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

Detecção de Fraude Assistida por Modelos de Linguagem em Dissertações Argumentativas

Marcos Paulo Rodrigues da Silva

JUIZ DE FORA
JULHO, 2026

Detecção de Fraude Assistida por Modelos de Linguagem em Dissertações Argumentativas

Marcos Paulo Rodrigues da Silva

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computução

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

JULHO, 2026

Detecção de Fraude Assistida por Modelos de Linguagem em Dissertações Argumentativas

Marcos Paulo Rodrigues da Silva

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 Infomática - PUC-Rio

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação - UFRJ

Marcelo de Oliveira Costa Machado
Doutor em Informática - UNIRIO

JUIZ DE FORA

6 DE JULHO, 2026

Resumo

A popularização dos grandes modelos de linguagem introduziu desafios sem precedentes para a integridade acadêmica, deslocando o foco do plágio tradicional para a geração sintética de textos. Em ambientes de avaliação supervisionados, esse cenário foi agravado pelo surgimento do *copy-typing*, um comportamento dissimulado no qual estudantes transcrevem manualmente textos gerados por inteligência artificial, inserindo ruídos, anomalias e edições intencionais com o objetivo de mascarar sua origem automática e dificultar a detecção por sistemas convencionais de identificação de plágio e autoria sintética. Diante dessa problemática, este trabalho tem como objetivo desenvolver uma solução computacional capaz de detectar fraudes em textos dissertativo-argumentativos e modelar padrões híbridos de autoria textual.

Neste contexto, este estudo propõe uma abordagem computacional para a detecção de fraudes educacionais baseada em um problema de classificação multiclasse, capaz de distinguir três categorias distintas de autoria: textos produzidos por seres humanos, textos gerados integralmente por inteligência artificial e textos resultantes de *copy-typing*. Para isso, foi desenvolvido um *pipeline* de processamento de linguagem natural baseado no modelo pré-treinado BERTimbau. Adicionalmente, investigou-se uma arquitetura híbrida baseada em *late fusion*, que incorpora características estilométricas extraídas *a priori* dos textos.

Os experimentos foram conduzidos em dois *corpora* distintos possibilitando avaliar o impacto da variabilidade dos dados na capacidade de generalização dos modelos. Os resultados demonstraram que o aumento da diversidade amostral produziu ganhos substanciais de desempenho. Além das métricas tradicionais de classificação, foram empregadas técnicas de calibração probabilística, ajuste de limiares de decisão e análise de incerteza, permitindo avaliar o comportamento probabilístico das predições e identificar cenários de elevada ambiguidade.

Por fim, foram incorporadas técnicas de inteligência artificial explicável, utilizando SHAP, LIME e *Integrated Gradients* para interpretar o processo de decisão e inves-

tigar os padrões linguísticos mais relevantes para a classificação. Os resultados indicam que a abordagem proposta constitui uma alternativa promissora para a identificação de conteúdos produzidos ou assistidos por inteligência artificial, contribuindo para o desenvolvimento de ferramentas de auditoria acadêmica mais transparentes, interpretáveis e alinhadas às novas demandas impostas pela inteligência artificial generativa.

Palavras-chave: Inteligência artificial, processamento de linguagem natural, detecção de fraude, *copy-typing*, BERTimbau, fusão tardia, inteligência artificial explicável.

Abstract

The increasing adoption of large language models has introduced unprecedented challenges to academic integrity, shifting the focus from traditional plagiarism to synthetic text generation. In supervised assessment environments, this scenario has been exacerbated by the emergence of *copy-typing*, a covert behavior in which students manually transcribe texts generated by artificial intelligence, inserting noise, anomalies, and intentional edits to mask their automated origin and hinder detection by conventional plagiarism detection systems and methods for identifying AI-generated content. In light of this problem, this work aims to develop a computational solution capable of detecting fraud in argumentative essays and modeling the hybrid authorship patterns exhibited by students.

