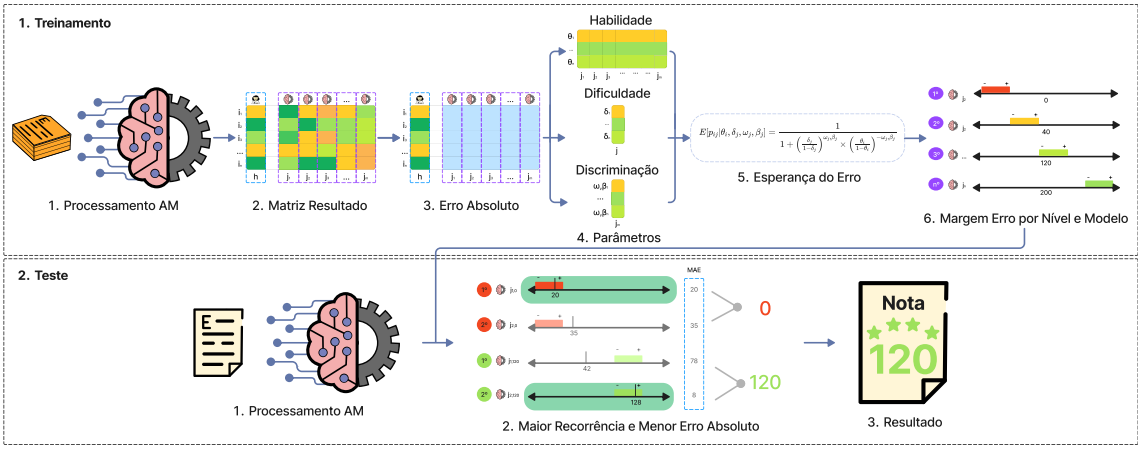


Figura 1. Visão geral das etapas da abordagem proposta.

A Figura 1 ilustra as etapas principais da abordagem proposta. Para calcular os parâmetros e a esperança do erro baseado na TRI, foi adotado o modelo BIRT-GD[11] proposto por [Ferreira-Junior et al. 2023]. Os parâmetros da TRI considerados no estudo foram: habilidade, dificuldade e discriminação. A habilidade refere-se à capacidade de um modelo de AM em prever corretamente as notas de coesão. A dificuldade é um parâmetro que indica quão complexo é para os modelos aprenderem os dados, refletindo a variabilidade das notas previstas em relação às notas reais. A discriminação, por sua vez, mede a precisão com que um modelo diferencia entre as faixas de notas de coesão, sendo crucial para identificar quais modelos possuem maior sensibilidade às variações das notas.

O modelo BIRT-GD é uma implementação de β^4-IRT que utiliza o método de otimização por gradiente descendente. Para o contexto deste estudo, o modelo BIRT-GD recebe um conjunto de dados, sendo X a matriz contendo o erro absoluto (p_{ij}) gerado entre a nota do avaliador humano e a nota prevista por cada modelo de AM. O índice (i) faz referência ao item, que corresponde à faixa de nota de coesão. A coluna de índice (j) representa o respondente, que se refere a cada modelo de AM utilizado. Para cada modelo de AM e faixa de nota de coesão, a esperança do erro fornece um valor esperado de erro, que é utilizado para definir os limites mínimos e máximos das previsões. Esses intervalos de confiança são essenciais para selecionar a previsão final da nota de coesão, pois permitem identificar a previsão mais frequente e confiável entre os modelos. Para garantir a consistência e a integridade dos resultados, a divisão adotada para treinamento, validação e teste é a mesma separação utilizada desde o *corpus* inicial.

