Em Direção à Pontuação Multimodal de Redações Narrativas

Hyan H. N. Batista[1], **Gabriel Augusto Barbosa**[1]

[1]Departamento de computação – Universidade Federal Rural do Pernambuco (UFRPE)
Recife – PE – Brasil

hyan.batista@ufrpe.br

gabriel.augusto@ufrpe.br

Abstract. *This article presents an attempt to use a multimodal model to evaluate elementary school students' essays, integrating textual and visual information through BERT and ViT models. The goal was to assign grades to the essays in different dimensions, such as Formal Register, Thematic Coherence, Text Typology, and Cohesion. Although the expectation was that the multimodal approach would improve classification, the results did not meet the expected performance. Precision, recall, and F1 scores, particularly macro metrics, indicate the model's difficulties in handling unbalanced classes and capturing the complexity of the tasks evaluated. The discussion points to potential improvements, including class balancing and more effective integration of textual and visual representations.*

Resumo. *Este artigo apresenta uma tentativa de utilização de um modelo multimodal para avaliar redações de estudantes do ensino fundamental, integrando informações textuais e visuais através dos modelos BERT e ViT. O objetivo foi atribuir notas às redações em diferentes dimensões, como Registro Formal, Coerência Temática, Tipologia Textual e Coesão. Embora a expectativa fosse que a abordagem multimodal trouxesse melhorias na classificação, os resultados não atingiram o desempenho esperado. As métricas de precisão, recall e F1, especialmente as macro, indicam dificuldades do modelo em lidar com classes desbalanceadas e em captar a complexidade das tarefas avaliadas. A discussão aponta para possíveis melhorias, incluindo o balanceamento das classes e uma integração mais eficaz das representações textuais e visuais.*

1. Introdução

Nos últimos anos, a área de modelos multimodais, que integra informações provenientes de diferentes modalidades de dados, como texto e imagem, tem se mostrado promissora em uma ampla gama de aplicações. Esses modelos combinam representações de múltiplas fontes para capturar nuances e complementariedades que um único tipo de dado pode não ser capaz de fornecer isoladamente. Um exemplo relevante dessa abordagem é o uso de redes neurais baseadas no BERT (Bidirectional Encoder Representations from Transformers) para processamento de texto e no ViT (Vision Transformer) para análise de imagens, ambos modelos de estado da arte em suas respectivas áreas.

No contexto da avaliação automática de redações de estudantes do ensino fundamental, a aplicação de modelos multimodais pode proporcionar melhorias substanciais na precisão das classificações. Além de avaliar o conteúdo textual das redações, que envolve

a análise de coerência, gramática, e estrutura argumentativa, a modalidade de imagem pode capturar informações adicionais, como a legibilidade da escrita, a organização visual do texto, e até mesmo traços comportamentais e estilísticos dos estudantes ao escrever. Integrar essas duas dimensões pode gerar uma compreensão mais rica do desempenho dos alunos, resultando em uma avaliação mais justa e detalhada.

A abordagem tradicional, que foca exclusivamente no texto, tende a negligenciar esses aspectos visuais. Um modelo multimodal, por outro lado, é capaz de preencher essa lacuna, trazendo possíveis avanços na capacidade de avaliar não apenas o conteúdo textual, mas também características visuais que podem refletir fatores relevantes para a qualidade da redação. Nesta tentativa, implementamos um modelo multimodal que combina BERT e ViT com o objetivo de atribuir notas de 1 a 5 para redações de estudantes do ensino fundamental, investigando como essa integração impacta a precisão e a robustez da classificação.

2. Trabalhos Relacionados

A presente seção apresenta uma análise dos trabalhos realizados na área. Inicialmente, são discutidas as pesquisas relacionadas à fusão de *features* multimodais no contexto do processamento de documentos digitais. Posteriormente, são abordados os estudos sobre AES (*Automated Essay Scoring*) e os métodos aplicados, com ênfase nas abordagens voltadas para a língua portuguesa.

