Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Avaliação Automática de Narrativas Produzidas por Alunos do 5º Ano do Ensino Fundamental

João Pedro Silva Freitas

JUIZ DE FORA
JULHO, 2026

Avaliação Automática de Narrativas Produzidas por Alunos do 5º Ano do Ensino Fundamental

João Pedro Silva Freitas

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

JULHO, 2026

Avaliação Automática de Narrativas Produzidas por Alunos do 5º Ano do Ensino Fundamental

João Pedro Silva Freitas

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Jairo Francisco de Souza
Doutor em Informática

Marcelo de Oliveira Costa Machado
Doutor em Informática - UNIRIO

Pedro Henrique Gasparetto Lugão
Doutor em Modelagem Computacional - LNCC

JUIZ DE FORA

13 DE JULHO, 2026

Resumo

A avaliação automática de redações em larga escala desempenha um papel fundamental no diagnóstico educacional, permitindo analisar de forma ágil o desempenho de redes de ensino. Contudo, o processamento de textos produzidos por alunos dos anos iniciais do Ensino Fundamental apresenta desafios complexos devido ao desbalanceamento das notas e à forte presença de desvios ortográficos. Este trabalho apresenta um estudo comparativo entre duas abordagens de processamento de linguagem natural (PLN) para a atribuição automática de notas em redações do gênero narrativo escritas por estudantes do 5º ano, avaliando quatro aspectos pedagógicos: Registro Formal, Tipologia Textual, Coerência Temática e Coesão. Foram investigados o modelo clássico *baseline* XGBoost, alimentado por características tabulares extraídas por uma implementação própria baseada no *NILC-Metrix*, e o modelo de aprendizado profundo *BERTimbau*, baseado na arquitetura *Transformers*. A avaliação de desempenho, mensurada predominantemente pela métrica Quadratic Weighted Kappa (QWK), revelou que o BERTimbau superou o baseline clássico (XGBoost) em todas as competências e métricas, com destaque para um ganho expressivo na predição da Coerência Temática, por conseguir processar relações semânticas profundas. Contudo, conclui-se que o modelo XGBoost ainda se mostra uma alternativa altamente viável e equilibrada para cenários de produção, oferecendo alta explicabilidade no mapeamento de dados, provando que a escolha do paradigma ideal depende das prioridades e objetivos do sistema de avaliação em larga escala

Palavras-chave: Avaliação Automática de Redações. Processamento de Linguagem Natural. Modelos de Linguagem Baseados em Transformers. Algoritmos de Aprendizado Supervisionado. Avaliação Educacional em Larga Escala.

Abstract

Large-scale automated essay scoring plays a fundamental role in educational diagnostics, enabling an agile analysis of school system performance. However, processing texts produced by early-grade Elementary School students presents complex challenges due to score imbalance and the high prevalence of orthographic deviations. This work presents a comparative study between two Natural Language Processing (NLP) approaches for automated grading of narrative essays written by 5th-grade students, evaluating four pedagogical aspects: Formal Register, Textual Typology, Thematic Coherence, and Cohesion. The classic *baseline* XGBoost model, powered by tabular features extracted through a proprietary implementation based on *NILC-Metrix*, and the deep learning model *BERTimbau*, based on the *Transformers* architecture, were investigated. The performance evaluation, predominantly measured by the Quadratic Weighted Kappa (QWK) metric, revealed that BERTimbau outperformed the classic baseline (XGBoost) across all competencies and metrics, highlighted by a significant gain in predicting Thematic Coherence due to its ability to process deep semantic relations. Nevertheless, it is concluded that the XGBoost model remains a highly viable and balanced alternative for production scenarios, offering high explainability in data mapping, proving that the choice of the ideal paradigm depends on the priorities and objectives of the large-scale assessment system.

Keywords: Automated Essay Scoring. Natural Language Processing. Transformer-Based Language Models. Supervised Learning Algorithms. Large-Scale Educational Assessment.

Agradecimentos

À minha mãe e a todos os meus familiares, pelo apoio até aqui, e ao meu falecido pai, pelos momentos que ficaram.

Ao meu orientador, pela paciência, e a todos os professores que contribuíram para a minha formação intelectual.

Aos colegas e à coordenação da Fundação CAEd, pelo suporte e pelas condições que viabilizaram a execução deste trabalho.

"Computadores fazem arte

Artistas fazem dinheiro, dinheiro".

Chico Science & Nação Zumbi

(Computadores Fazem Arte)

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computação
UFJF	Universidade Federal de Juiz de Fora
AES	Automated Essay Scoring (Avaliação Automática de Redações)
CAEd	Centro de Políticas Públicas e Avaliação da Educação
PLN	Processamento de Linguagem Natural
ENEM	Exame nacional do Ensino médio
QWK	*Quadratic Weighted Kappa*
NILC	Núcleo Interinstitucional de Linguística Computacional
LSA	Latent Semantic Analysis (Análise Semântica Latente)
BERT	Bidirectional Encoder Representations from Transformers
TRI	Teoria da Resposta ao Item
POS	Part of Speech

1 Introdução

A aquisição e o desenvolvimento da linguagem escrita representam um dos marcos mais importantes na trajetória educacional de um estudante. O 5º ano do Ensino Fundamental consolida-se como uma etapa crucial nesse processo, assinalando a transição da alfabetização inicial para o letramento pleno, momento em que a criança deixa de apenas decodificar símbolos para assumir o papel de autora na produção autônoma de sentidos (SOARES, 2004). Para aferir a eficácia dessa consolidação e garantir a qualidade do ensino, redes educacionais de todo o país recorrem a avaliações em larga escala, frequentemente apoiadas por instituições especializadas, como a Fundação Centro de Políticas Públicas e Avaliação da Educação (Fundação CAEd).

Contudo, a aplicação de avaliações em larga escala que exigem a produção de textos (redações) esbarra em um gargalo logístico e financeiro. Corrigir milhares de redações de forma estritamente manual exige um tempo de resposta demorado e um custo alto para as instituições. Além da barreira da escalabilidade, a avaliação manual é uma atividade suscetível à vieses humanos, tornando a padronização dos critérios de correção entre múltiplos avaliadores humanos um desafio persistente (OLIVEIRA et al., 2023).

Nesse cenário, a Computação surge como uma alternativa viável e promissora. Através dos avanços no Processamento de Linguagem Natural (PLN), os sistemas de Avaliação Automática de Redações (do inglês, *Automated Essay Scoring* - AES) surgem como uma solução tecnológica proposta por Page (1966) em seu trabalho pioneiro. Ao traduzir características textuais em modelos preditivos, as ferramentas de AES têm a capacidade de automatizar ou apoiar o processo de correção com alta consistência, garantindo a escalabilidade necessária para atender às demandas das avaliações em larga escala de forma eficiente (MARINHO; ANCHIÊTA; MOURA, 2021).

1.1 Problema

Embora a literatura internacional documente o sucesso de modelos de AES em diversos cenários, a grande maioria desses trabalhos concentra-se em ensaios acadêmicos maduros, produzidos por estudantes do ensino médio ou ensino superior em exames de alta complexidade (como o Exame Nacional do Ensino Médio - ENEM) (AMORIM; VELOSO, 2018; MARINHO; ANCHIÊTA; MOURA, 2021). Essa concentração evidencia uma lacuna crítica quando o público-alvo se desloca para os anos iniciais da educação básica no contexto brasileiro. Resta investigar como os modelos no estado da arte se desempenham ao processar textos infantis. Ademais, no contexto da avaliação educacional, o processamento dessas redações transcende a atribuição de uma nota global, exigindo a avaliação analítica de múltiplas competências estruturantes da escrita,como o registro formal, a tipologia textual, a coerência e a coesão, adotando-se neste trabalho as métricas padronizadas pela Fundação CAEd.

Associado a este duplo desafio,a imprevisibilidade do público-alvo e a multiplicidade de competências avaliativas, emerge um duelo de paradigmas computacionais. De um lado, encontram-se as abordagens clássicas, baseadas em regras de engenharia de atributos manuais (*Feature Engineering*), que extraem características textuais simples como contagem de palavras até características mais complexas como indices de leiturabilidade por meio de ferramentas como a NILC-Metrix (LEAL et al., 2023) e alimentam algoritmos de aprendizado de máquina, como o XGBoost (FILHO et al., 2023). Do outro lado, o paradigma moderno traz as abordagens de *Deep Learning* contextual, impulsionadas pela arquitetura *Transformer* (como o modelo BERTimbau), que dispensam a extração manual ao aprenderem a semântica de forma bidirecional e densa (SOUZA; NOGUEIRA; LOTUFO, 2020).

Diante disso, formula-se a seguinte questão: **em que medida abordagens baseadas em engenharia explícita de características textuais comparam-se a modelos de aprendizado profundo contextual na tarefa de avaliação automática de múltiplas competências em redações narrativas do 5º ano do ensino fundamental brasileiro?**

1.2 Justificativa

A relevância do presente trabalho sustenta-se sobre dois pilares complementares: o eixo social/pedagógico e o eixo científico.

No pilar da relevância social e pedagógica, destaca-se a construção de um modelo computacional capaz de avaliar competências específicas da escrita, como Registro Formal, Tipologia Textual, Coesão e Coerência (OLIVEIRA et al., 2025). As abordagens propostas viabilizam, em conjunto com os serviços realizados pela Fundação CAEd, a geração de relatórios diagnósticos detalhados para as redes de ensino. Essa automação auxilia gestores e professores na identificação das principais deficiências dos alunos do 5º ano, fundamentando, assim, intervenções pedagógicas direcionadas.

No pilar da relevância científica, este estudo ataca diretamente a escassez de trabalhos de AES focados estritamente na língua portuguesa do Brasil aplicada às séries iniciais (MARINHO; ANCHIÊTA; MOURA, 2021). Enquanto o desenvolvimento de sistemas para o inglês conta com vastos recursos, e o português começa a consolidar bases para o ensino médio, a exploração do ensino básico no campo da inteligência artificial educacional ainda é pequena (AMORIM; VELOSO, 2018). Ao investigar o contraste entre as abordagens de *Machine Learning* tradicional e o *Deep Learning* sobre um corpus narrativo inédito (OLIVEIRA et al., 2025), este trabalho contribui metodologicamente para preencher essa lacuna, expandindo as fronteiras da literatura nacional na área de processamento de linguagem natural educacional.

1.3 Objetivos

1.3.1 Objetivo Geral

A fim de propor uma solução para o problema mencionado, este trabalho se propõe a desenvolver e avaliar sistemas de Automatic Essay Scoring (AES) para a correção de textos narrativos em português brasileiro de alunos do 5º ano do ensino fundamental, com foco na predição de notas para quatro competências de correção: Registro Formal, Tipologia Textual, Coerência Temática e Coesão.

1.3.2 Objetivos Específicos

Para viabilizar o alcance do objetivo geral desta pesquisa, o problema central foi decomposto em etapas metodológicas sequenciais. Os objetivos específicos listados a seguir formam o percurso do estudo, estruturando a evolução da pesquisa desde a padronização dos dados brutos até a modelagem e a avaliação comparativa dos algoritmos. Essa decomposição garante que a investigação do desempenho dos modelos frente à escrita infantil seja conduzida de forma sistemática e reprodutível:

- Consolidar e padronizar um corpus de redações em língua portuguesa provenientes da Fundação CAEd e de datasets de referência da literatura;
- Implementar um pipeline unificado de extração de características linguísticas baseado nas métricas computacionais da Coh-Metrix;
- Desenvolver e otimizar modelos preditivos clássicos, adotando o algoritmo XGBoost como representante do estado da arte para o processamento de dados tabulares de alta dimensionalidade (classificação multiclasse e regressão contínua);
- Investigar o paradigma de *Deep Learning* por meio da Transferência de Aprendizado (*Transfer Learning*) do modelo BERTimbau, explorando a superioridade de arquiteturas *Encoder-only* na compreensão de contexto bidirecional;
- Comparar o desempenho quantitativo (via Quadratic Weighted Kappa) e qualitativo das abordagens, analisando as limitações e sensibilidades de cada modelo frente às particularidades da escrita infantil.

1.4 Organização do Trabalho

Para proporcionar uma leitura estruturada do percurso investigativo, o restante deste documento está organizado em quatro capítulos principais.

O Capítulo 2 apresenta a Fundamentação Teórica e os Trabalhos Relacionados. A seção estabelece o panorama histórico do Aprendizado de Máquina e da Avaliação Automática de Redações (AES), introduzindo os conceitos de Processamento de Linguagem Natural, o contraste arquitetural entre algoritmos baseados em árvores e *Transformers*.

O Capítulo 3 detalha a Metodologia. O texto descreve o processo de integração das fontes de dados, as etapas de pré-processamento e seleção de características, o delineamento experimental das arquiteturas de modelagem clássica (XGBoost) e profunda, e os protocolos de avaliação adotados.

O Capítulo 4 expõe os Resultados e a Discussão. O capítulo analisa o impacto empírico da redução de dimensionalidade e avalia o desempenho quantitativo de cada abordagem preditiva isoladamente. A seção é encerrada com uma discussão comparativa ampla entre os paradigmas de engenharia de características e *Deep Learning*.

Por fim, o Capítulo 5 apresenta a Conclusão da pesquisa, sintetizando os achados, retomando o atendimento aos objetivos propostos e delineando perspectivas para trabalhos futuros.

2 Fundamentação Teórica e Trabalhos relacionados

2.1 Introdução Histórica sobre Machine Learning

A evolução do aprendizado de máquina (*Machine Learning*) revolucionou diversas áreas da computação, baseando-se historicamente no aprendizado supervisionado, paradigma no qual os algoritmos aprendem a mapear dados de entrada para saídas a partir de conjuntos de dados previamente rotulados (GOODFELLOW; BENGIO; COURVILLE, 2016). Em sua trajetória, algoritmos baseados em agrupamento de árvores de decisão, notadamente o XGBoost (*eXtreme Gradient Boosting*), tornaram-se o estado da arte para dados estruturados tradicionais e extração de características devido à sua eficiência em tarefas de classificação e regressão (FILHO et al., 2023).

Posteriormente, a introdução das redes neurais profundas (*Deep Learning*) alterou radicalmente o cenário do processamento de dados não estruturados, como imagens e linguagem natural (LECUN; BENGIO; HINTON, 2015). Esse avanço culminou no surgimento da arquitetura *Transformer*, que substituiu os tradicionais modelos recorrentes por mecanismos de autoatenção (*self-attention*), permitindo o processamento paralelo das sequências e a captura de dependências de longo alcance em textos (VASWANI et al., 2017). Essa arquitetura fundamentou o desenvolvimento dos grandes modelos de linguagem natural (LLMs), consolidando uma nova era no aprendizado supervisionado focado em processamento textual (DEVLIN et al., 2019).