Em linhas gerais, a abordagem proposta, denominada TRI-Multiregressor, pode ser resumida nas seguintes etapas:

1. Processamento do Treinamento:
    1.1. Os modelos de AM considerados para compor a abordagem são treinados individualmente utilizando o conjunto de treinamento;

[11] `https://pypi.org/project/birt-gd/`

1.2. Após o treinamento, os modelos são aplicados nos dados de validação e as notas estimadas por cada modelo em cada redação do conjunto de validação são representadas em uma matriz;
1.3. O erro absoluto é calculado para cada redação e modelo de AM;
1.4. O conjunto de resultados do erro absoluto para cada modelo em cada faixa de nota (0-200) é dado como entrada para o modelo BIRT-GD para o cálculo dos parâmetros da TRI;
1.4.1. Os valores de discriminação e dificuldade são computados para cada faixa de nota de coesão;
1.4.2. O valor da habilidade é calculado para cada modelo e faixa de nota de coesão.
1.5. A esperança do erro é estimada pelo BIRT-GD para cada modelo e faixa de nota de coesão utilizando os parâmetros da TRI;
1.6. Para cada modelo e nota, são gerados os limites mínimos e máximos por meio da esperança do erro, gerando um intervalo de confiança.
2. Processamento do Teste:
2.1. Dada uma redação, os modelos de AM são aplicados individualmente para estimar a sua nota de coesão;
2.2. A nota de coesão final da redação é selecionada com base na maior frequência de previsões dentro do intervalo de confiança do modelo AM. Em caso de empate na frequência, a nota é selecionada com base no menor erro absoluto;
2.3. Por fim, a nota final selecionada é atribuída à redação.

3.5. Avaliação dos resultados

Nesta etapa, foram avaliados o desempenho dos resultados de diferentes algoritmos tradicionais de AM, de combinação e a abordagem proposta baseada na TRI. Para garantir a consistência e a integridade dos resultados da análise, adotamos a metodologia de validação cruzada estratificada com dez subconjuntos, 10-*fold Stratified Cross Validation*. As seguintes medidas de avaliação foram utilizadas: coeficiente linear de *Kappa* [Cohen 1960], coeficiente *Quadratic Weighted Kappa* (QWK) [Cohen 1968], coeficiente de correlação de *Pearson* [Pearson 1896] e matriz de confusão.

Selecionamos essas medidas de avaliação, pois elas são comumente usadas 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. 2023a, Oliveira et al. 2023b, de Lima et al. 2023]. Embora o coeficiente de *Kappa*, o QWK e a matriz de confusão sejam tradicionalmente utilizados em problemas de classificação, neste contexto, eles também se aplicam aos problemas de regressão, pois os valores previstos pelos regressores foram categorizados em faixas de notas predefinidas nas orientações do ENEM. As métricas de *Kappa* e QWK foram calculadas utilizando a biblioteca Scikit-learn[12], e a correlação de *Pearson* foi calculada usando a biblioteca SciPy[13].

Dado que os valores previstos pelos algoritmos de regressão são contínuos, foi aplicada a estratégia de conversão sugerida em [Marinho et al. 2022b] para correlacionar os valores preditos com as notas de coesão do ENEM. O estudo seguiu o seguinte padrão:

[12] https://scikit-learn.org/
[13] https://scipy.org/

(i) valor menor que 20 indica nota 0; **(ii)** valor maior ou igual a 20 e menor que 60 indica nota 40; **(iii)** valor maior ou igual a 60 e menor que 100 indica nota 80; **(iv)** valor maior ou igual a 60 e menor que 100 indica nota 120; **(v)** valor maior ou igual a 60 e menor que 100 indica nota 160; e **(vi)** valor maior que 180 indica nota 200.

Os coeficientes *Kappa* são interpretados seguindo as diretrizes apresentadas por [Landis and Koch 1977], que apontam: **(i)** valores menores que 0,20 sugerem um nível baixo de concordância; **(ii)** valores entre 0,21 e 0,4 indicam um nível razoável de concordância; **(iii)** valores entre 0,41 e 0,6 representam um nível moderado de concordância; **(iv)** valores entre 0,61 e 0,8 sugerem um bom nível de concordância; e **(v)** valores entre 0,81 e 1,0 indicam um nível muito alto de concordância.