O uso de técnicas de fusão de *features* multimodais, principalmente no contexto específico de construção de *Vision Language Models*, não é algo novo na literatura. [Jain and Wigington 2019], por exemplo, explorou o uso de um modelo multimodal *early fusion* para classificar automaticamente documentos de imagem. A abordagem demonstrou desempenho superior à métodos que faziam uso de *Transfer Learning*. [Su et al. 2023], por sua vez, desenvolveram um sistema de AES para avaliar redações manualmente escritas de estudantes chineses usando um modelo que fundia *features* multimodais de linguagem e imagem.

No que se refere ao estudo de sistemas AES para Língua Portuguesa, existem uma série de estudos abordando o uso de métodos de *Machine Learning* na detecção automática de elementos essenciais para a avaliação de textos narrativos. [Batista et al. 2022], por exemplo, aplicou métodos de *Machine Learning* para identificar computacionalmente a presença de clímax em textos narrativos de estudantes do ensino fundamental. Na mesma linha, utilizou técnicas de *Generative AI* e outros algoritmos de inteligência computacional para detectar categorias e elementos narrativos.

Os trabalhos citados acima focam em elementos específicos de um texto que são cruciais no processo de *assesment* dos estudantes. Entretanto, existem trabalhos que apresentam sistemas de AES completos. Se baseando no conjunto de dados introduzido por [Mello et al. 2024], por exemplo, [da Silva Filho et al. 2023] combinou diferentes algoritmos de *Machine Learning* com um conjunto de *features* linguísticas construídas manualmente. O sistema de AES alcançou níveis de concordância semelhante ao de dois anotadores humanos. Usando esse mesmo conjunto de dados, [Ribeiro et al. 2024] explorou o uso de modelos pré-treinados baseados na arquitetura *Transformer*.

Na literatura, portanto, há muitos trabalhos explorando o uso de métodos de *Deep Learning*, *Machine Learning* e *Multimodal Feature Fusion* para processar documentos de

imagem e redações. Entretanto, há uma escassez de trabalhos explorando a aplicação de métodos de *Deep Learning* empregando fusão de *features* multimodais para o desenvolvimento de sistemas de AES para o *assesment* de textos de alunos do ensino fundamental em Língua Portuguesa.

3. Metodologia

3.1. Dados

O dataset utilizado é uma extensão do dataset do KAGGLE, um dataset de avaliação de redações criado a partir de redações do ensino fundamental do 5º ao 9º ano. Este dataset é composto por imagens das redações, textos anotados manualmente, e avaliações destes textos. Cada texto é avaliado em quarto competências, criadas a partir de um guia de correção feito pela UFAL. Cada uma das competências avalia um aspecto único do texto. São consideradas competências:

- Registro Formal: Avalia o nível da escrita do aluno em relação a norma culta da língua portuguesa, aspectos como erros ortográficos e de pontuações são avaliados nesse quesito.
- Coerência Temática: Avalia do nível de adesão ao tema proposto para a escrita da redação, o aluno deve atender ao tema proposto acrescentando informações, mas sem fugir ao tema.
- Tipologia Textual: O tipo textual proposto no dataset é uma breve narrativa, o aluno deve criar uma história que tenha estrutura narrativa, com introdução, desenvolvimento, clímax e um desfecho. Este aspecto avalia a completude dos elementos narrativos na redação.
- Coesão: Avalia o nível de coesão do texto, entre cada sentença, o aluno deve manter uma sequencia lógica de parágrafos.

Como mostra a figura 1, a distribuição dos dados de cada competência é bem diferente. Por exemplo, a competência de Tipologia Textual tem seu pico de frequência no nível 4, enquanto as competências de Coesão, Coerência Temárica e Registro Formal têm mais exemplos no nível 3. Já o nível 5 quase não possui exemplares de redações, mostrando que falta, nessa base, redações de nota máxima.