2.2 Panorama Histórico do AES

A Avaliação Automática de Redações (do inglês *Automated Essay Scoring* - AES) é o campo de pesquisa dedicado à criação de sistemas computacionais capazes de atribuir notas a produções textuais de forma automática ou com mínima intervenção humana

(MARINHO; ANCHIÊTA; MOURA, 2021). A evolução dessas tecnologias pode ser classificada em três gerações principais:

A primeira geração do AES foi marcada pelo trabalho pioneiro de Ellis B. Page, em 1966, impulsionado pela necessidade de reduzir a extenuante carga de trabalho dos professores na correção de textos (PAGE, 1966). O sistema desenvolvido por ele, denominado *Project Essay Grade* (PEG), baseava-se em heurísticas estatísticas superficiais (PAGE, 1966). O modelo de Page avaliava correlações entre as características intrínsecas da escrita ("*trins*") e aproximações quantitativas mensuráveis pelo computador ("*proxes*"), como o tamanho do texto, tamanho das palavras e a quantidade de vírgulas, para simular a nota que seria atribuída por um juiz humano (PAGE, 1966).

A segunda geração introduziu o Processamento de Linguagem Natural clássico e ferramentas estatísticas mais sofisticadas. Essa fase é caracterizada pela engenharia de características (*Feature Engineering*), na qual extraem-se atributos linguísticos complexos (como riqueza lexical, sintaxe, frequência de conectivos e métricas do *Coh-Metrix*) para alimentar algoritmos de aprendizado de máquina clássicos (AMORIM; VELOSO, 2018). O *e-rater* é um exemplo clássico que passou a usar múltiplas regressões lineares sobre aspectos focados em gramática, mecânica, estilo e desenvolvimento (ATTALI; BURSTEIN, 2006). Na literatura brasileira, trabalhos empregaram extensas extrações de características com algoritmos de ML como o *Random Forest* e o próprio XGBoost para prever as notas de competências específicas (como registro formal e coesão) nas redações (FILHO et al., 2023).

A terceira geração caracteriza-se pela adoção das redes neurais profundas (*Deep Learning*) e modelos de linguagem pré-treinados, que diminuem a dependência da extração manual de características por meio da representação vetorial (embeddings) do texto inteiro (NGUYEN; DERY, 2016). A introdução da arquitetura *Transformer* abriu portas para o uso de modelos como o BERT, cujos mecanismos de atenção bidirecionais conseguem apreender nuances semânticas contextuais robustas (DEVLIN et al., 2019; VASWANI et al., 2017). No Brasil, a adaptação desse modelo resultou no BERTimbau, voltado à linguagem em português (SOUZA; NOGUEIRA; LOTUFO, 2020). Estudos recentes na área de AES têm reportado que tais modelos de terceira geração, baseados em *Transfor-*

mers, muitas vezes superam a precisão das gerações anteriores ou são combinados a elas para avaliar coesão, coerência e demais competências de redações de forma extremamente acurada (OLIVEIRA et al., 2023).

2.3 Aspectos Avaliados e o Contexto Pedagógico

A produção de narrativas é uma das práticas mais basilares na formação textual da criança. O gênero narrativo caracteriza-se por relatar uma sequência de acontecimentos no tempo e no espaço, vivenciados por personagens reais ou fictícios. Conforme defendem Pasquier e Dolz (1996) *apud* Soares (1999), a instrução voltada à produção textual deve englobar um conjunto de habilidades de escrita provenientes de diversos gêneros textuais, incluindo a narrativa.

O presente trabalho insere-se no contexto pedagógico da avaliação de narrativas produzidas por alunos do 5º ano do ensino fundamental, etapa que encerra os anos iniciais da educação básica. Esta fase escolar é um momento crucial no processo de alfabetização, pois exige que o estudante passe da simples codificação e decodificação gráfica para a produção autônoma de sentidos (SOARES, 2004). A avaliação automatizada dessas produções textuais fundamenta-se em quatro competências principais, alinhadas aos critérios de avaliação de larga escala (OLIVEIRA et al., 2025):

- **Tipologia Textual:** A avaliação da tipologia textual verifica se o estudante compreendeu e executou adequadamente o gênero solicitado. O avaliador, analisa se o aluno produziu, de fato, uma história contendo as partes estruturantes do enredo narrativo e os elementos essenciais (narrador, personagens, tempo e espaço) (OLIVEIRA et al., 2025). Avalia-se, assim, se houve o domínio do gênero ou se o aluno desviou do formato.

- **Registro Formal:** O registro formal diz respeito à adequação do texto à norma-padrão da Língua Portuguesa. Nesta competência, o avaliador analisa os desvios gramaticais, a precisão ortográfica, a correta concordância verbal e nominal, a regência e o uso adequado de pontuação (FILHO et al., 2023).

- **Coesão Textual:** A coesão textual pode ser compreendida como o conjunto de

mecanismos linguísticos responsáveis por garantir a conexão linear e significativa entre as palavras e frases na superfície do texto. Sob a perspectiva de Halliday e Hasan (1976) *apud* Koch (1989), essa engrenagem é dividida em cinco mecanismos fundamentais: a referência (pessoal, demonstrativa e comparativa), a substituição, a elipse, a conjunção e a coesão lexical (manifestada por meio de repetições, sinonímias e hiperonímias). Cada ocorrência desses recursos na estrutura textual estabelece um elo de sentido, denominado na literatura como "laço" ou "elo coesivo" (KOCH, 1989). Expandindo essa visão, Beaugrande e Dressler (1981) *apud* Koch (1989) apontam que a coesão concerne diretamente ao modo como os componentes da superfície do texto se conectam em uma sequência linear por meio de dependências gramaticais. Complementarmente, Marcuschi (1983) *apud* Koch (1989) define os fatores coesivos como estruturadores dessa sequência superficial, argumentando que o conceito ultrapassa os limites sintáticos para atuar como uma espécie de semântica da sintaxe textual. Conclui-se, portanto, que a coesão textual engloba todos os processos de sequencialização que asseguram e tornam recuperável a ligação linguística entre os elementos textuais (KOCH, 1989).

- **Coerência Textual:** Definir a coerência textual de forma isolada é uma tarefa complexa; por esse motivo, ela é frequentemente definida a partir de um conjunto de fatores. Conforme aponta Koch e Travaglia (1997), a coerência envolve a inteligibilidade do texto e a sua capacidade de permitir a construção de um sentido lógico. Sob essa ótica, para que um texto seja considerado coerente, é indispensável que o leitor consiga calcular o seu significado global, estabelecendo conexões sólidas e uma relação de unidade entre todos os elementos que compõem a estrutura textual (KOCH; TRAVAGLIA, 1997).

2.4 Processamento de Linguagem Natural e Engenharia de Características

Historicamente, a extração de características textuais (*Feature Engineering*) divide-se em várias frentes de análise, englobando a análise de legibilidade, riqueza lexical, densidade

sintática e coesão textual (AMORIM; VELOSO, 2018). Para viabilizar a transição do texto bruto para uma representação matemática que os algoritmos de *Machine Learning* possam processar, a literatura apoia-se em diversas ferramentas e bibliotecas de Processamento de Linguagem Natural, como o SpaCy, o NLTK e os dicionários do LIWC (*Linguistic Inquiry and Word Count*) (CAMELO; JUSTINO; MELLO, 2020; ROSA et al., 2025). Dentre as ferramentas disponíveis, destaca-se a Coh-Metrix.

A Coh-Metrix é um sistema computacional robusto de análise textual, amplamente utilizado na área educacional, projetado para extrair métricas aprofundadas sobre a coesão, a coerência e a legibilidade de textos (LEAL et al., 2023). Devido à sua capacidade de extrair atributos complexos, a ferramenta ganhou grande destaque em tarefas de Avaliação Automática de Redações (AES), modelando computacionalmente propriedades linguísticas que vão além das contagens de superfície (CAMELO; JUSTINO; MELLO, 2020).

Para permitir a análise precisa de textos redigidos no Brasil, como narrativas e dissertações escolares, surgiram importantes iniciativas de adaptação das métricas originais para o nosso idioma, consolidando ferramentas como o Coh-Metrix-Port e, posteriormente, a NILC-Metrix (SCARTON; ALUÍSIO, 2010; LEAL et al., 2023). O processo de adaptação do Coh-Metrix para a língua portuguesa é complexo e exige a integração de diversas ferramentas e recursos de PLN de alto desempenho, tais como etiquetadores morfossintáticos (*POS taggers*, como o MXPOST) e analisadores sintáticos (*parsers*), essenciais para lidar com as particularidades da nossa gramática (SCARTON; ALUÍSIO, 2010). A operação dessas ferramentas divide-se em seções especializadas que traduzem os aspectos pedagógicos em variáveis numéricas permitindo contabilizar a informação das palavras (incidência de verbos, advérbios, pronomes, etc.). Mensura-se também a diversidade lexical do estudante por meio da razão de palavras únicas e índices de inteligibilidade do documento, fatores diretamente ligados à facilidade de compreensão do texto (SCARTON; ALUÍSIO, 2010)

Para a quantificação computacional da coesão textual, o sistema analisa a coesão referencial buscando sobreposições explícitas de palavras entre as orações e o uso de pronomes (LEAL et al., 2023). Além disso, incorporam também Índices de leiturabilidade

adaptadas da língua inglesa como Flesch adaptado (LEAL et al., 2023) e métricas adicionais exclusivas para a contagem de múltiplos tipos de marcadores discursivos, mapeando a incidência de conjunções alternativas, conclusivas, explicativas, concessivas e condicionais (CAMELO; JUSTINO; MELLO, 2020).

Por fim, no aspecto mais abstrato a coerência textual, a ferramenta ultrapassa as métricas tradicionais aplicando modelos de representação como a Análise Semântica Latente (LSA - *Latent Semantic Analysis*) (SCARTON; ALUÍSIO, 2010; CAMELO; JUSTINO; MELLO, 2020). Essa técnica calcula o grau de similaridade semântica entre diferentes sentenças e parágrafos de uma mesma redação, convertendo o texto em vetores e calculando o cosseno entre eles para verificar a distância (LEAL et al., 2023).

Finalmente, após a extração de todas essas métricas por ferramentas como a Coh-Metrix e outros recursos de PLN, o conjunto resultante de variáveis numéricas é consolidado em um vetor de características (*feature vector*) que representa matematicamente cada redação (AMORIM; VELOSO, 2018).

2.5 Abordagens de Modelagem: XGBoost *versus* Transformers

A transição da segunda para a terceira era da Avaliação Automática de Redações (AES) é fundamentalmente marcada pela mudança no paradigma de como as máquinas processam a linguagem. O contraste entre os modelos clássicos de aprendizado de máquina e as modernas redes neurais profundas ilustra a evolução do processamento dependente de extração manual para a aprendizagem analítica de representações (OLIVEIRA et al., 2023).

O algoritmo XGBoost é um algoritmo de árvore de decisão que aprimora a lógica por tras dos algoritmos clássicos da técnica de *Gradient Boosting*, construindo um comitê (*ensemble*) de árvores sequenciais de forma aditiva: cada nova árvore no modelo é otimizada matematicamente para corrigir os erros residuais deixados pelas árvores anteriores, resultando em um modelo preditivo eficiente (CHEN; GUESTRIN, 2016).

Entretanto, do ponto de vista computacional, algoritmos não necessariamente

conseguem compreender um texto da mesma forma que um ser humano (ATTALI; BURSTEIN, 2006), logo é necessário que entradas numéricas estritamente estruturadas sejam fornecidas para que o algoritmo consiga predizer a nota. É por esse motivo que a modelagem tradicional depende de forma crucial das *features* tabulares extraídas na etapa de Processamento de Linguagem Natural (AMORIM; VELOSO, 2018).

Por consequência, essa abordagem tradicional deixa de captar o contexto e o significado real das palavras na frase. Isso gera a necessidade de modelos capazes de analisar o sentido de cada termo dentro do texto, evitando que informações valiosas sejam ignoradas pelo sistema.

A terceira era do AES rompeu com a limitação da engenharia manual ao introduzir a arquitetura *Transformer* (VASWANI et al., 2017). Diferente dos modelos sequenciais clássicos e árvores de decisão, o *Transformer* é uma arquitetura de aprendizado profundo (*Deep Learning*) que processa os tokens (pedaços de palavras) de uma redação de forma paralela e em sua totalidade, utilizando uma inovação matemática chamada mecanismo de autoatenção (*self-attention*) (VASWANI et al., 2017).

O mecanismo de autoatenção calcula uma matriz de pesos que relaciona cada palavra da redação com todas as outras palavras do mesmo texto simultaneamente (VASWANI et al., 2017). Ao processar um token, o *self-attention* computa o produto escalar entre representações vetoriais de consulta (*query*), chave (*key*) e valor (*value*), determinando matematicamente o nível de relevância que o modelo deve atribuir aos tokens vizinhos e distantes para compreender o contexto daquela palavra específica (VASWANI et al., 2017). O resultado é a geração de *embeddings* (vetores densos) dinâmicos e contextualizados (DEVLIN et al., 2019). Essa característica elimina a necessidade de engenharia manual de atributos textuais, contrastando diretamente com a abordagem estática de extração de características empregada no pipeline unificado do *Coh-Metrix* e XGBoost.

Ao utilizar modelos baseados na arquitetura *Transformer*, como o BERT, a abordagem metodológica adota o paradigma da transferência de aprendizado (*Transfer Learning*) (DEVLIN et al., 2019). Essa técnica consiste em reaproveitar um modelo previamente treinado em extensos *corpora* textuais, do qual se herda o mapeamento sintático e semântico geral do idioma, para resolver uma tarefa específica. A adaptação para o novo

domínio ocorre por meio de um treinamento adicional e restrito (poucas épocas), focado exclusivamente no conjunto de dados rotulados alvo, que, neste estudo, corresponde às redações infantis e suas respectivas notas (DEVLIN et al., 2019)

2.6 Métricas e avaliação

A métrica principal e mais importante adotada neste trabalho é o *Quadratic Weighted Kappa* (QWK). Historicamente, o QWK consolidou-se como a métrica oficial de avaliação em sistemas de avaliação automática de redações (DOEWES; KURDHI; SAXENA, 2023; AMORIM; VELOSO, 2018). Diferente da simples concordância percentual, o QWK leva em conta a gravidade do erro através da aplicação de pesos quadráticos (DOEWES; KURDHI; SAXENA, 2023).

Por exemplo, em uma escala de notas de 1 a 5, prever a nota 4 quando a nota real é 5 representa um erro menor. No entanto, prever a nota 1 quando a real é 5 é uma falha grave do modelo. O peso quadrático penaliza severamente os erros quando as predições estão muito distantes da nota real dada pelo avaliador humano. Isso torna o QWK a métrica ideal para lidar com a escala de notas do ensino fundamental (DOEWES; KURDHI; SAXENA, 2023; OLIVEIRA et al., 2023). O valor do QWK varia de 0 (quando a concordância ocorre apenas por puro acaso) a 1 (concordância perfeita entre a máquina e o humano) (DOEWES; KURDHI; SAXENA, 2023). Na literatura da área, sistemas que alcançam ou passam de $0,70$ de QWK são considerados aceitáveis para uso em cenários reais ao lado de corretores humanos (DOEWES; KURDHI; SAXENA, 2023).