A correlação de *Pearson* pode ser avaliada segundo as seguintes diretivas: **(i)** um valor de 0 indica a inexistência de relação linear; **(ii)** os valores de +1 e -1 correspondem a uma correlação perfeita, seja ela positiva ou negativa; **(iii)** valores que variam entre 0 e 0,3 (ou entre 0 e -0,3) sugerem uma correlação fraca, positiva ou negativa; **(iv)** valores situados entre 0,3 e 0,7 (ou entre -0,3 e -0,7) indicam uma correlação moderada, positiva ou negativa; e **(v)** valores que se situam entre 0,7 e 1,0 (ou entre -0,7 e -1,0) representam uma forte correlação, seja ela positiva ou negativa.

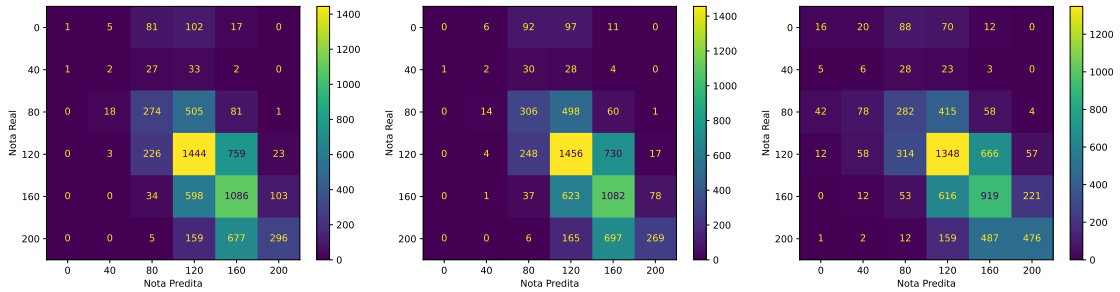
4. Resultados

Na Tabela 3 são apresentados os resultados dos experimentos realizados para responder à pergunta de pesquisa PP. Os resultados foram obtidos por meio da abordagem tradicional usando os algoritmos de AM individualmente, os métodos de combinação clássicos e a TRI-Multiregressor proposta. Foram utilizadas as métricas de avaliação *Kappa* Linear, *Kappa* Ponderado Quadrático (QWK) e a correlação de *Pearson*. Os melhores resultados em cada abordagem foram destacados em negrito.

Tabela 3. Resultados dos experimentos.

	Abordagem	Kappa	QWK	Pearson
Tradicional	Bayesian Ridge	0,230 (0,018)	0,508 (0,023)	0,554 (0,024)
	Cat Boost Regressor	0,247 (0,016)	**0,533 (0,028)**	**0,572 (0,029)**
	Decision Tree Regressor	0,113 (0,020)	0,352 (0,029)	0,352 (0,029)
	Extra Trees Regressor	0,218 (0,021)	0,492 (0,030)	0,551 (0,029)
	LGBM Regressor	0,236 (0,018)	0,525 (0,022)	0,560 (0,024)
	Linear Regression	0,248 (0,015)	0,525 (0,019)	0,559 (0,019)
	MLP Regressor	0,173 (0,030)	0,490 (0,036)	0,502 (0,041)
	Random Forest	0,217 (0,019)	0,489 (0,024)	0,545 (0,022)
	SVR	**0,250 (0,019)**	0,528 (0,022)	0,569 (0,023)
	XGB Regressor	0,209 (0,021)	0,506 (0,022)	0,527 (0,021)
Combinação	Stacked Linear Regression	**0,253 (0,018)**	**0,538 (0,019)**	**0,580 (0,019)**
	Stacked SVR	0,233 (0,021)	0,518 (0,021)	0,554 (0,020)
	Voting Regressor	0,246 (0,012)	0,524 (0,018)	0,570 (0,019)
Proposta	TRI-Multiregressor	**0,421 (0,019)**	**0,581 (0,021)**	**0,587 (0,020)**