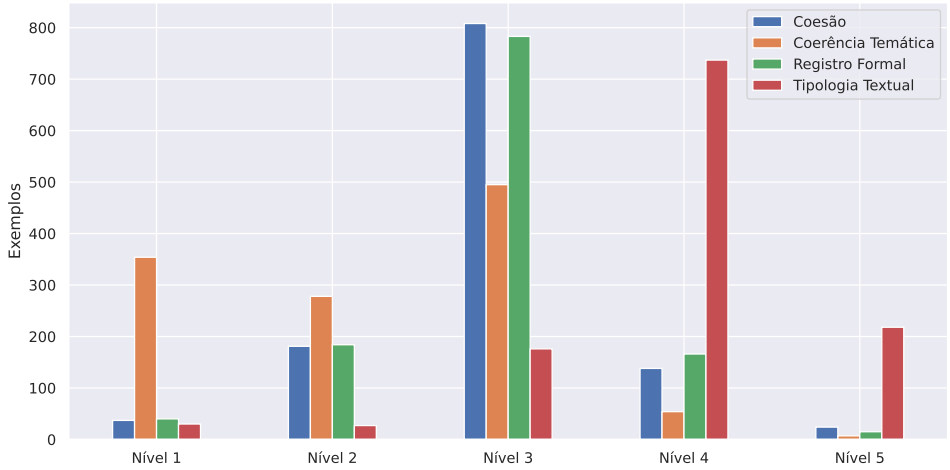

Figure 1. Redações Agrupadas por Nível

	Encoder	Acc	Macro Prec	Weighted Prec	Macro Recall	Weighted Recall	Macro F_1	Weighted F_1	Kappa
SVC	**TF-IDF**	0.670	0.169	0.461	0.246	0.670	0.246	0.546	-0.012
SVC	**BERT**	0.650	0.446	0.646	0.444	0.650	0.444	0.636	0.274
SVC	**LBP**	0.360	0.341	0.664	0.375	0.360	0.375	0.400	0.128
SVC	**ViT**	0.600	0.164	0.447	0.220	0.600	0.220	0.512	-0.067
RF	**TF-IDF**	0.680	0.223	0.481	0.296	0.680	0.296	0.562	0.049
RF	**BERT**	0.720	0.411	0.591	0.388	0.720	0.388	0.639	0.253
RF	**LBP**	0.600	0.266	0.550	0.285	0.600	0.285	0.572	0.139
RF	**ViT**	0.670	0.169	0.461	0.246	0.670	0.246	0.546	-0.012
DT	**TF-IDF**	0.560	0.277	0.535	0.283	0.560	0.283	0.542	0.088
DT	**BERT**	0.540	0.334	0.555	0.364	0.540	0.364	0.538	0.136
DT	**LBP**	0.530	0.270	0.523	0.259	0.530	0.259	0.523	0.043
DT	**ViT**	0.500	0.217	0.505	0.227	0.500	0.227	0.498	0.004
ET	**TF-IDF**	0.670	0.169	0.461	0.246	0.670	0.246	0.546	-0.012
ET	**BERT**	0.680	0.170	0.463	0.250	0.680	0.250	0.551	0.000
ET	**LBP**	0.660	0.255	0.520	0.285	0.660	0.285	0.578	0.109
ET	**ViT**	0.670	0.169	0.461	0.246	0.670	0.246	0.546	-0.012
XGB	**TF-IDF**	0.590	0.431	0.624	0.392	0.590	0.392	0.603	0.205
XGB	**BERT**	0.660	0.412	0.604	0.400	0.660	0.400	0.623	0.234
XGB	**LBP**	0.600	0.305	0.577	0.306	0.600	0.306	0.584	0.165
XGB	**ViT**	0.630	0.248	0.501	0.274	0.630	0.274	0.555	0.040
MLP	**TF-IDF**	0.650	0.171	0.466	0.239	0.650	0.239	0.542	-0.002
MLP	**BERT**	0.640	0.379	0.582	0.405	0.640	0.405	0.604	0.205
MLP	**LBP**	0.460	0.280	0.586	0.310	0.460	0.310	0.484	0.104
MLP	**ViT**	0.550	0.196	0.471	0.236	0.550	0.236	0.504	-0.027

Table 1. Pontuações nas métricas dos vários classificadores e *encoders* para a competência de coesão.