Como métrica secundária, este trabalho utiliza a Acurácia, que representa o percentual de concordância exata. A acurácia mede a proporção de vezes em que o modelo acertou exatamente a mesma nota dada pelo avaliador humano (GOODFELLOW; BENGIO; COURVILLE, 2016). Foi considerada como secundária por dois motivos: primeiro, ela trata todos os erros da mesma forma, não diferenciando um erro leve de um erro grave; segundo, a acurácia pode dar uma falsa impressão de bom desempenho quando a base de dados é desbalanceada. Por isso, ela serve apenas como um suporte para a análise do QWK.

Por fim, o trabalho também utiliza o *F1-Score* (especificamente em sua versão

Macro). O *F1-Score* é uma métrica que faz um balanço entre a precisão e a sensibilidade do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016). O motivo de sua inclusão neste estudo, é que ele avalia o desempenho do algoritmo classe por classe. Dessa forma, ele impede que o desbalanceamento das notas mascare os erros do modelo nas notas mais raras (como as notas 1 e 5).

2.7 Trabalhos Relacionados

O desenvolvimento de sistemas de Avaliação Automática de Redações (AES) para a língua portuguesa tem apresentado avanços significativos nos últimos anos, evoluindo de abordagens baseadas em engenharia de características para modelos de aprendizado profundo. Esta seção destaca os principais estudos da literatura que fundamentam teórica e metodologicamente este projeto de monografia, o qual visa avaliar narrativas de alunos do 5º ano utilizando tanto um algoritmo de *Machine Learning* clássico quanto um representante da arquitetura *Transformer*, *Encoder-Only*.

A pesquisa em AES no Brasil teve um de seus marcos fundacionais com o trabalho de Amorim e Veloso (2018), que desenvolveram um sistema de classificação multiaspecto pioneiro focado em redações do Exame Nacional do Ensino Médio (ENEM). Utilizando regressão linear e um conjunto de 19 características (*features*) extraídas manualmente, como contagem de erros gramaticais e diversidade lexical, os autores comprovaram a viabilidade da avaliação automatizada para o português, embora seu foco fosse restrito a textos dissertativo-argumentativos de alunos do ensino médio (AMORIM; VELOSO, 2018).

A partir de abordagens como a de Amorim e Veloso (2018), as ferramentas de Processamento de Linguagem Natural para o português brasileiro vêm sendo aprimoradas. Nesse sentido, destaca-se o ecossistema *NILC-Metrix*, proposto por Leal et al. (LEAL et al., 2023). Este sistema representa uma evolução direta da ferramenta pioneira desenvolvida por Scarton e Aluísio (SCARTON; ALUÍSIO, 2010) no Núcleo Interinstitucional de Linguística Computacional (NILC), consolidando 200 métricas computacionais para a extração de propriedades linguísticas de textos escritos e falados no português brasileiro (LEAL et al., 2023). Os autores demonstram o potencial dessas métricas, que abrangem

desde índices descritivos superficiais até análises de coesão semântica latente (*LSA*) e diversidade lexical, na classificação da complexidade textual. Para a presente monografia, as contribuições de Leal et al. (LEAL et al., 2023) e Scarton e Aluísio (SCARTON; ALUÍSIO, 2010) são fundamentais, visto que fornecem a base tecnológica indispensável para a etapa de engenharia de características do pipeline clássico.

No que tange à avaliação específica de competências textuais, Oliveira et al. (OLIVEIRA et al., 2023) investigaram a predição da coesão em redações em português e inglês, contrastando modelos clássicos (como o *CatBoost*) com modelos baseados em *Deep Learning* (BERT). Os resultados demonstraram que, embora os *Transformers* alcancem uma correlação superior com os avaliadores humanos, os modelos tradicionais oferecem uma vantagem significativa na explicabilidade via valores SHAP (OLIVEIRA et al., 2023). Este estudo inspira diretamente o contraste metodológico deste trabalho, que propõe comparar a explicabilidade das métricas tabulares do XGBoost com a representação contextual profunda do modelo BERTimbau na avaliação das competências.

Avançando ainda mais na mensuração da coesão, Rosa et al. (ROSA et al., 2025) propuseram uma abordagem inovadora combinando *Machine Learning* com a Teoria da Resposta ao Item (TRI). Ao aplicar a TRI para ajustar e combinar as predições de múltiplos algoritmos de regressão e modelos baseados em BERT, os autores conseguiram mitigar vieses e melhorar significativamente a acurácia da avaliação da coesão textual em redações brasileiras (ROSA et al., 2025).

Enquanto a maioria dos estudos foca em redações do ENEM, o desafio de avaliar textos de crianças dos anos iniciais tem ganhado tração recentemente. Da Silva Filho et al. (FILHO et al., 2023) abordaram a correção do Registro Formal especificamente em redações narrativas produzidas por alunos do 5º ao 9º ano. Ao extraírem características focadas na gramática formal da língua portuguesa, como concordância e regência verbal/nominal, e utilizarem algoritmos de agrupamento de árvores (como o *Extra-Trees* e o XGBoost), os autores alcançaram um nível de concordância equivalente ao de anotadores humanos (FILHO et al., 2023).

O pilar direto para a realização da presente monografia é o trabalho de Oliveira et al. (OLIVEIRA et al., 2025), que publicou a primeira base de dados aberta de redações

narrativas focada no ensino fundamental brasileiro. O corpus introduzido pelos autores contém 1.235 redações de alunos do 5º ano, avaliadas manualmente por especialistas em quatro competências principais: Registro Formal, Tipologia Textual, Coerência Temática e Coesão (OLIVEIRA et al., 2025).

Este projeto de monografia se posiciona exatamente na interseção dessas pesquisas recentes. Utilizando a base pública de narrativas infantis de Oliveira et al. (OLIVEIRA et al., 2025) complementada por um novo corpus fornecido pela Fundação CAEd, este trabalho estende o estado da arte ao construir modelos preditivos para as quatro competências pedagógicas (Registro Formal, Tipologia Textual, Coerência e Coesão). Inspirando-se no histórico de extração de *features* de Amorim e Veloso (AMORIM; VELOSO, 2018) e Leal et al. (LEAL et al., 2023), e no contraste metodológico de Oliveira et al. (OLIVEIRA et al., 2023), o projeto avaliará a eficácia do *Machine Learning* clássico (XGBoost) contra o estado da arte em *Deep Learning* (BERTimbau) no desafiador contexto da escrita em fase de alfabetização.

3 Metodologia

3.1 Fontes de Dados e Processo de Integração

Para o desenvolvimento e a avaliação dos modelos de *Automatic Essay Scoring* (AES) propostos neste trabalho, a consolidação do *corpus* foi dividida em duas etapas principais: a coleta e amostragem de dados provenientes da Fundação CAEd e, em seguida, a integração estrutural com uma base de dados pública de referência na literatura (OLIVEIRA et al., 2025).

O passo inicial consistiu na extração de uma amostra de redações pertencentes a quatro projetos distintos de avaliação em larga escala, aplicados e corrigidos pela Fundação CAEd. O público-alvo de todos esses projetos foram alunos do 5º ano do Ensino Fundamental. Cada projeto avaliativo possuía um tema diferente, a partir do qual os estudantes eram incentivados a redigir um texto do gênero narrativo com base em um texto motivador.

Visando garantir a representatividade das diferentes habilidades de escrita e mitigar problemas de desbalanceamento de classes durante o treinamento dos algoritmos, adotou-se uma técnica de amostragem estratificada.

Para isso, foram consideradas candidatas à amostra apenas as redações enquadradas em um dos cinco padrões de desempenho com base em sua nota final consolidada (métrica global, distinta das notas individuais por competência). Selecionaram-se, de forma aleatória, 40 textos por padrão. A expectativa inicial era compor uma amostra de 200 redações por projeto, totalizando 800. Contudo, devido à insuficiência de textos em um nível de um projeto específico, o conjunto de dados final extraído do CAEd totalizou 778 redações.

Originalmente, as redações brutas extraídas da Fundação CAEd encontravam-se armazenadas em formato de imagem, contendo os textos manuscritos e as rubricas dos alunos. Uma vez que os algoritmos de Processamento de Linguagem Natural requerem cadeias de caracteres para operar, foi indispensável realizar a conversão desse material.

Após a etapa de coleta, o processamento das redações, passando de imagem para texto, foi realizado de forma manual, contando com o auxílio da ferramenta de transcrição de palavras do modelo de inteligência artificial Gemini. A fim de assegurar o rigor metodológico, todas as transcrições geradas passaram por uma etapa subsequente de revisão humana, garantindo que os desvios gramaticais e ortográficos originais dos estudantes fossem preservados para não enviesar a análise do Registro Formal. Esse suporte tecnológico foi utilizado para otimizar o processo de leitura e garantir a fidelidade da transcrição ao documento físico original. Concluída a transcrição, a amostra coletada foi padronizada e estruturada em um formato tabular, configurada com 6 colunas principais para viabilizar o processamento computacional:

- **ID:** Identificador único da redação;
- **Texto:** A transcrição do texto narrativo produzido pelo aluno;
- **Registro Formal:** Nota atribuída à obediência da norma-padrão;
- **Coerência Temática:** Nota atribuída ao desenvolvimento lógico da história;
- **Tipologia Textual:** Nota atribuída ao respeito à estrutura da narrativa;
- **Coesão:** Nota atribuída à amarração e à conexão do texto.

As quatro últimas variáveis correspondem aos alvos (*labels*) preditivos do estudo. Elas representam as notas finais em cada competência, atribuídas por corretores humanos especialistas da fundação, variando em uma escala discreta de 1 a 5.

Para enriquecer o volume de dados e fortalecer a modelagem do AES, essa base estruturada foi unida ao *benchmark dataset* público de redações narrativas disponibilizado por Oliveira et al. (OLIVEIRA et al., 2025). Visto que o conjunto de dados público engloba o mesmo perfil sociodemográfico (alunos do 5º ano), o mesmo gênero textual (narrativas) e adota os exatos mesmos critérios de avaliação e escala de notas (1 a 5 nas quatro competências citadas) (OLIVEIRA et al., 2025) com avaliadores qualificados, o processo de concatenação ocorreu de maneira direta. A etapa de integração resultou na consolidação de uma base de dados final composta por um total de 2013 registros textuais prontos para as subsequentes fases de pré-processamento.

A consolidação do *corpus* resultou em um conjunto de dados final ilustrado na Figura 3.1, os aspectos de *Registro Formal* e *Coesão* apresentam uma distribuição centralizada, com forte concentração na nota 3. Em contrapartida, a competência de *Tipologia Textual* exibe distribuição deslocada para a direita, com evidente acúmulo de registros na nota 4, enquanto a *Coerência Temática* manifesta uma distribuição mais centradas nas faixas iniciais, caracterizando um leve deslocamento para a esquerda.

Complementarmente, as estatísticas descritivas gerais referentes à extensão dos textos processados são apresentadas a seguir:

- **Tamanho Mínimo do Texto:** 32 caracteres;
- **Tamanho Máximo do Texto:** 2285 caracteres;
- **Tamanho Médio do Texto:** $757,71$ caracteres;
- **Mediana do Tamanho:** $706,00$ caracteres;

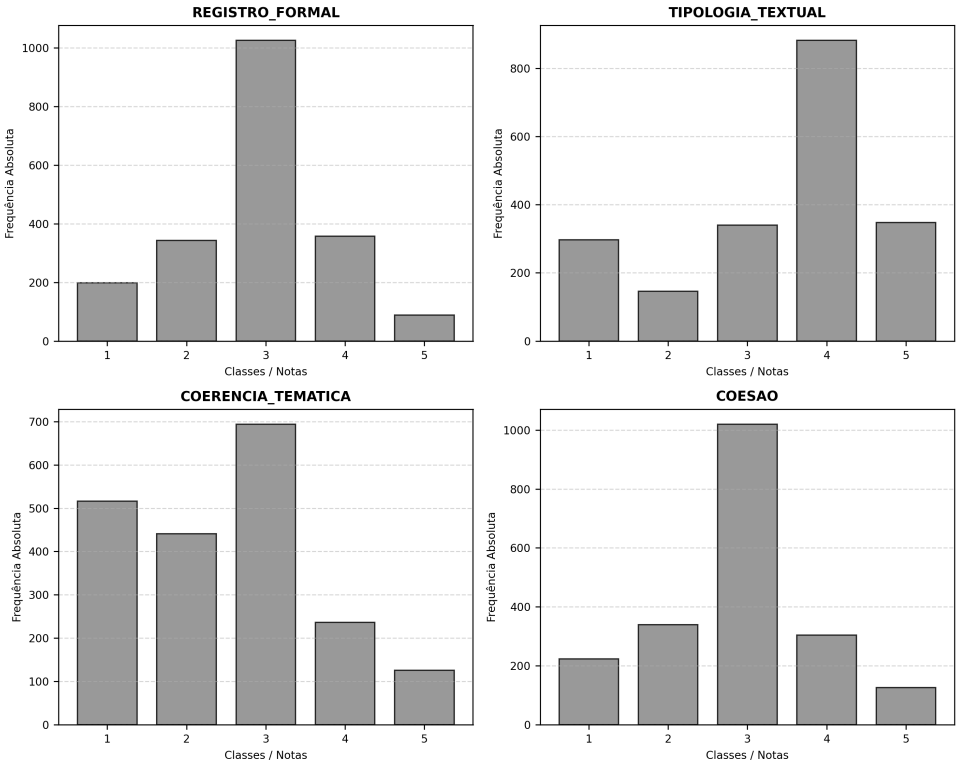

Figura 3.1: Distribuição de frequências das notas (1 a 5) por competência avaliada no corpus consolidado.

3.2 Pré-processamento e Extração de Características

Após a consolidação da base de dados, a etapa seguinte da metodologia consistiu na conversão dos textos brutos em representações matemáticas consumíveis pelos algoritmos de aprendizado de máquina. Para otimizar o fluxo de dados, optou-se por implementar um *pipeline* unificado de pré-processamento e extração de características (*Feature Engineering*), desenvolvido inteiramente em linguagem Python.

A implementação deste processamento foi estruturada em uma classe personalizada denominada `CohMetrix`, inspirada nas métricas computacionais da ferramenta NILC-Metrix (LEAL et al., 2023). Para realizar as análises morfológicas, sintáticas e léxicas de forma acurada, o *pipeline* integrou bibliotecas da literatura de Processamento de Linguagem Natural:

- **Processamento Morfossintático (spaCy):** A fundação do pré-processamento utilizou a biblioteca `spaCy`, empregando o modelo avançado de linguagem para o português (`pt_core_news_lg`). Essa ferramenta foi responsável por executar a tokenização profunda das redações, a lematização das palavras e, fundamentalmente, a etiquetagem morfossintática (*POS Tagging*). O modelo identificou e mapeou classes gramaticais estruturais, como substantivos (*NOUN*), verbos (*VERB*), pronomes (*PRON*) e adjetivos (*ADJ*), bem como conectivos de superfície, como preposições (*ADP*) e conjunções (*CCONJ*).