Analisando os resultados obtidos na medida de concordância do *Kappa*, vê-se que a abordagem proposta do TRI-Multiregressor apresentou o melhor desempenho, com um



(a) SVR (b) Stacked LR (c) TRI-Multiregressor

Figura 2. Matrizes de confusão dos algoritmos com melhor desempenho.

valor de $0,421$. Esse valor é superior em relação a qualquer algoritmo tradicional, com o SVR como o mais próximo, mas com desempenho bem inferior $(0,250)$, e também à combinação usando *Stacking* com regressão linear, que obteve um desempenho de $0,253$. Essa diferença indica que o TRI-Multiregressor apresenta uma concordância mais forte entre as predições geradas e as notas reais de coesão.

No que se refere à medida QWK, a abordagem por meio da TRI se destaca com um valor de $0,581$, enquanto os demais algoritmos não ultrapassam $0,538$. O QWK é uma métrica amplamente usada em trabalhos de avaliação automática de redações, pois não apenas leva em conta a concordância, como também pondera a importância dos erros. Isso torna a métrica mais robusta e indica que a TRI é eficaz na avaliação da qualidade da nota de coesão.

A correlação de *Pearson* avalia a correlação linear entre as predições e os valores verdadeiros. A abordagem proposta atingiu um resultado de $0,587$. Em comparação, a maioria dos outros modelos apresenta correlações moderadas, com valores que não ultrapassam a marca de $0,580$. Isso implica um nível moderado entre as predições do modelo e os valores reais, o que é um sinal de que a abordagem tende a ser mais precisa.

Na Figura 2, foram apresentadas as matrizes de confusão considerando o melhor modelo da abordagem tradicional (Figura 2a), da combinação (Figura 2b) e da TRI (Figura 2c), considerando o coeficiente linear de *Kappa*. As matrizes de confusão demonstram o número de previsões corretas ou erradas por nota. É possível observar que os modelos apresentaram maior precisão ao estimar as redações com notas 120 e 160, que foram as mais comuns no *corpus* Essay-BR estendido. As demais notas (0, 40 e 200), embora tenham fornecido predições corretas, apresentaram uma proporção de erro maior.

Por fim, é importante notar que, mesmo nos casos de predições incorretas, esses erros ocorrem geralmente em notas adjacentes. Por exemplo, uma redação com nota real 160 sendo estimada como 120 ou 200. Essas discordâncias são comuns, sendo observadas até mesmo entre examinadores humanos, considerando a dificuldade e subjetividade da tarefa de avaliação da coesão textual. Os modelos tradicionais de AM encontraram dificuldades nas redações com notas 0, 40 e 200, que foram pouco representativas no *corpus*. No entanto, ao se empregar a abordagem por meio da TRI, as previsões de saída foram ajustadas nos limites da esperança do erro, o que permitiu um aumento da precisão dos resultados em níveis extremos de coesão. A melhora nos resultados reforça a importância de explorar novas abordagens em experimentos futuros.

5. Considerações Finais

Tarefas de produção textual, em especial a escrita de redações, estão entre as ferramentas mais populares utilizadas por professores para avaliar o aprendizado de alunos. Neste trabalho, realizamos uma análise comparativa de abordagens tradicionais e combinadas de algoritmos de aprendizado de máquina com o objetivo de estimar as notas relacionadas à coesão em redações no contexto do ENEM. Os experimentos foram realizados usando o *corpus* Essay-BR estendido [Marinho et al. 2022b]. Foram extraídas 325 características referentes à coesão textual baseadas em abordagens da literatura, como Coh-Metrix, LIWC, entre outras.