	Encoder	Acc	Macro Prec	Weighted Prec	Macro Recall	Weighted Recall	Macro F_1	Weighted F_1	Kappa
SVC	**TF-IDF**	0.610	0.416	0.552	0.432	0.610	0.432	0.542	0.398
SVC	**BERT**	0.590	0.362	0.500	0.414	0.590	0.414	0.531	0.374
SVC	**LBP**	0.190	0.101	0.097	0.279	0.190	0.279	0.113	0.010
SVC	**ViT**	0.300	0.179	0.242	0.213	0.300	0.213	0.258	-0.078
RF	**TF-IDF**	0.600	0.403	0.531	0.421	0.600	0.421	0.532	0.380
RF	**BERT**	0.590	0.389	0.523	0.421	0.590	0.421	0.535	0.376
RF	**LBP**	0.350	0.276	0.352	0.276	0.350	0.276	0.347	0.060
RF	**ViT**	0.320	0.205	0.279	0.229	0.320	0.229	0.284	-0.041
DT	**TF-IDF**	0.460	0.330	0.466	0.327	0.460	0.327	0.456	0.225
DT	**BERT**	0.370	0.297	0.379	0.285	0.370	0.285	0.358	0.085
DT	**LBP**	0.270	0.191	0.236	0.227	0.270	0.227	0.244	-0.033
DT	**ViT**	0.280	0.241	0.273	0.308	0.280	0.308	0.262	-0.021
ET	**TF-IDF**	0.620	0.406	0.534	0.436	0.620	0.436	0.540	0.407
ET	**BERT**	0.650	0.355	0.499	0.456	0.650	0.456	0.560	0.458
ET	**LBP**	0.380	0.292	0.370	0.298	0.380	0.298	0.369	0.083
ET	**ViT**	0.320	0.225	0.286	0.231	0.320	0.231	0.278	-0.060
XGB	**TF-IDF**	0.530	0.373	0.510	0.380	0.530	0.380	0.515	0.308
XGB	**BERT**	0.490	0.324	0.450	0.348	0.490	0.348	0.458	0.248
XGB	**LBP**	0.410	0.329	0.408	0.328	0.410	0.328	0.400	0.142
XGB	**ViT**	0.250	0.246	0.304	0.190	0.250	0.190	0.247	-0.109
MLP	**TF-IDF**	0.550	0.455	0.599	0.428	0.550	0.428	0.541	0.338
MLP	**BERT**	0.550	0.365	0.497	0.400	0.550	0.400	0.515	0.322
MLP	**LBP**	0.160	0.083	0.094	0.159	0.160	0.159	0.110	-0.070
MLP	**ViT**	0.290	0.223	0.279	0.223	0.290	0.223	0.276	-0.062

Table 2. Pontuações nas métricas dos vários classificadores e *encoders* para a competência de coerência temática.

	Encoder	Acc	Prec_m	Prec_w	Recall_m	Recall_w	F_{1_m}	F_{1_w}	Kappa
SVC	**TF-IDF**	0.640	0.149	0.420	0.227	0.640	0.227	0.507	-0.009
SVC	**BERT**	0.690	0.403	0.660	0.433	0.690	0.433	0.666	0.397
SVC	**LBP**	0.190	0.199	0.539	0.214	0.190	0.214	0.204	0.086
SVC	**ViT**	0.600	0.196	0.480	0.228	0.600	0.228	0.528	0.054
RF	**TF-IDF**	0.650	0.228	0.491	0.248	0.650	0.248	0.541	0.065
RF	**BERT**	0.740	0.461	0.685	0.415	0.740	0.415	0.682	0.410
RF	**LBP**	0.570	0.187	0.491	0.231	0.570	0.231	0.523	0.104
RF	**ViT**	0.630	0.150	0.422	0.223	0.630	0.223	0.505	-0.008
DT	**TF-IDF**	0.440	0.255	0.499	0.253	0.440	0.253	0.450	0.029
DT	**BERT**	0.530	0.399	0.634	0.349	0.530	0.349	0.562	0.229
DT	**LBP**	0.460	0.227	0.498	0.202	0.460	0.202	0.471	0.027
DT	**ViT**	0.500	0.229	0.493	0.278	0.500	0.278	0.492	0.085
ET	**TF-IDF**	0.640	0.149	0.420	0.227	0.640	0.227	0.507	-0.009
ET	**BERT**	0.640	0.154	0.434	0.227	0.640	0.227	0.517	0.029
ET	**LBP**	0.600	0.181	0.469	0.234	0.600	0.234	0.527	0.087
ET	**ViT**	0.640	0.149	0.420	0.227	0.640	0.227	0.507	-0.009
XGB	**TF-IDF**	0.620	0.373	0.606	0.366	0.620	0.366	0.605	0.256
XGB	**BERT**	0.650	0.280	0.579	0.311	0.650	0.311	0.607	0.269
XGB	**LBP**	0.530	0.203	0.497	0.224	0.530	0.224	0.511	0.097
XGB	**ViT**	0.580	0.180	0.458	0.226	0.580	0.226	0.505	0.030
MLP	**TF-IDF**	0.630	0.172	0.449	0.230	0.630	0.230	0.520	0.027
MLP	**BERT**	0.670	0.427	0.653	0.416	0.670	0.416	0.652	0.348
MLP	**LBP**	0.380	0.253	0.603	0.303	0.380	0.303	0.424	0.160
MLP	**ViT**	0.530	0.172	0.447	0.195	0.530	0.195	0.481	-0.025