- **Análise Léxica e Semântica (NLTK e SciPy):** Em complemento ao `spaCy`, o *pipeline* utilizou a biblioteca `NLTK` para análises morfológicas em nível de radical, aplicando o algoritmo `SnowballStemmer` parametrizado para o idioma português. Para o cálculo de medidas de complexidade e densidade de informação no texto, integrou-se o módulo `scipy.stats.entropy`, permitindo avaliar a distribuição probabilística do vocabulário do estudante.

A integração dessas ferramentas em uma única etapa de execução garantiu que as abstrações teóricas detalhadas na fundamentação (como a contagem de pronomes para avaliação de coesão referencial e a análise de POS Tagging para o registro formal) fossem consolidadas em variáveis numéricas baseadas em funções específicas, que podem ser

agrupadas nas seguintes categorias:

- **Coesão Referencial:** Possui métricas que mapeiam a presença de elementos necessários para construir cadeias de correferência. Elas calculam a sobreposição de palavras de conteúdo e de radicais entre sentenças vizinhas. Textos mais longos demandam mais essas conexões para ajudar o leitor a ligar as partes do texto (LEAL et al., 2023).

- **Coesão Semântica:** Baseia-se na Análise Semântica Latente (*LSA*), considerando a sobreposição de palavras que possuem proximidade ou relação de significado. A coocorrência de termos é a base para capturar essas relações semânticas (Landauer et al., 1997 *apud* (LEAL et al., 2023)).

- **Complexidade Sintática:** Traz medidas de proporção que envolvem índices estruturais (como os de Yngve e Frazier) e tipos de orações. Permite investigar a complexidade estrutural do texto, identificando o uso de ordem não-canônica (diferente de Sujeito-Verbo-Objeto), voz passiva, orações subordinadas, relativas e adverbiais (LEAL et al., 2023).

- **Conectivos:** Foca nas palavras que servem explicitamente para ajudar o leitor a estabelecer ligações coesas e lógicas entre as partes e parágrafos do texto (LEAL et al., 2023).

- **Diversidade Lexical:** É calculada por meio da razão entre o número de palavras únicas (*types*) e o total de palavras do texto (*tokens*). Essa diversidade é inversamente proporcional à coesão: quanto menor a variedade de palavras diferentes usadas, maior tende a ser a coesão do texto (McNamara et al., 2014 *apud* Leal et al. (2023)).

- **Índices de Leiturabilidade:** Reúne fórmulas clássicas da literatura utilizadas para avaliar a facilidade de leitura e compreensão de um texto, tais como o Índice de Legibilidade de Flesch, a fórmula adaptada de Dale-Chall, o índice Gunning's Fog, Brunet e Honoré (LEAL et al., 2023).

- **Informações Morfossintáticas de Palavras:** Avalia a densidade de palavras funcionais e de conteúdo no texto e por sentença. Essa categoria detalha a proporção e as flexões de classes gramaticais específicas, como adjetivos, advérbios, verbos, substantivos, preposições e pronomes (LEAL et al., 2023).

- **Medidas Descritivas:** Agrupa as estatísticas básicas que descrevem a estrutura física do texto, tais como a contagem total de palavras, parágrafos e sentenças, além de médias, valores máximos, mínimos e o desvio padrão do comprimento das frases (LEAL et al., 2023).

- **Simplicidade Textual:** Reúne métricas focadas em medir o quão fácil ou acessível um texto se apresenta. Para isso, o sistema calcula a proporção exata de sentenças curtas, médias, longas e muito longas em relação ao total de frases do documento (LEAL et al., 2023).

O resultado desse processo gerou uma base tabular final, onde cada linha representa uma redação vetorizada por suas propriedades linguísticas, pronta para alimentar o algoritmo clássico de predição.

3.3 Análise Exploratória e Seleção de Características

Após a extração das *features*, foi realizada uma análise exploratória dos dados para encontrar características com alta colinearidade e enriquecer o processo de experimentação. A consolidação dessas métricas resultou em um vetor original de alta dimensionalidade. No entanto, o uso irrestrito de todas as *features* gera redundância e ruído, o que pode prejudicar o desempenho do modelo por meio do *overfitting* (sobreajuste). Para isolar os sinais mais relevantes, a seleção de características operou em dois estágios de filtragem:

O primeiro filtro consistiu na análise de colinearidade por meio do Coeficiente de Correlação de Pearson. Estabeleceu-se um limiar de corte de $|r| > 0,7$. Quando duas métricas apresentavam uma correlação superior a este limiar, adotou-se uma heurística de seleção focada na representatividade do atributo: verificou-se a quantidade de correlações que cada variável do par possuía com o restante do conjunto. Optou-se por reter a *feature*

com o maior número de correlações superiores ao limiar, assumindo-a como representante daquele agrupamento de variáveis colineares, e remover as demais. Essa etapa garantiu a redução imediata da dimensionalidade do *dataset*. **Limitações da etapa:** Cabe ressaltar que o critério de seleção adotado priorizou a retenção de variáveis centrais aos clusters de colinearidade por inspeção. Essa heurística difere de métodos baseados na eliminação da variável com maior correlação média global, o que introduz um risco de descarte subótimo em cascata, configurando-se como uma limitação desta etapa de pré-processamento.

Após um treinamento preliminar do algoritmo XGBoost com as *features* sobreviventes ao Filtro 1, extraiu-se o ranqueamento de importância de cada variável (*Feature Importance*). O critério de corte adotado foi a remoção total de atributos que apresentaram grau de importância estritamente igual a zero. Ao remover características que o modelo considerou irrelevantes para os nós de decisão, o algoritmo foi simplificado, reduzindo o custo computacional e preservando apenas o núcleo de *features* que efetivamente balizam a predição das notas das redações.

Após um treinamento preliminar do algoritmo XGBoost com as *features* sobreviventes ao Filtro 1, extraiu-se o ranqueamento de importância de cada variável (*Feature Importance*) como critério para o Filtro 2. Esta é uma métrica nativa do algoritmo que quantifica a contribuição relativa de cada atributo (ganho de informação) para a redução do erro nas ramificações das árvores de decisão. O critério de corte adotado foi a remoção total de atributos que apresentaram grau de importância estritamente igual a zero. Ao descartar variáveis irrelevantes para os nós de decisão, o algoritmo foi simplificado, reduzindo o custo computacional e preservando o núcleo preditivo. O impacto quantitativo detalhado dessas remoções sobre a dimensionalidade do *dataset* é reportado na seção Resultados e Discussão.

Para validar a eficácia da seleção de atributos, o delineamento experimental incluiu um teste comparativo de referência (*baseline*). O algoritmo XGBoost foi inicialmente treinado e avaliado utilizando o vetor original completo sob as mesmas configurações de particionamento. Esse procedimento estabeleceu um referencial de desempenho para contrastar com o modelo treinado sobre o subconjunto reduzido, permitindo isolar o impacto da redução de dimensionalidade.

3.4 Arquiteturas de Modelagem Computacional e Experimentos

Após a preparação dos dados e a seleção de características, a etapa de modelagem focou no treinamento e na otimização dos algoritmos preditivos. Para investigar o contraste entre a Inteligência Artificial clássica e a moderna, o delineamento experimental optou por não realizar um *benchmark* exaustivo de algoritmos, mas sim confrontar os dois paradigmas por meio de seus representantes mais robustos e validados para a língua portuguesa.

No paradigma clássico, selecionou-se o algoritmo XGBoost (*Extreme Gradient Boosting*). A escolha justifica-se pela supremacia comprovada dos métodos de *gradient boosting* sobre dados tabulares estruturados de alta dimensionalidade (como os vetores gerados pelo *pipeline* do Coh-Metrix).

No paradigma de *Deep Learning*, a escolha pelo BERT (*Bidirectional Encoder Representations from Transformers*), especificamente o modelo pré-treinado BERTimbau, baseia-se na natureza da arquitetura frente à tarefa. A avaliação de redações exige a compreensão semântica profunda do texto, o que requer modelos do tipo *Encoder-only*. Ao contrário de arquiteturas autoregressivas (*Decoders*, focadas em geração de texto da esquerda para a direita), o BERT captura o contexto bidirecional simultaneamente.

3.4.1 Modelagem com Aprendizado de Máquina Clássico (XGBoost)

A etapa de modelagem computacional para a Avaliação Automática de Redações apresenta duas possíveis formas de se enxergar a tarefa: a predição das notas pode ser tratada tanto como um problema de regressão quanto como um problema de classificação multiclasse. Na literatura, grande parte dos trabalhos modela a tarefa como regressão, assumindo que as notas progridem em um espectro contínuo de proficiência (OLIVEIRA et al., 2023; AMORIM; VELOSO, 2018). No entanto, considerando que o *corpus* deste trabalho (textos narrativos do 5º ano) adota uma rubrica de correção estritamente discreta, variando em níveis categóricos de 1 a 5 para cada competência, a modelagem via classificação direta das classes também se mostra uma abordagem possível (FILHO et al.,

2023).

Para garantir uma investigação metodológica abrangente, optou-se por implementar e avaliar ambas as abordagens utilizando o algoritmo *Extreme Gradient Boosting* (XGBoost), operando sobre os vetores de características (*features*) extraídos na etapa anterior. Os *pipelines* de treinamento foram desenvolvidos em linguagem Python e encapsulados em duas frentes distintas.

Antes da submissão aos modelos, o conjunto de dados final foi dividido na proporção de 75% para o conjunto de treinamento e 25% para o conjunto de teste. A fim de preservar a distribuição das classes estabelecida na etapa de amostragem, adotou-se a separação estratificada. Adicionalmente, fixou-se *random seed* = 42 para assegurar a reprodutibilidade exata da divisão. Ressalta-se que a otimização de hiperparâmetros e a validação cruzada foram aplicadas estritamente sobre o conjunto de treinamento, isolando o conjunto de teste como dados inéditos para a avaliação final de generalização dos modelos.

Na primeira abordagem, a tarefa foi formulada estritamente como a predição de categorias independentes, utilizando o módulo `XGBClassifier`. O modelo foi configurado para tratar cada nota (1 a 5) como uma classe isolada. Para lidar com o natural desbalanceamento do *dataset* (onde notas extremas, como 1 e 5, possuem menor frequência), a implementação da classificação adotou recursos como o cálculo de pesos por classe (`compute_class_weight`) e o particionamento estratificado (`StratifiedKFold`, com $K = 3$), que divide os dados em K grupos garantindo que a proporção de amostras de cada classe seja mantida em todas as divisões de treino e teste. Isso garantiu que o algoritmo penalizasse mais severamente os erros cometidos nas classes minoritárias durante o treinamento. A otimização de hiperparâmetros nesta frente foi conduzida usando buscas exaustivas (`GridSearchCV`), tendo como função objetivo a minimização da métrica *mlogloss* (Multi-class Log Loss), a qual calcula a entropia cruzada multilogarítmica negativa entre as probabilidades previstas e as classes reais.

Na segunda abordagem, os vetores foram submetidos a um modelo preditivo contínuo utilizando o módulo `XGBRegressor`. Neste cenário, o algoritmo busca ajustar uma função matemática que minimize o erro quadrático em relação à nota real dada pelo

corretor humano. O *pipeline* implementado incorporou o redimensionamento dos dados e utilizou o método de busca em grade (`GridSearchCV`) associado à validação cruzada (*K-Fold*) para explorar o espaço de hiperparâmetros e encontrar a configuração ideal para as árvores de decisão. Contudo, devido à natureza discreta das rubricas, os valores contínuos gerados pela regressão (ex: 3,4 ou 4,1) foram submetidos a um arredondamento matemático para a classe inteira mais próxima imediatamente após a etapa de inferência.

Para garantir a reprodutibilidade do experimento e o controle de complexidade estrutural em ambas as abordagens, definiu-se um domínio restrito para a busca exaustiva de hiperparâmetros. O espaço de busca priorizou o balanceamento entre a capacidade de generalização e a mitigação de *overfitting*, explorando diferentes profundidades de árvore (`max_depth`), taxas de convergência (`learning_rate`) e termos de regularização para a poda estrutural (`gamma`) e penalização de pesos L2 (`reg_lambda`). A Tabela 3.1 detalha as configurações testadas.

Tabela 3.1: Espaço de busca de hiperparâmetros submetido à validação cruzada.

Hiperparâmetro	**Descrição Teórica**	**Domínio de Busca**
`n_estimators`	Número total de árvores construídas	$\{30, 50, 100, 125, 150\}$
`max_depth`	Profundidade máxima de cada árvore	$\{2, 3, 4, 5, 6\}$
`learning_rate`	Taxa de aprendizado (*shrinkage*)	$\{0.01, 0.05, 0.1\}$
`gamma` (γ)	Redução mínima de perda para ramificação	$\{0.1, 0.5, 1.0\}$
`reg_lambda` (λ)	Termo de regularização L2 nos pesos	$\{1.0, 5.0, 10.0\}$

3.4.2 Modelagem com Aprendizado Profundo (*Transformers*)

Para contrastar com a abordagem clássica de engenharia de características, a segunda etapa deste trabalho consiste na implementação de uma arquitetura baseada em *Deep Learning*. Especificamente, adotou-se o modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), uma versão da arquitetura BERT (DEVLIN et al., 2019) pré-treinada massivamente para a língua portuguesa do Brasil. Diferentemente do XGBoost, que dependia da extração computacional de métricas pela Coh-Metrix, esta abordagem consome as transcrições das redações em seu formato bruto (texto livre), delegando ao mecanismo de autoatenção a tarefa de capturar a semântica e os desvios linguísticos de forma contextual.

O *pipeline* de implementação e treinamento foi desenvolvido em linguagem Python, utilizando o ecossistema `PyTorch` em conjunto com a biblioteca `transformers` da Hugging Face. A configuração do modelo foi estruturada nas seguintes etapas metodológicas:

O primeiro passo do fluxo consistiu em adequar as redações ao formato esperado pela rede neural. Utilizou-se o módulo `AutoTokenizer` correspondente ao BERTimbau para segmentar as palavras em subunidades (*subwords*). O processo de formatação dos dados foi encapsulado em uma classe personalizada `TextDataset` (herdando de `torch.utils.data.Dataset`), responsável por padronizar as entradas. Dada a limitação da janela de contexto da arquitetura BERT, o comprimento máximo das sequências (`max_len`) foi fixado em 512 *tokens* (DEVLIN et al., 2019). Redações que ultrapassaram esse limite foram truncadas, enquanto as mais curtas receberam preenchimento artificial (*padding*).