In this context, this research proposes a computational approach for detecting educational fraud based on a multiclass classification problem, capable of distinguishing three distinct categories of authorship: human-written texts, AI-generated texts, and *copy-typed* texts. To this end, a natural language processing *pipeline* was developed based on the pre-trained BERTimbau model. Additionally, a hybrid architecture based on *late fusion* was investigated, which incorporates stylometric features extracted *a priori* from the texts.

The experiments were conducted on two distinct *corpora* enabling the evaluation of the impact of data variability on the models' generalization ability. The results demonstrated that increasing sample diversity led to substantial performance gains. In addition to traditional classification metrics, probabilistic calibration techniques, decision threshold tuning, and uncertainty analysis were employed, allowing the evaluation of the probabilistic behavior of model predictions and the identification of highly ambiguous scenarios.

Finally, explainable artificial intelligence techniques were incorporated, using SHAP, LIME, and *Integrated Gradients* to interpret the decision-making process and investigate the linguistic patterns most relevant to classification. The results indicate

that the proposed approach constitutes a promising alternative for identifying content generated entirely or partially by AI, contributing to the development of academic auditing tools that are more transparent, interpretable, and aligned with the new demands imposed by generative artificial intelligence.

Keywords: Artificial intelligence, natural language processing, fraud detection, copy-typing, BERTimbau, late fusion, explainable artificial intelligence.

Agradecimentos

Agradeço a Deus, por me guiar ao longo de toda esta trajetória.

Agradeço à minha família, em especial meus pais Rosely e Jalmir e meu irmão Marcus Vinícius, pelo apoio e amor incondicional ao longo desses anos.

Agradeço aos meus amigos pelo suporte nas dificuldades encontradas.

Por fim, agradeço aos professores do Departamento de Ciência da Computação pelos seus ensinamentos, em especial ao orientador Jairo Francisco de Souza pelas valiosas contribuições e por auxiliar meu crescimento acadêmico.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

API	*Application Programming Interface*
BERT	*Bidirectional Encoder Representations from Transformers*
brWaC	*Brazilian Web as Corpus*
ENEM	Exame Nacional do Ensino Médio
GPT	*Generative Pre-trained Transformer*
IA	Inteligência Artificial
IDE	Ambiente de Desenvolvimento Integrado
IG	*Integrated Gradients*
LGPD	Lei Geral de Proteção de Dados
LIME	*Local Interpretable Model-Agnostic Explanations*
LLM	*Large Language Models*
MTLD	*Measure of Textual Lexical Diversity*
PLN	Processamento de Linguagem Natural
RGPD	Regulamento Geral sobre a Proteção de Dados
SHAP	*SHapley Additive exPlanations*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TTR	*Type-Token Ratio*
XAI	*eXplainable Artificial Intelligence*

1 Introdução

A avaliação da aprendizagem em ambientes virtuais tem enfrentado historicamente o desafio de garantir a integridade acadêmica, com foco predominante na detecção de plágio tradicional, no qual a cópia entre pares é identificada por similaridade textual. No entanto, a rápida popularização dos Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*), como o ChatGPT, introduziu um novo paradigma. Para os estudantes, essas ferramentas podem atuar como facilitadoras ao auxiliar na superação de bloqueios criativos, sintetizar vastos volumes de pesquisa em resumos concisos e apoiar a preparação para avaliações por meio da geração de exames práticos e *feedbacks* personalizados (KASNECI et al., 2023). Para os docentes, a inteligência artificial oferece suporte na elaboração de planos de aula, materiais didáticos e na provisão de acompanhamento guiado por meio de orientações, *feedbacks* e intervenções em tempo real, sendo um recurso promissor para o gerenciamento de turmas numerosas (REGE; YARMOLUK, 2023). Apesar desses benefícios, essa mesma tecnologia impõe desafios sem precedentes. Por possuírem capacidades notáveis na geração de textos coerentes e de alta qualidade, os LLMs permitem que estudantes completem tarefas e exames *online* de forma totalmente automatizada. Tal fato levanta preocupações sobre a validade dos instrumentos de avaliação, a integridade acadêmica e o real desenvolvimento do pensamento crítico e das habilidades de resolução de problemas dos alunos (RAHMAN; WATANOBE, 2023; LO, 2023).