Table 3. Pontuações nas métricas dos vários classificadores e *encoders* para a competência de registro formal.

4. Discussão

Os resultados obtidos com o modelo multimodal para a avaliação das redações de estudantes do ensino fundamental, apresentados na tabela acima, não foram tão promissores quanto o esperado. Embora a expectativa fosse que a combinação de informações textuais e visuais trouxesse melhorias significativas na capacidade de classificação, o desempenho geral do modelo ficou aquém, principalmente nas métricas de precisão, recall e F1 para várias das categorias avaliadas.

4.1. Registro Formal

Com uma precisão média ponderada de 64,3%, essa categoria apresentou o melhor desempenho entre todas. Ainda assim, as métricas macro, que consideram a performance por classe, mostram uma queda expressiva, com uma precisão de apenas 42,4% e um F1-score de 40,4%. Esse contraste entre as médias ponderada e macro pode indicar que o modelo está melhor classificado em uma ou duas categorias majoritárias, mas tem dificuldades nas categorias menos representadas, resultando em uma avaliação desequilibrada.

Approach	Cohesion		Formal Register		Text Typology		Thematic Coherence	
	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1
SVC + TF-IDF	-0.012	0.546	-0.009	0.507	**0.176**	0.548	0.398	0.542
SVC + BERT	**0.274**	0.636	0.397	0.666	0.065	0.461	0.374	0.531
SVC + LBP	0.128	0.400	0.086	0.204	-0.030	0.053	0.010	0.113
SVC + ViT	-0.067	0.512	0.054	0.528	0.000	0.460	-0.078	0.258
RF + TF-IDF	0.049	0.562	0.065	0.541	0.161	0.539	0.380	0.532
RF + BERT	0.253	**0.639**	**0.410**	**0.682**	0.016	0.478	0.376	0.535
RF + LBP	0.139	0.572	0.104	0.523	-0.083	0.398	0.060	0.347
RF + ViT	-0.012	0.546	-0.008	0.505	0.159	0.541	-0.041	0.284
DT + TF-IDF	0.088	0.542	0.029	0.450	0.031	0.441	0.225	0.456
DT + BERT	0.136	0.538	0.229	0.562	-0.018	0.427	0.085	0.358
DT + LBP	0.043	0.523	0.027	0.471	0.077	0.451	-0.033	0.244
DT + ViT	0.004	0.498	0.085	0.492	-0.014	0.403	-0.021	0.262
ET + TF-IDF	-0.012	0.546	-0.009	0.507	0.176	0.548	0.407	0.540
ET + BERT	0.000	0.551	0.029	0.517	-0.042	0.448	**0.458**	**0.560**
ET + LBP	0.109	0.578	0.087	0.527	0.020	0.461	0.083	0.369
ET + ViT	-0.012	0.546	-0.009	0.507	**0.176**	0.548	-0.060	0.278
XGB + TF-IDF	0.205	0.603	0.256	0.605	0.128	0.511	0.308	0.515
XGB + BERT	0.234	0.623	0.269	0.607	0.020	0.464	0.248	0.458
XGB + LBP	0.165	0.584	0.097	0.511	0.009	0.426	0.142	0.400
XGB + ViT	0.040	0.555	0.030	0.505	0.091	0.512	-0.109	0.247
MLP + TF-IDF	-0.002	0.542	0.027	0.520	0.287	**0.599**	0.338	0.541
MLP + BERT	0.205	0.604	0.348	0.652	-0.012	0.443	0.322	0.515
MLP + LBP	0.104	0.484	0.160	0.424	0.017	0.097	-0.070	0.110
MLP + ViT	-0.027	0.504	-0.025	0.481	0.127	0.521	-0.062	0.276