Para fins de simetria experimental, optou-se por modelar a tarefa no *Transformer* também como um problema de classificação multiclasse de notas discretas (escala de 1 a 5). Para a etapa de Transferência de Aprendizado, aplicou-se a classe nativa `AutoModelForSequenceClassification`. Essa configuração preserva os pesos originais do BERTimbau nas camadas de atenção, adicionando um cabeçalho de classificação (*classification head*) linear no topo do *token* especial `[CLS]` (DEVLIN et al., 2019), cujos pesos são inicializados aleatoriamente e treinados para prever a nota de cada competência.

A otimização dos parâmetros da rede foi conduzida pelo algoritmo`AdamW` (*Adam with Weight Decay*). A escolha justifica-se pela característica do método em desacoplar a regularização L2 (decaimento de pesos) da atualização baseada no gradiente, mitigando o sobreajuste (*overfitting*) de forma mais eficaz em arquiteturas *Transformer*. Adicionalmente, adotou-se uma política de escalonamento dinâmico para a taxa de aprendizado, estruturada com uma fase inicial de aquecimento linear (*warmup*) seguida de decaimento gradativo. Essa estratégia é metodologicamente necessária no treinamento de Transformers para evitar divergências prematuras nas primeiras épocas, garantindo que os pesos do cabeçalho de classificação se estabilizem antes do modelo convergir para os mínimos locais.

Além disso, para lidar com o desbalanceamento das notas no *dataset* (conforme

evidenciado na Análise Exploratória), incorporou-se a função `compute_class_weight` da biblioteca *Scikit-learn*. Os pesos calculados para cada classe foram integrados à função de perda (*Loss Function*) do PyTorch, forçando a rede neural a aplicar uma penalidade maior a erros cometidos nas classes minoritárias (como as notas 1 e 5). Para viabilizar o alto custo computacional dessas operações matriciais, todo o processo de treinamento e inferência foi paralelizado utilizando aceleração de *hardware* via GPU (dispositivo `cuda`).

3.5 Protocolo e Métricas de Avaliação

Para avaliar o desempenho dos modelos desenvolvidos (XGBoost e BERTimbau), foram consideradas as duas métricas explicadas na fundamentação teórica: o QWK, como métrica principal, além da acurácia e do *F1-Score* como métricas de suporte. O modelo com a melhor nota de concordância (humano versus máquina) será o escolhido, mas a decisão final também levou em conta as seguintes análises práticas:

- **Presença de erros graves (extremos):** Foi verificado se os modelos cometeram erros absurdos na escala de notas, como prever nota 1 para um texto que o corretor humano avaliou como 5, ou prever 5 para um texto que era nota 1. Modelos que evitam esse tipo de erro são mais confiáveis para o uso real.

- **Comparação das distribuições:** Analisou-se se o formato da distribuição das notas geradas pelo modelo é parecido com o formato do *corpus* original. Isso serve para garantir que o modelo não está apenas "chutando" as notas que mais aparecem para inflar a acurácia, mas sim aprendendo o comportamento real dos corretores.

4 Resultados e Discussão

Este capítulo apresenta os experimentos realizados para a AES no corpus de narrativas do 5º ano do ensino fundamental. A apresentação dos resultados segue o fluxo da metodologia, iniciando-se pela análise dos impactos da filtragem e redução de dimensionalidade dos dados tabulares. Em seguida, realiza-se a avaliação individual da abordagem baseada em aprendizado de máquina clássico (ML Clássico). O capítulo culminará na discussão comparativa entre os paradigmas de classificação e regressão, estabelecendo a linha de base (*baseline*) para os experimentos subsequentes com modelos baseados em *Deep Learning*.

4.1 Impacto da Redução de Dimensionalidade e Relevância de Atributos

A etapa de engenharia de características, fundamentada pelas métricas do Coh-Metrix-PT-BR (LEAL et al., 2023) e bibliotecas de Processamento de Linguagem Natural, resultou em um espaço vetorial inicial de alta dimensionalidade. Esta seção unifica a análise dessas *features* de forma comparativa, avaliando como a seleção espacial afetou o desempenho preditivo dos algoritmos.

O processo de extração computacional gerou um vetor original composto por 108 *features* linguísticas para cada redação. Para mitigar o risco de *overfitting* (sobreajuste) e a redundância informacional, aplicaram-se sucessivamente dois filtros ao conjunto de dados. O Filtro 1 eliminou 30 variáveis por apresentarem alta colinearidade (Correlação de Pearson com $|r| > 0,7$), restando 78 atributos. Em seguida, o Filtro 2 descartou 7 variáveis cujo ganho de informação (*Feature Importance*) no modelo XGBoost foi estritamente nulo. A aplicação rigorosa desses estágios reduziu o espaço vetorial de 108 para 71 *features* finais, otimizando o treinamento ao remover variáveis que apenas adicionavam ruído às fronteiras de decisão do classificador.

Para compreender os critérios de avaliação do modelo, extraiu-se o ranqueamento

global de importância (método *Gain*) do algoritmo XGBoost Classificador. A representação visual das *features* mais determinantes para cada aspecto pode ser observada nas Figuras 4.1.

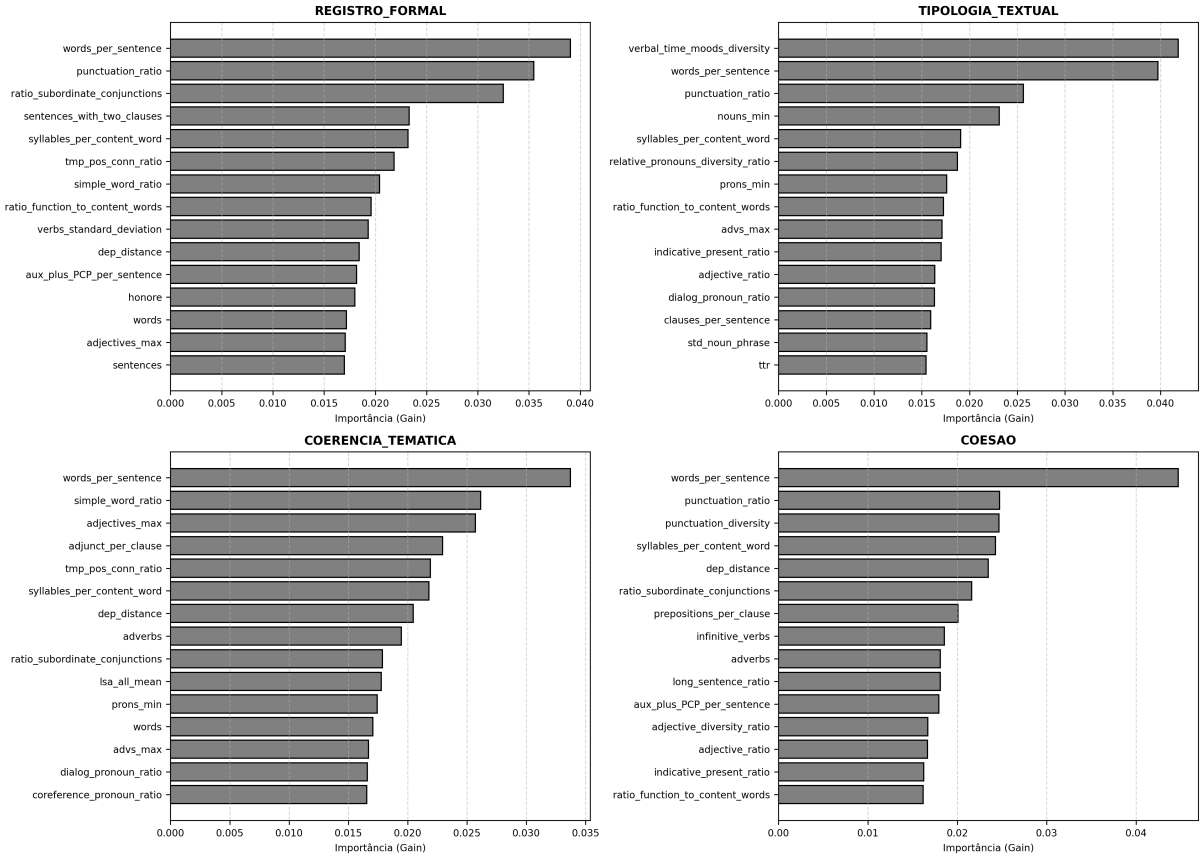

Figura 4.1: Top 15 Features Mais Importantes geradas pelo XGBoost Classificador para cada aspecto.

A análise revela um padrão notável: a característica `words_per_sentence` (tamanho médio das sentenças) dominou a importância global em praticamente todas as competências (Coesão, Coerência, Registro Formal), seguida de perto pela proporção de pontuação (`punctuation_ratio`). Do ponto de vista pedagógico, esse comportamento do algoritmo faz total sentido para o corpus avaliado. Narrativas produzidas por alunos do 5º ano frequentemente sofrem com problemas de segmentação de frases. Estudantes com menor proficiência tendem a escrever parágrafos inteiros sem utilizar pontuação final, unindo ideias ininterruptamente com a conjunção "e"(*run-on sentences*).

Dessa forma, a alta importância atribuída ao tamanho das sentenças e ao volume de pontuação não indica, isoladamente, uma maior maturidade de escrita. Em vez disso, a extensão do texto atua como um estabilizador estatístico: redações mais

longas fornecem uma amostra lexical e sintática mais robusta, viabilizando um cálculo mais preciso e confiável das demais métricas pelo algoritmo, o que refina a qualidade da predição. Para a competência de Tipologia Textual, a diversidade de tempos e modos verbais (`verbal_time_moods_diversity`) desponta como a *feature* de maior peso preditivo. Esse comportamento é coerente com a estrutura do gênero narrativo, cuja progressão exige o domínio e a alternância de tempos verbais para o encadeamento lógico dos acontecimentos.

4.2 Avaliação da Abordagem de Aprendizado de Máquina Clássico (XGBoost)

Esta seção apresenta os resultados de desempenho obtidos pelo algoritmo *Extreme Gradient Boosting* (XGBoost) sobre o conjunto de dados tabulares. Conforme o delineamento estabelecido na metodologia, a avaliação dos experimentos é dividida nas duas frentes de modelagem adotadas: classificação multiclasse e regressão contínua.

A otimização dos hiperparâmetros do `XGBClassifier` foi conduzida de forma independente para cada competência avaliada, visando maximizar o ganho informacional das árvores de decisão diante das características intrínsecas de cada nota. A Tabela 4.1 apresenta a configuração final encontrada pelo método *GridSearchCV*.

Tabela 4.1: Melhores hiperparâmetros encontrados para o XGBoost (Classificação).

Competência	**gamma**	**LR**	**MD**	**NE**	**RL**
Coesão	0.1	0.1	4	100	1.0
Coerência Temática	0.1	0.1	4	100	5.0
Tipologia Textual	0.1	0.1	4	100	1.0
Registro Formal	0.1	0.1	4	100	5.0

* LR: **learning_rate**; MD: **max_depth**; NE: **n_estimators**; RL: **reg_lambda**

Aplicando essas configurações sobre as 71 *features* selecionadas, os resultados preditivos evidenciaram a capacidade do modelo em avaliar os textos de forma automatizada. A Tabela 4.2 exibe as métricas de desempenho obtidas.

Conforme o delineamento experimental, avaliou-se o impacto da seleção de características através do contraste entre o modelo *baseline* (vetor completo com 108 variáveis)

Tabela 4.2: Métricas de Desempenho do XGBoost Classificador por Aspecto.

Aspecto Avaliado	**QWK**	**Acurácia Global**	**F1-Score (Macro)**
Registro Formal	0.5488	0.4960	0.3842
Tipologia Textual	0.4477	0.4306	0.3692
Coerência Temática	0.4215	0.4028	0.3644
Coesão	0.4144	0.5159	0.4367

e o modelo simplificado (71 variáveis). Para a tarefa de classificação, o modelo *baseline* alcançou um QWK médio de 0,464800 entre as quatro competências. O modelo simplificado manteve o desempenho com um QWK médio de 0,458100, ressaltando-se que nenhuma das competências individuais apresentou uma degradação superior a 0,05. Esse comportamento evidencia que os filtros foram eficazes em eliminar variáveis ruidosas e gerar um sistema computacionalmente mais leve sem comprometer a eficácia geral do modelo.

Destaca-se o Registro Formal como o aspecto melhor avaliado ($QWK = 0,5488$), justificado pelo fato de desvios ortográficos e gramaticais serem facilmente quantificáveis de forma exata por meio de análise de *POS tagging*. Por outro lado, percebe-se que a Coerência foi o aspecto com menor desempenho com um QWK por volta de $0,4215$. Isto colabora com a hipótese de que a engenharia de *features*, embora eficaz, esbarra no limite da compreensão do texto, não conseguindo transcender essa barreira apenas com contagens matemáticas superficiais sem um contexto semântico.

A análise detalhada do comportamento do *XGBoost* Classificador pode ser realizada por meio da visualização das matrizes de confusão geradas para cada competência, conforme apresentado na Figura 4.2. De forma geral, observa-se que o modelo apresenta uma forte tendência a concentrar suas previsões nas notas intermediárias e majoritárias (especialmente na nota 3), o que é um reflexo direto da distribuição da base de dados original.

As maiores dificuldades do classificador tradicional ficam evidentes nos aspectos de Tipologia Textual e Coerência Temática. Na matriz de Tipologia Textual, nota-se uma tendência acentuada de previsões voltadas para a nota 1. Esse comportamento pode ser explicado pela grande presença de erros de ortografia e desvios gramaticais característicos dessa etapa escolar, o que provavelmente gerou um viés nas regras de decisão do algoritmo baseado em contagens de palavras e a identificação dos tempos verbais.

Por outro lado, ao analisar as matrizes dos demais aspectos, constata-se que a incidência de erros crassos e distantes da diagonal principal foi pequena. Esse padrão é visível principalmente no triângulo inferior das matrizes, onde o modelo atribuiu uma nota ligeiramente superior à nota real do corretor humano. Sob uma perspectiva pedagógica focada no estudante, esses desvios podem ser interpretados como erros de menor impacto prático, uma vez que o aluno avaliado pelo sistema automático não seria prejudicado por uma subestimação severa de seu desempenho.