O uso indevido dessas tecnologias acarreta consequências severas que ameaçam os pilares do ecossistema educacional. Esse impacto pode ser observado em três frentes principais:

- A degradação da qualidade da educação: Ao gerar respostas ou soluções completas simplesmente enviando *prompts* para uma inteligência artificial (IA) em vez de estudar conceitos, compreender os problemas, e obter conhecimento através de seus próprios erros e acertos, os estudantes perdem oportunidades valiosas de aprendizado. A abstenção do esforço cognitivo impede o desenvolvimento do pensamento

crítico e da compreensão genuína da disciplina, dificultando a aplicação dos conceitos em novos cenários (LO, 2023);

- Vantagem injusta: alunos que utilizam essas ferramentas concluem tarefas rapidamente e com esforço mínimo, gerando um ambiente de aprendizado desequilibrado que desmoraliza os estudantes comprometidos com a integridade acadêmica e corrói os princípios de equidade (COTTON; COTTON; SHIPWAY, 2024);

- Dano à integridade das instituições: quando as avaliações deixam de refletir as reais habilidades dos alunos, a credibilidade de todo o processo de certificação é colocada em xeque. Essa erosão afeta não apenas a reputação da instituição, mas também as perspectivas futuras dos estudantes na busca por emprego ou ingresso em programas de pós-graduação (COTTON; COTTON; SHIPWAY, 2024).

Diferente do plágio convencional, o texto gerado por IA é muitas vezes único e semanticamente coerente, tornando ineficazes as ferramentas tradicionais de comparação de respostas. Em resposta, a comunidade acadêmica e a indústria desenvolveram classificadores treinados para distinguir texto humano de texto sintético. Contudo, muitos desses sistemas abordam o problema de maneira estritamente binária (Humano *versus* IA) e são restritos a um conjunto fixo de modelos, falhando frequentemente em generalizar para novos LLMs que surgem continuamente (WANG; LI, 2025).

Mais criticamente, a abordagem binária ignora o comportamento real e dissimulado dos estudantes durante a fraude. Em ambientes virtuais de exames monitorados, onde funções que permitem copiar o texto diretamente estão desabilitadas, os alunos recorrem à digitação manual das respostas geradas pela IA, um processo conhecido como *copy-typing* (NIU et al., 2024). Durante essa transcrição, ocorrem modificações intencionais ou acidentais, como a introdução de erros de digitação, omissões, substituição de sinônimos e edições sintáticas. Esse comportamento híbrido degrada os sinais estatísticos do texto sintético, reduzindo drasticamente a eficácia dos detectores convencionais, o que motiva a transição de um modelo binário para uma classificação de múltiplas classes, visando capturar a real complexidade do fenômeno

A problemática se agrava em itens de resposta construída (perguntas abertas),

nos quais a variabilidade das respostas aceitáveis é ampla, e, especialmente, redações. A literatura recente aponta que o desempenho dos detectores é sensível a características intrínsecas do texto, tais como o comprimento do documento, o domínio textual, as estruturas linguísticas empregadas e o grau de edição ou parafraseamento (WEBER-WULFF et al., 2023; CHAKRABORTY et al., 2023). Em particular, o tamanho da resposta afeta diretamente a capacidade preditiva do modelo, pois textos muito curtos não fornecem evidências estatísticas suficientes para distinguir, de forma confiável, padrões de escrita humana e de geração automática, aumentando a probabilidade de erros de classificação. Em contrapartida, em textos mais extensos, os padrões probabilísticos se estabilizam, tornando as anomalias detectáveis. Adicionalmente, fatores contextuais do estudante, como a idade e o nível de proficiência, influenciam a complexidade linguística esperada. Um adolescente, por exemplo, tende a utilizar uma linguagem mais simples do que a estrutura sofisticada gerada por um LLM, criando um contraste que pode ser explorado.