Diferentes tipos de modelos de AM foram treinados, abordando o problema como uma tarefa de regressão. Alguns dos modelos utilizados foram Bayesian Ridge, Cat Boost, Regressão Linear, entre outros. A abordagem proposta combinou as predições dos modelos de AM por meio da TRI. Essa abordagem demonstra ser particularmente adequada para capturar a probabilidade da esperança do erro por modelo de AM e faixa de nota, levando em consideração parâmetros como a dificuldade e a discriminação. Os resultados experimentais obtidos indicaram que a abordagem proposta obteve o melhor desempenho geral com um *Kappa* de $0,421$, um QWK de $0,581$ e uma correlação de *Pearson* moderada de $0,587$.

Ainda que este estudo apresente resultados encorajadores, é necessário considerar algumas limitações. Primeiro, embora tenhamos explorado uma ampla gama de algoritmos tradicionais de AM, não incluímos modelos de linguagem de aprendizado profundo, nem métodos relacionados ao ajuste de parâmetros e ao balanceamento de dados. Segundo, o *corpus* utilizado possui características semelhantes quanto ao tipo de texto em português (dissertativo-argumentativo), o que indica a necessidade de avaliar a proposta em termos de generalização. Por fim, este estudo não teve a intenção de analisar a aplicação prática do método proposto, que incluiria o desenvolvimento de uma ferramenta de análise de aprendizagem e avaliar a satisfação de professores e alunos.

Para futuros estudos, planejamos: **(i)** incorporar modelos de regressão baseados no modelo *Bidirectional Encoder Representations from Transformer* (BERT); **(ii)** investigar uma abordagem híbrida por meio da TRI que combine modelos que extraem características linguísticas e abordagens que utilizam diretamente o texto; **(iii)** aplicar a abordagem a um *corpus* contendo outros tipos de produções textuais, como textos narrativos; e **(iv)** expandir a análise para outras competências comumente consideradas em processos de avaliação de redações.

Referências

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Camelo, R., Justino, S., and de Mello, R. F. L. (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. SBC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Crossley, S. A., Kyle, K., and Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1):14–27.

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.

Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

Ferreira-Junior, M., Reinaldo, J. T., Filho, T. M. S., Neto, E. A. L., and Prudencio, R. B. (2023). β^4-irt: A new β^3-irt with enhanced discrimination estimation. *arXiv e-prints*.

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

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., and Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 404–414.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, 43(1):277–303.

Grama, D. F. (2022). *Elementos coesivos do português brasileiro em córpus de redações nos moldes do Enem: um estudo para a elaboração da CoTex*. PhD thesis.

Halliday, M. A. and Hasan, R. (1976). Cohesion in english. *Longman*.

Klebanov, B. B. and Madnani, N. (2022). *Automated Essay Scoring*. Springer Nature.

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

Koch, I. G. V. (2010). *A coesão textual*, volume 22. São Paulo Contexto.

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

Leal, S., Duran, M., Scarton, C., Hartmann, N., and Aluísio, S. (2021). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese.

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. (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.

Moraes, J. V., Reinaldo, J. T., Ferreira-Junior, M., Silva Filho, T., and Prudêncio, R. B. (2022). Evaluating regression algorithms at the instance level using item response theory. *Knowledge-Based Systems*, 240:108076.

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

Oliveira, H., Mello, R. F., Miranda, P., Alexandre, B., Cordeiro, T., Bittencourt, I. I., and Isotani, S. (2023b). Classificaçao ou regressão? avaliando coesão textual em redações no contexto do enem. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1226–1237. SBC.

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.

Palma, D. and Atkinson, J. (2018). Coherence-based automatic essay assessment. *IEEE Intelligent Systems*, 33(5):26–36.

Pearson, K. (1896). Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London Series A*, 187:253–318.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Shah, F. P. and Patel, V. (2016). A review on feature selection and feature extraction for text classification. In *International conference on wireless communications, signal processing and networking (WiSPNET)*, pages 2264–2268. IEEE.

Travaglia, L. C. (2018). Tipologia textual e ensino de língua. *Domínios de Lingu@gem*, 12(3):1336–1400.

Uto, M., Aomi, I., Tsutsumi, E., and Ueno, M. (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies*.