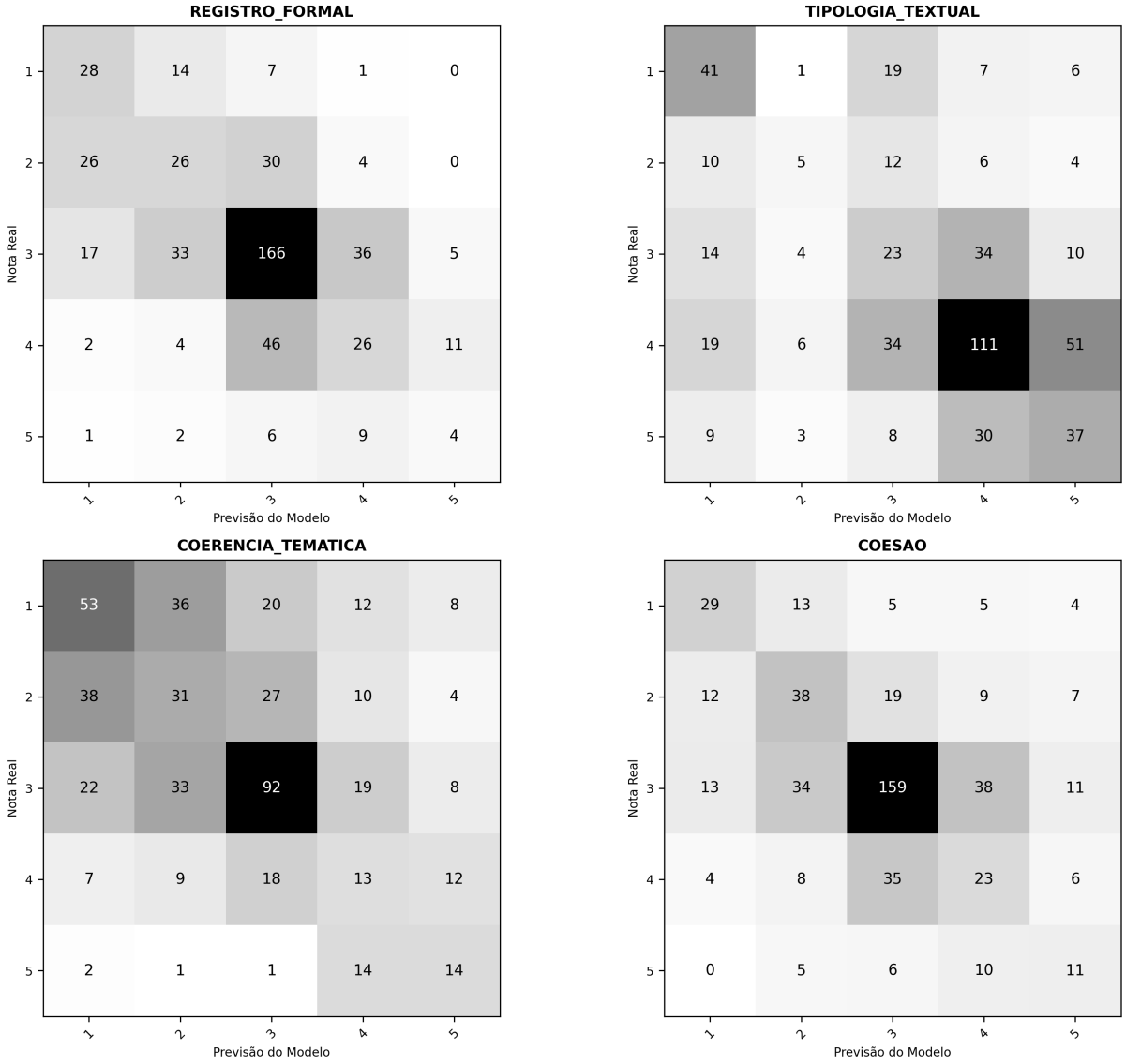

Figura 4.2: Matrizes de confusão obtidas pelo modelo XGBoost Classificador para os aspectos avaliados na escala de 1 a 5. Disposição dos gráficos: Registro Formal (superior esquerdo), Tipologia Textual (superior direito), Coerência Temática (inferior esquerdo) e Coesão (inferior direito).

Como contraponto, o `XGBRegressor` foi treinado sobre o espaço tabular contendo

também 71 características selecionadas pelo processo de filtragem de *features* [1]. Visto que a saída gerada pela regressão é um número contínuo, aplicou-se uma etapa de limiarização e arredondamento padrão para converter as previsões matemáticas aos níveis discretos reais da rubrica de correção (1 a 5). Os hiperparâmetros ótimos localizados para este paradigma encontram-se na Tabela 4.3.

Diferente do que ocorreu na classificação, a redução da quantidade de características para o modelo de regressão causou uma ligeira queda nos resultados preditivos. Apesar disso, optou-se por manter o modelo simplificado com as 71 *features* a fim de evitar o sobreajuste (*overfitting*) e garantir uma comparação justa entre as duas abordagens.

Tabela 4.3: Melhores hiperparâmetros encontrados para o XGBoost (Regressão).

Competência	**gamma**	**LR**	**MD**	**NE**	**RL**
Coesão	1.0	0.1	3	100	10.0
Coerência Temática	0.1	0.1	4	100	5.0
Tipologia Textual	1.0	0.05	4	100	10.0
Registro Formal	0.1	0.05	4	100	10.0

* LR: **learning_rate**; MD: **max_depth**; NE: **n_estimators**; RL: **reg_lambda**

Observa-se que a regressão exigiu um termo de regularização L2 (`reg_lambda`) muito mais severo (variando de 5.0 a 10.0) em relação à classificação e também um gamma mais alto para Coesão e Tipologia Textual resultando em um modelo mais conservador. Os resultados de avaliação sobre as notas discretizadas estão apresentados na Tabela 4.4.

Tabela 4.4: Métricas de Desempenho do XGBoost Regressor (após limiarização).

Aspecto Avaliado	**QWK**	**Acurácia Global**	**F1-Score (Macro)**
Tipologia Textual	0.4396	0.4127	0.2593
Coerência Temática	0.4133	0.3651	0.2432
Registro Formal	0.4041	0.5397	0.3070
Coesão	0.3900	0.5397	0.3040

Apesar de o F1-Score e a Acurácia apresentarem oscilações (com viés de predição concentrado nas classes majoritárias, o QWK manteve-se competitivo. No entanto, ao contrastar com a Tabela 4.2, conclui-se que a Classificação provou-se globalmente superior para o *corpus* analisado, especialmente na avaliação de Registro Formal, onde obteve quase

[1] Cumpre destacar que o quantitativo idêntico ao do modelo de classificação configurou-se como uma coincidência do processo de corte, uma vez que os atributos selecionados para a regressão não coincidem com os da abordagem de classificação.

0, 15 pontos a mais no QWK. Isso indica que, para o sistema de rubricas desta base (anos iniciais, com notas de 1 a 5), estabelecer fronteiras de decisão discretas funcionou melhor do que tentar ajustar uma curva de erro contínua.

A análise visual do comportamento do *XGBoost Regressor*, apresentada na Figura 4.3, evidencia um fenômeno marcante em todas as quatro matrizes: a incapacidade absoluta do modelo em predizer a nota máxima (nota 5). A coluna correspondente a essa nota encontra-se completamente zerada em todas as competências, o que demonstra que o algoritmo tendeu a suavizar suas previsões em direção à média.

Esse comportamento pode ser explicado pelo desbalanceamento das classes combinado com o método de arredondamento das previsões contínuas. As notas das extremidades (1 e 5) foram amplamente preteridas pelo sistema porque, matematicamente, os intervalos válidos para que a limiarização retorne essas notas são bem menores em comparação com os intervalos das notas intermediárias (2, 3 e 4). Associado à escassez de dados nos extremos, esse achatamento de intervalos impediu o modelo de atingir as notas limites da escala.

Por fim, a matriz de *Coerência Temática* (quadrante inferior esquerdo) expõe a maior inconsistência do modelo de regressão: o algoritmo viciou suas predições na nota 2, distribuindo seus erros de forma generalizada. O modelo previu nota 2 para 78 redações que eram nota 1, para 73 redações que eram nota 3 e até mesmo para 13 redações que eram nota 4. Esses resultados consolidam a conclusão de que, para este *corpus*, a abordagem por classificação discreta é substancialmente superior à regressão, pois esta última foi severamente penalizada pelo desbalanceamento das notas, tornando-se incapaz de mapear a real amplitude das notas atribuídas pelos avaliadores humanos.

De toda forma, pela transparência metodológica, os fluxos e resultados de ambas as abordagens (regressão e classificação) foram mantidos na seção de resultados, servindo como a base de referência (*baseline*) tradicional que antecede os experimentos com a arquitetura *Transformer*.

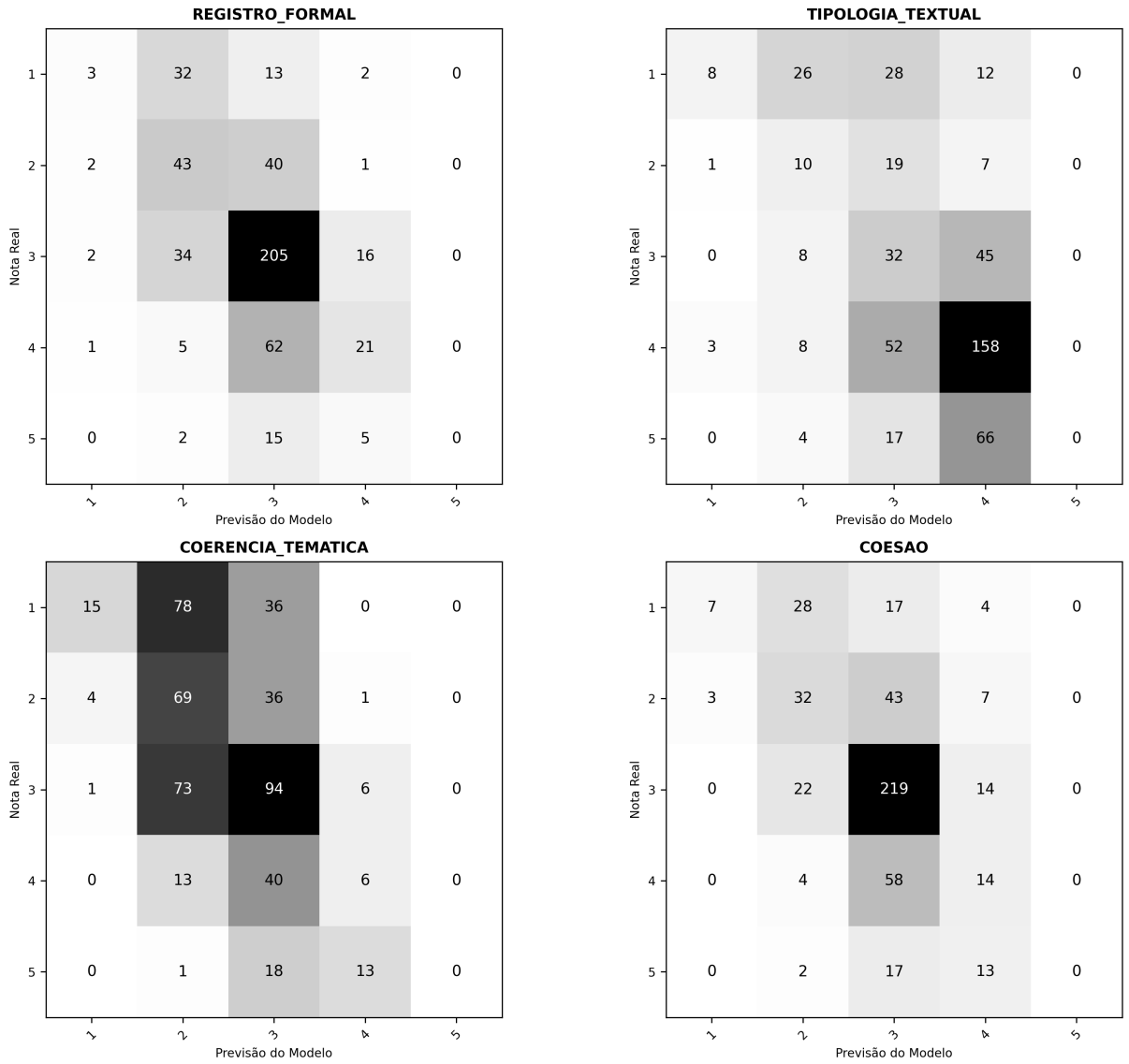

Figura 4.3: Matrizes de confusão obtidas pelo modelo XGBoost Regressor para os aspectos avaliados na escala de 1 a 5. Disposição dos gráficos: Registro Formal (superior esquerdo), Tipologia Textual (superior direito), Coerência Temática (inferior esquerdo) e Coesão (inferior direito).

4.3 Avaliação da Abordagem de Aprendizado Profundo (*Transformers*)

Como contraponto à modelagem baseada em extração manual de características (XGBoost + Coh-Metrix), esta seção avalia o desempenho da abordagem de aprendizado profundo, o estado da arte em Processamento de Linguagem Natural. O foco recai sobre o *Transfer learning* (Transferência de Aprendizado) do modelo de linguagem pré-treinado BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). Diferente da abordagem anterior, que dependia de variáveis tabulares, o *Transformer* consome a transcrição do texto bruto,

utilizando seu mecanismo de autoatenção (*self-attention*) bidirecional para apreender as nuances semânticas e sintáticas diretamente da sequência de palavras (DEVLIN et al., 2019).

A adaptação do BERTimbau para a tarefa de AES exigiu uma parametrização rigorosa para lidar com a complexidade da rede neural e as características do *dataset*. Utilizou-se a variante `neuralmind/bert-base-portuguese-cased` (BERTimbau Base), preservando a diferenciação de maiúsculas e minúsculas, um fator relevante para a análise gramatical. O tamanho máximo de sequência (`max_len`) foi fixado no limite arquitetural de 512 *tokens*, com truncamento e preenchimento forçado (*padding*) para padronização dos tensores de entrada.

A análise do processo de tokenização realizado pela biblioteca *transformers* revelou propriedades importantes sobre o comportamento estrutural do *corpus*. Ao avaliar as 2.013 redações, observou-se que o comprimento médio dos textos após a tokenização foi de 218,06 tokens, com uma mediana de 199,00 tokens. O documento mais curto apresentou apenas 17 tokens, enquanto a redação mais longa atingiu o pico de 1.205 tokens. Como pode ser observado na Figura 4.4, a grande maioria dos textos concentra-se bem abaixo do teto operacional de 512 tokens utilizado. Esse cenário é altamente positivo, visto que apenas 37 redações (o equivalente a 1,84% do total da base) ultrapassaram esse limite e precisaram ser cortadas, indicando que o modelo conseguiu processar o contexto integral da quase totalidade dos estudantes sem perda relevante de informação.

Outro fator de destaque diz respeito à cobertura e ao alinhamento do vocabulário infantil com o modelo pré-treinado. Do vocabulário global do BERTimbau, composto por 29.794 tokens únicos, o *dataset* analisado mobilizou um total de 8.417 tokens únicos, o que representa uma cobertura de 28,25% do dicionário do modelo. Diante da natureza dos textos de alunos do 5º ano, tipicamente ricos em desvios ortográficos, a quantidade de tokens desconhecidos gerados foi extremamente baixa, totalizando apenas 223 ocorrências de tokens `[UNK]` (*unknown*).