Table 4. Comparação de múltiplas abordagens com diferentes métricas.

	Acc	Precision_w	Recall_w	F1_w	Precision_m	Recall_m	F1_m
Registro formal	0.643981	0.642576	0.643981	0.641287	0.424591	0.396777	0.404402
Coerência Temática	0.624575	0.622286	0.624575	0.620606	0.399331	0.414174	0.398246
Tipologia Textual	0.521856	0.543915	0.521856	0.529059	0.368106	0.360199	0.355740
Coesão	0.650679	0.646430	0.650679	0.646595	0.430058	0.395604	0.404842

Table 5. Resultados do Modelo Multi Modal testado

4.2. Coerência Temática

O modelo teve desempenho semelhante ao do Registro Formal, com uma precisão ponderada de 62,4%, mas, novamente, as métricas macro indicam uma baixa performance nas classes menos frequentes. A pontuação macro de F1 é de apenas 39,8%, o que demonstra dificuldade do modelo em generalizar bem para diferentes tipos de redações.

4.3. Tipologia Textual

Esta categoria apresentou o pior desempenho, com uma precisão ponderada de apenas 52,1% e um F1 ponderado de 52,9%. O modelo mostrou particular dificuldade em distinguir as diferentes tipologias textuais, o que pode ser atribuído à complexidade dessa tarefa, já que essa classificação depende de aspectos tanto textuais quanto visuais que talvez o modelo multimodal não tenha capturado de maneira adequada.

4.4. Coesão

Embora tenha atingido uma precisão ponderada de 65%, a mais alta entre as dimensões avaliadas, as métricas macro novamente indicam uma dificuldade em capturar corretamente as variações entre classes. A precisão macro de 43% e o F1 de 40,4% refletem um problema similar ao observado em outras categorias.

5. Conclusão

Os resultados indicam que, apesar da inclusão de informações visuais, o modelo multimodal não foi capaz de atingir as expectativas de melhoria na classificação. A discrepância entre as métricas ponderadas e macro revela que o modelo não conseguiu lidar bem com as diferentes classes, especialmente nas menos frequentes. Para melhorar o desempenho do modelo, algumas abordagens podem ser exploradas, como o ajuste mais fino na arquitetura multimodal e a inclusão de mais dados de treinamento que ajudem o modelo a captar as nuances entre as diferentes classes. Além disso, uma melhor integração entre as modalidades de texto e imagem, talvez por meio de redes mais profundas e específicas para cada tipo de dado, pode levar a uma representação mais rica e significativa dos documentos.

References

Batista, H. H., Barbosa, G. A., Miranda, P., Santos, J., Isotani, S., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Detecção automática de clímax em produções de textos narrativos. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 932–943. SBC.

da Silva Filho, M. W., Nascimento, A. C., Miranda, P., Rodrigues, L., Cordeiro, T., Isotani, S., Bittencourt, I. I., and Mello, R. F. (2023). Automated formal register scoring of student narrative essays written in portuguese. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 1–11. SBC.

Jain, R. and Wigington, C. (2019). Multimodal document image classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–77. IEEE.

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024). Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Ribeiro, E., Mamede, N., and Baptista, J. (2024). Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17.

Su, T., Wang, J., You, H., and Wang, Z. (2023). Multimodal scoring model for handwritten chinese essay. In *International Conference on Document Analysis and Recognition*, pages 505–519. Springer.