O treinamento foi conduzido ao longo de 4 épocas (*epochs*). Para viabilizar o treinamento na memória da GPU, o tamanho do lote (*batch size*) físico foi definido como 8, mas aliado a 4 passos de acumulação de gradiente (`accumulation_steps = 4`),

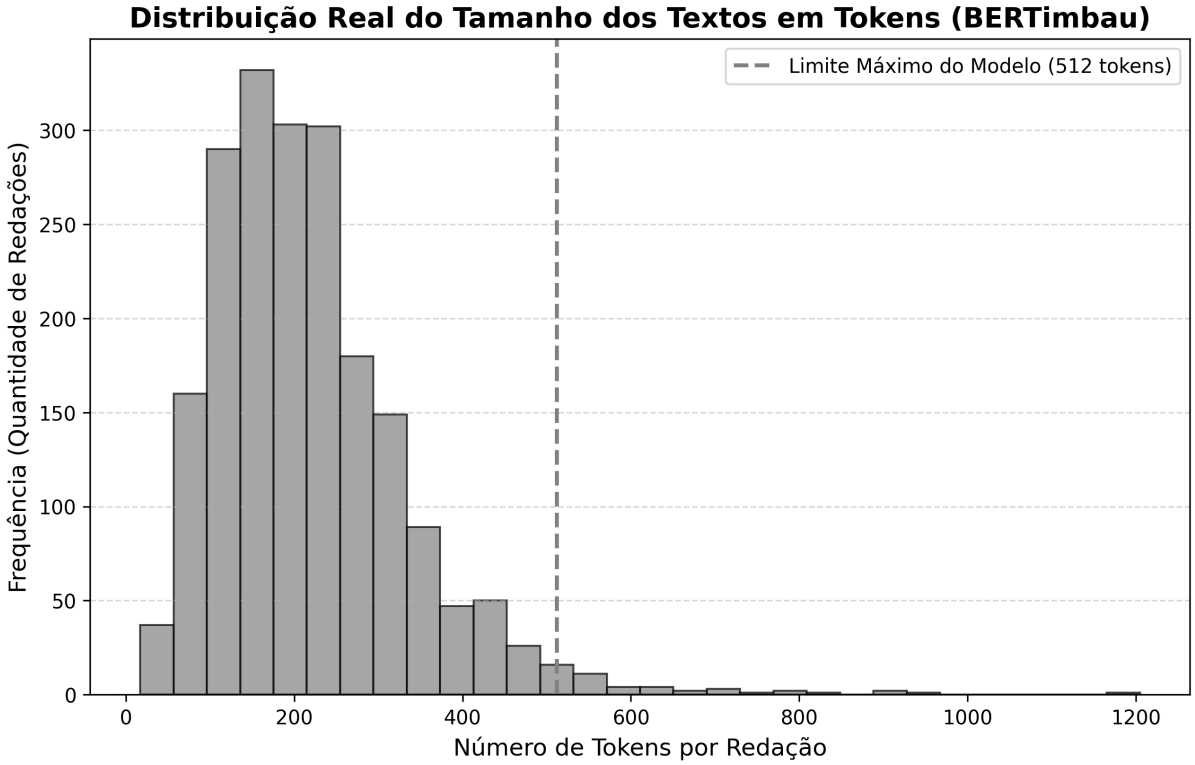

Figura 4.4: distribuição de redações por quantidade de tokens.

resultando em um tamanho de lote efetivo de 32. A otimização dos pesos foi realizada pelo algoritmo `AdamW`, configurado com uma taxa de aprendizado (*learning rate*) conservadora de 2×10^{-5} e um fator de decaimento de peso (*weight decay*) de $0,01$. Para a estabilidade inicial da rede, aplicou-se um *scheduler* com *warmup* linear durante os primeiros 10% dos passos de treinamento.

Após a adaptação da rede, o modelo foi submetido ao conjunto de teste seguindo o mesmo protocolo de avaliação estabelecido para os modelos clássicos. A Tabela 4.5 consolida as métricas preditivas alcançadas pelo BERTimbau para as quatro competências pedagógicas.

Tabela 4.5: Métricas de Desempenho do modelo BERTimbau (Transformer) por Aspecto Avaliado.

Aspecto Avaliado	**QWK**	**Acurácia Global**	**F1-Score (Macro)**
Coerência Temática	**0.6431**	**0.5595**	**0.4786**
Tipologia Textual	0.5994	0.4583	0.4224
Registro Formal	0.5951	0.5456	0.4488
Coesão	0.5707	0.5655	0.4962

A análise das predições do *Transformer* revela um salto quantitativo notável em relação ao XGBoost. O modelo BERTimbau apresentou um desempenho superior e mais

equilibrado em todas as frentes de avaliação. O aspecto mais discrepante desses resultados é o comportamento do modelo diante da Coerência Temática. Enquanto o XGBoost (ML Clássico) atingiu um QWK de $0,4215$ para esta competência, o BERTimbau alcançou $0,6431$ — uma evolução de mais de 22 pontos percentuais na métrica principal.

Este resultado reforça a hipótese levantada na fundamentação teórica deste trabalho e na literatura recente (OLIVEIRA et al., 2023): a coerência, por representar a conexão mental que exige um contexto, não pode ser mensurada eficazmente por contagens estatísticas de superfície (como as do Coh-Metrix consumidas pelo XGBoost). O BERTimbau, ao utilizar camadas de atenção bidirecional para capturar o contexto global do documento, conseguiu simular a percepção humana de sentido e contradição intrínseca da história infantil de forma muito mais acurada.

Além disso, as métricas de Tipologia Textual ($QWK = 0,5994$), Registro Formal ($QWK = 0,5951$) e Coesão ($QWK = 0,5707$) aproximam-se do limiar de $0,60$, o qual é considerado na literatura um indicativo de moderada concordância com o avaliador humano (DOEWES; KURDHI; SAXENA, 2023). Embora o F1-Score Macro ainda indique desafios na classificação das categorias minoritárias (como evidenciado nas matrizes de confusão 4.5), o peso quadrático do QWK demonstra que os erros cometidos pelo BERTimbau foram, em sua imensa maioria, erros "brandos"(predizer a nota vizinha imediata, como um 3 em vez de um 4), provando a viabilidade e a superioridade da abordagem de aprendizado profundo para a correção de redações de alunos dos anos iniciais.

Constata-se que o *BERTimbau* apresentou um comportamento semelhante ao do *baseline* no aspecto de Tipologia Textual, mantendo a tendência de concentrar um volume expressivo de predições na nota 1. Esse cenário levanta a hipótese de que esse viés não decorre apenas de limitações no tratamento do texto, mas pode ser uma característica inerente à própria tarefa de classificação discreta sob este *corpus*. Reafirma-se esse argumento ao considerar que, diferentemente do modelo tradicional, o *Transformer* utiliza um método de tokenização superior e dinâmico, capaz de capturar com maior precisão a morfologia e o contexto das palavras.

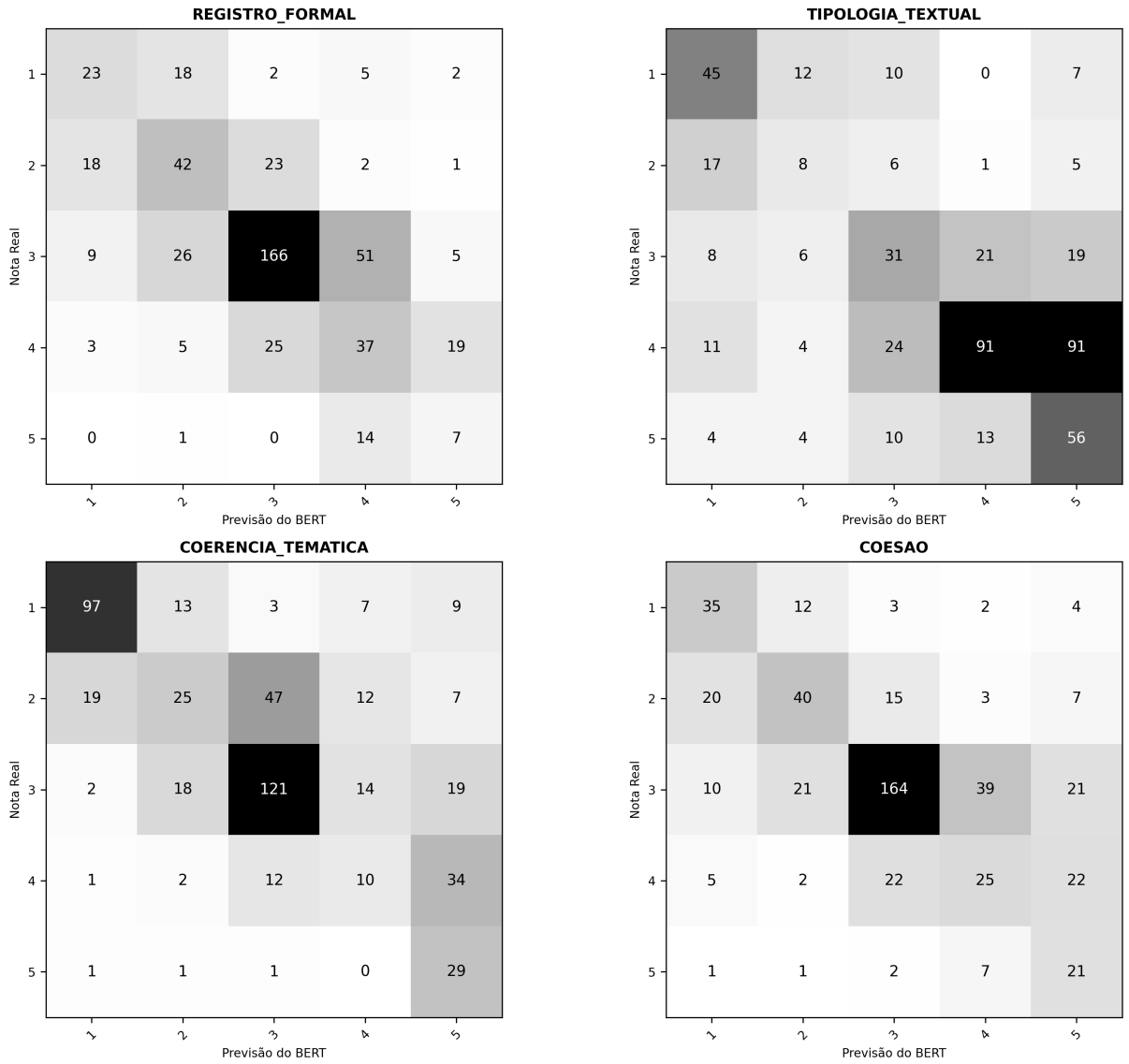

Figura 4.5: Matrizes de confusão obtidas pelo modelo BERTimbau para os aspectos avaliados na escala de 1 a 5. Disposição dos gráficos: Registro Formal (superior esquerdo), Tipologia Textual (superior direito), Coerência Temática (inferior esquerdo) e Coesão (inferior direito).

4.4 Discussão Geral: Engenharia de Características versus *Deep Learning*

A Tabela 4.6 consolida o confronto direto entre a abordagem clássica de aprendizado de máquina (*XGBoost* alimentado com características do *NILC-Metrix*) e o modelo baseado em aprendizado profundo (*BERTimbau*). Os resultados revelam que o *BERTimbau* superou o *baseline* tradicional em absolutamente todos os aspectos e métricas avaliadas.

O salto de desempenho mais expressivo ocorreu no aspecto de Coerência Temática, onde o índice QWK saltou de $0,4215$ para $0,6431$ (um ganho real de mais de $0,22$ pon-

tos). Esse comportamento corrobora a hipótese inicial deste trabalho: enquanto os modelos clássicos dependem de contagens e propriedades estatísticas superficiais da superfície textual que falham em capturar o sentido global, a arquitetura baseada em mecanismos de autoatenção do *Transformer* consegue processar com sucesso as relações semânticas profundas e o contexto entre as palavras, mostrando-se substancialmente superior para a tarefa de avaliação automática em redações dos anos iniciais.

Tabela 4.6: Comparação de Desempenho entre o Baseline (XGBoost) e o Modelo Proposto (BERTimbau).

	QWK		**Acurácia Global**		**F1-Score (Macro)**	
Aspecto Avaliado	**XGB**	**BERT**	**XGB**	**BERT**	**XGB**	**BERT**
Registro Formal	0.5488	**0.5951**	0.4960	**0.5456**	0.3842	**0.4488**
Tipologia Textual	0.4477	**0.5994**	0.4306	**0.4583**	0.3692	**0.4224**
Coerência Temática	0.4215	**0.6431**	0.4028	**0.5595**	0.3644	**0.4786**
Coesão	0.4144	**0.5707**	0.5159	**0.5655**	0.4367	**0.4962**

* **XGB**: XGBoost Classificador; **BERT**: BERTimbau.

Embora o desempenho superior do *BERTimbau* fosse esperado devido à robustez arquitetural dos *Transformers*, faz-se indispensável discutir os resultados além das métricas. Uma das principais hipóteses para a divergência de desempenho se encontra na fase inicial do fluxo de processamento: a tokenização. O modelo proposto utiliza o algoritmo *WordPiece* (DEVLIN et al., 2019), que opera com subpalavras, permitindo ao modelo decompor termos grafados incorretamente sem perder a associação morfológica. Aliado aos mecanismos de autoatenção, esse processo viabiliza a captura de contextos e dependências semânticas de longo alcance. Em contrapartida, para que a extração de características do modelo *baseline* atingisse o seu potencial máximo, ela dependeria de uma tokenização e de uma classificação morfológica muito mais robustas na base de dados. Como o desenvolvimento dessas ferramentas de extração do zero não foi o foco principal deste trabalho, essa etapa acabou limitando o desempenho do modelo clássico.

Por outro lado, a abordagem *baseline* baseada no XGBoost garante uma transparência significativamente maior no mapeamento dos dados, permitindo inspecionar a importância de cada *feature* tabular gerada pelo *NILC-Metrix* na composição da nota final. Em cenários de alta sensibilidade, como as avaliações educacionais de larga escala, essa interpretabilidade é valiosa, pois fornece justificativas auditáveis para notas inespera-

das. O modelo *Transformer*, em sentido oposto, opera sob o paradigma de "caixa-preta", no qual as decisões preditivas são diluídas em milhões de parâmetros.

Adicionalmente, torna-se relevante comparar as distribuições de predição de cada aspecto, apresentadas nas Figuras 4.6 e 4.7, com a distribuição real do *corpus* original ilustrada na Figura 3.1. Essa análise visual permite identificar um fenômeno que as métricas globais e as matrizes de confusão não evidenciam de forma direta: o modelo *baseline* (XGBoost) aproximou-se da distribuição real com maior precisão do que o *Transformer*.

Essa aderência do *baseline* à curva real pode sinalizar um comportamento ambíguo. Por um lado, indica que o modelo clássico conseguiu mapear as proporções macroscópicas das notas. Por outro lado, há o risco de o algoritmo estar sofrendo com um sobreajuste (*overfitting*) direcionado à distribuição marginal dos dados, funcionando mais como um reprodutor das frequências das classes do que como um avaliador de proficiência individual. Apesar dessa limitação de generalização fina, esse comportamento confere um ponto positivo ao XGBoost quando pensado para aplicações em avaliações de larga escala, como as conduzidas pelo CAED, onde o objetivo primordial reside no diagnóstico estatístico da rede de ensino como um todo, e não necessariamente na precisão cirúrgica da nota de um aluno isolado.

Contudo, essa característica levanta um questionamento crítico sobre a robustez temporal do modelo tradicional: se o algoritmo tende a reproduzir rigidamente a distribuição do *dataset* de treino, ele seria capaz de capturar mudanças estruturais ao longo do tempo? Em um cenário ideal, onde políticas públicas melhorem a qualidade da educação nos anos seguintes e desloquem a curva de notas reais para a direita (em direção a scores mais altos), um modelo condicionado à distribuição antiga falharia em detectar essa evolução, tendendo a forçar os novos dados dentro do padrão histórico memorizado. Portanto, a aparente vantagem macro do *baseline* deve ser encarada com cautela frente à flexibilidade adaptativa demonstrada pelas redes neurais profundas.

Por fim, os resultados indicam que a escolha do modelo ideal depende diretamente do objetivo da aplicação prática. Se a prioridade do sistema for o entendimento refinado de critérios complexos como a Coerência, o *BERTimbau* consolida-se como a abordagem mais robusta. Contudo, ao considerar os fatores de alta interpretabilidade, a capacidade de

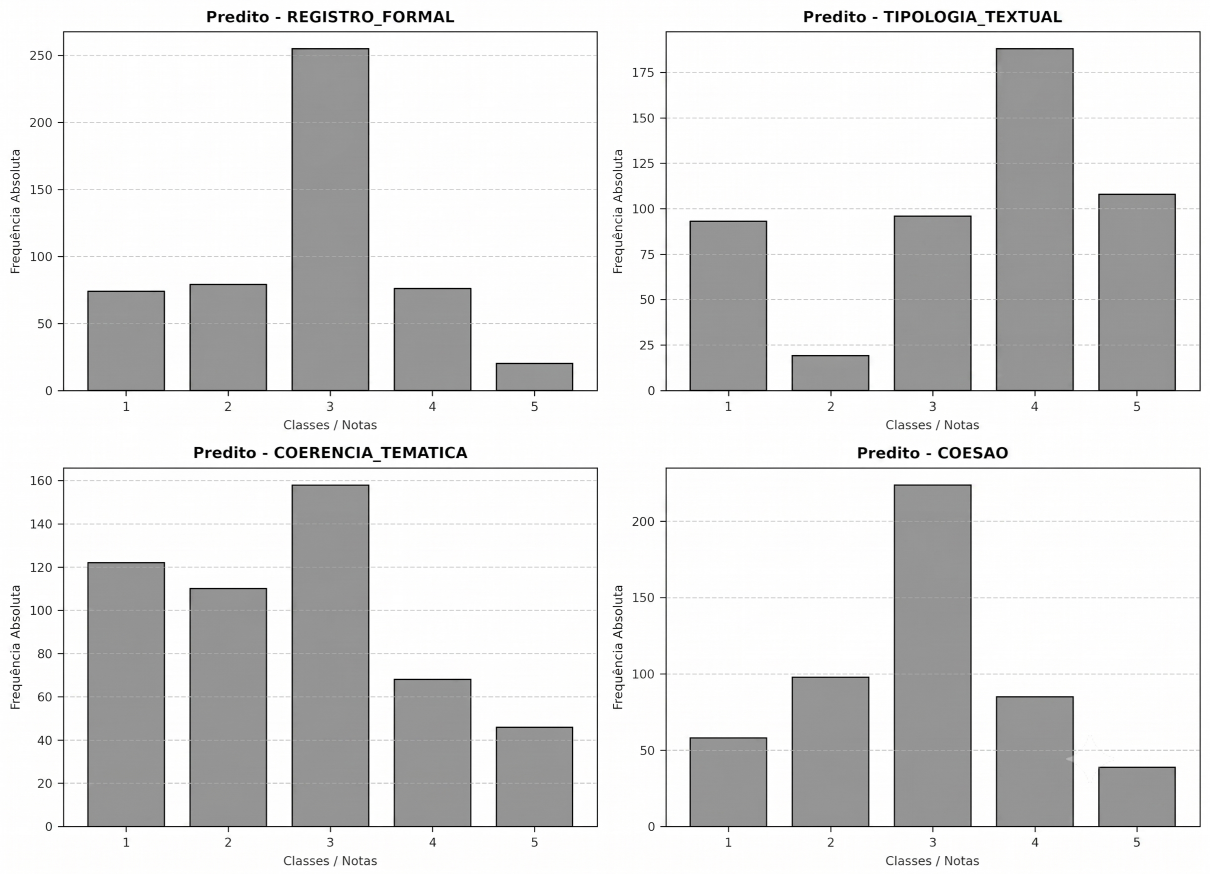

Figura 4.6: Distribuição das classes preditas obtidas pelo modelo XGBoost Classificador para os aspectos avaliados na escala de 1 a 5. Disposição dos gráficos: Registro Formal (superior esquerdo), Tipologia Textual (superior direito), Coerência Temática (inferior esquerdo) e Coesão (inferior direito).

replicar a distribuição macroscópica das notas e o baixo custo computacional, o *XGBoost* afirma-se como a solução mais viável e equilibrada para o ambiente de produção em avaliações de larga escala. Dessa forma, as análises realizadas demonstram que ambos os paradigmas possuem relevância prática dentro do cenário de avaliação automática de redações.

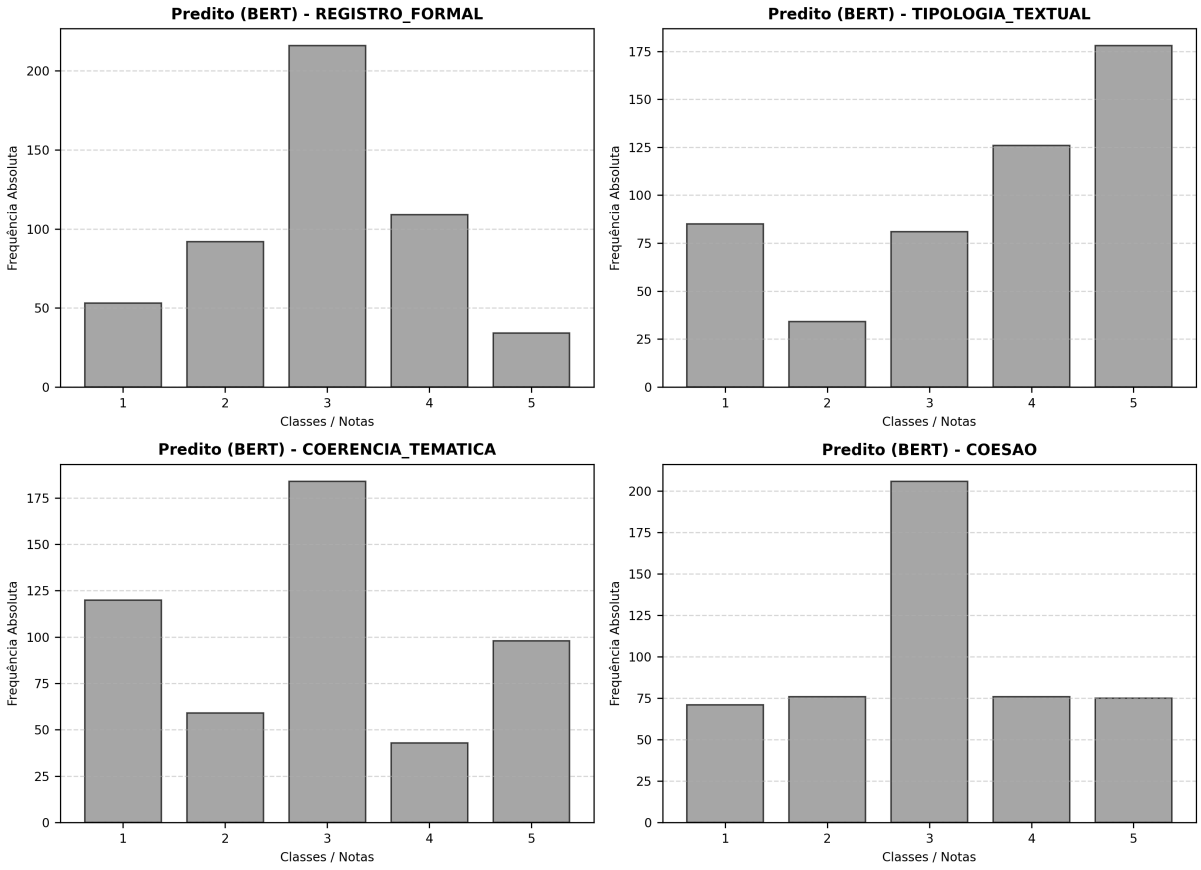

Figura 4.7: Distribuição das classes preditas obtidas pelo modelo BERTimbau para os aspectos avaliados na escala de 1 a 5. Disposição dos gráficos: Registro Formal (superior esquerdo), Tipologia Textual (superior direito), Coerência Temática (inferior esquerdo) e Coesão (inferior direito).

5 Conclusão

Conclui-se que o desenvolvimento deste trabalho permitiu alcançar a totalidade dos objetivos propostos inicialmente. Para além da validação acadêmica das arquiteturas de aprendizado de máquina, a investigação gerou diagnósticos fundamentais para subsidiar uma futura aplicação prática dessas tecnologias no contexto da Fundação CAEd. Os resultados obtidos estabelecem uma base sólida de evidências empíricas, servindo como ponto de partida para análises aprofundadas de viabilidade técnica e operacional na automação da avaliação de redações nos anos iniciais do Ensino Fundamental.

Apesar das contribuições alcançadas, este trabalho apresenta limitações técnicas e metodológicas que devem ser contextualizadas. Sob a perspectiva da generalização, a investigação delimitou-se a um público-alvo bastante específico: produções textuais do gênero narrativo escritas por alunos do 5º ano do Ensino Fundamental brasileiro. Trata-se de um recorte focado, que não reflete a totalidade de gêneros textuais, etapas de escolaridade ou variações linguísticas existentes. Por esse motivo, os modelos treinados não devem ser aplicados diretamente a outros contextos sem o devido reajuste.

Contudo, essas restrições de escopo não invalidam os resultados obtidos. O trabalho sustenta sua validade científica por ter seguido uma metodologia experimental rigorosa, tradicional e totalmente replicável, cobrindo desde a extração manual de características até o treinamento e a comparação direta das arquiteturas de aprendizado de máquina. Os experimentos reuniram evidências empíricas consistentes de que ambas as abordagens, clássica e profunda, são viáveis para o tratamento do problema proposto, consolidando este projeto como uma prova de conceito válida e um ponto de partida seguro para futuras expansões.

Como sugestões para trabalhos futuros a serem desenvolvidos por pesquisas subsequentes, recomenda-se, primeiramente, a expansão do escopo metodológico para outros públicos-alvo, aplicando os modelos em bases de dados de diferentes anos escolares e gêneros textuais. Além disso, aponta-se como um caminho promissor o desenvolvimento de modelos especialistas independentes para cada aspecto da grade de correção. Essa

abordagem permitiria criar pipelines de modelagem e customizações dedicadas a otimizar os resultados de cada critério individualmente, em vez de tratá-los de forma unificada.

Paralelamente, sugere-se investigar a viabilidade de arquiteturas híbridas que integrem a análise contextual profunda dos *Transformers* com a interpretabilidade de extratores especializados de características tabulares, buscando unir o melhor de ambos os paradigmas. Por fim, reforça-se a relevância de dedicar esforços ao refinamento da tokenização clássica, por meio da criação de analisadores morfológicos focados na linguagem infantil, mitigando o impacto dos desvios ortográficos na extração do modelo *baseline*. Com essas perspectivas, este trabalho encerra sua contribuição, mapeando caminhos técnicos claros para o avanço da avaliação automática de redações no cenário educacional brasileiro.

Bibliografia

AMORIM, E. C. F.; VELOSO, A. A multi-aspect analysis of automatic essay scoring for brazilian portuguese. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO (SBC). *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*. [S.l.], 2018. p. 11–20.

ATTALI, Y.; BURSTEIN, J. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, v. 4, n. 3, 2006.

CAMELO, R.; JUSTINO, S.; MELLO, R. F. Coh-metrix PT-BR: uma API web de análise textual para a educação. In: SBC. *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação (WCBIE 2020)*. [S.l.], 2020. p. 179–186.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. (KDD '16), p. 785–794. Disponível em: ⟨http://dx.doi.org/10.1145/2939672.2939785⟩.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

DOEWES, A.; KURDHI, N. A.; SAXENA, A. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In: INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY (IEDMS). *Proceedings of the 16th International Conference on Educational Data Mining*. [S.l.], 2023. p. 103–113.

FILHO, M. W. da S. et al. Automated formal register scoring of student narrative essays written in portuguese. In: SBC. *Anais do Congresso Brasileiro de Informática na Educação (CBIE 2023)*. [S.l.], 2023.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

KOCH, I. V. *A coesão textual*. São Paulo: Editora Contexto, 1989. 11–21 p.

KOCH, I. V.; TRAVAGLIA, L. C. *A Coerência Textual*. 8. ed. São Paulo: Editora Contexto, 1997.

LEAL, S. E. et al. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources & Evaluation*, 2023. Disponível em: ⟨https://doi.org/10.1007/s10579-023-09693-w⟩.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

MARINHO, J. C.; ANCHIÊTA, R. T.; MOURA, R. S. Essay-BR: a brazilian corpus of essays. In: SBC. *Anais do III Dataset Showcase Workshop (DSW)*. [S.l.], 2021. p. 53–64.

NGUYEN, H.; DERY, L. *Neural networks for automated essay grading*. [S.l.], 2016. CS224n Project Report.

OLIVEIRA, H. et al. Towards explainable prediction of essay cohesion in portuguese and english. In: ACM. *LAK23: 13th International Learning Analytics and Knowledge Conference*. [S.l.], 2023. p. 509–519.

OLIVEIRA, H. et al. A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in brazilian portuguese. *Data in Brief*, Elsevier, v. 60, p. 111526, 2025.

PAGE, E. B. The imminence of... grading essays by computer. *The Phi Delta Kappan*, Phi Delta Kappa International, v. 47, n. 5, p. 238–243, 1966.

ROSA, B. A. et al. Enhancing essay cohesion assessment: A novel item response theory approach. *Preprint*, 2025.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. In: *Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. [S.l.: s.n.], 2010.

SOARES, M. Letramento e alfabetização: as muitas facetas. *Revista Brasileira de Educação*, n. 25, p. 5–17, 2004.

SOARES, M. E. *A produção de textos na escola*. [S.l.]: Universidade Federal do Ceará, 1999.

SOUZA, F. C.; NOGUEIRA, R. F.; LOTUFO, R. A. BERTimbau: pretrained BERT models for brazilian portuguese. In: *Brazilian Conference on Intelligent Systems*. [S.l.]: Springer, 2020. p. 403–417.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. p. 5998–6008.