1.1 Justificativa

O desenvolvimento de novas abordagens para a detecção de fraude justifica-se pela necessidade urgente de resguardar a validade dos instrumentos de avaliação educacional e garantir que o desenvolvimento do pensamento crítico dos alunos não seja prejudicado pelo uso indevido de tecnologias gerativas (RAHMAN; WATANOBE, 2023; LO, 2023). Apesar de existirem classificadores no mercado, a literatura e a prática demonstram que a abordagem estritamente binária (Humano *versus* IA) tornou-se insuficiente frente à constante evolução dos modelos de linguagem e às limitações das ferramentas de detecção comerciais (WEBER-WULFF et al., 2023; WANG; LI, 2025).

O principal motivador técnico deste trabalho é a constatação de que os alunos adotaram táticas adversárias dissimuladas para burlar esses sistemas, destacando-se o processo de *copy-typing*. Esse comportamento híbrido ofusca os padrões estatísticos procurados pelos detectores convencionais, limitando a capacidade de distinção do modelo e gerando alta taxa de falsos negativos (CHAKRABORTY et al., 2023). Logo, expandir o escopo de detecção para um modelo multiclasse é uma evolução necessária para espelhar o comportamento real do usuário e mitigar as vulnerabilidades dos sistemas atuais.

Além da urgência técnica, a motivação desta pesquisa estende-se à esfera ética. A aplicação de ferramentas de detecção em cenários acadêmicos pode resultar em punições severas, reprovações e danos irreparáveis à trajetória do estudante. Acusações de desonestidade baseadas puramente na probabilidade de um algoritmo são consideradas problemáticas e eticamente inaceitáveis (XIE; WU; CHAKRAVARTY, 2023). Sendo assim, este trabalho também é motivado pela premissa de aliar precisão à explicabilidade algorítmica, buscando garantir que o modelo forneça evidências interpretáveis que sustentem a sua classificação e promovam um processo de auditoria justo, transparente e passível de defesa.

1.2 Objetivos

Este trabalho propõe o desenvolvimento de modelos computacionais para a detecção de fraudes textuais em exames *online*, visando mitigar o impacto do uso indevido de IA generativa na educação. O objetivo central é superar a abordagem binária convencional por meio de uma classificação multiclasse, que categorize as respostas em: Humano, Inteligência Artificial e Copy-typing. Com essa expansão, busca-se modelar com maior fidelidade o comportamento dissimulado dos estudantes, garantindo maior confiabilidade e explicabilidade às ferramentas de auditoria acadêmica.

Os objetivos específicos deste trabalho são:

- Construir e estruturar um conjunto de dados composto por textos humanos, sintéticos e híbridos (*copy-typing*), garantindo a representatividade necessária para a extração de características linguísticas discriminativas;
- Avaliar o desempenho de classificadores baseados na arquitetura *transformer* na distinção e classificação correta das três categorias textuais propostas;
- Analisar o impacto de variáveis intrínsecas e contextuais, como o tamanho das respostas (extensão em *tokens*) e a complexidade linguística, na acurácia das predições;
- Implementar e validar mecanismos de explicabilidade algorítmica capazes de extrair e apresentar as evidências linguísticas que fundamentam a classificação do modelo.

2 Fundamentação Teórica

O desenvolvimento de uma solução computacional capaz de identificar comportamentos híbridos e dissimulados, como o *copy-typing*, exige uma fundamentação teórica multidisciplinar. A compreensão destes conceitos é essencial não apenas para justificar as decisões arquiteturais e os métodos de implementação adotados, mas também para situar este trabalho na intersecção entre a linguística computacional, o comportamento cognitivo e o aprendizado de máquina. Neste capítulo, estabelece-se o alicerce da pesquisa delineando os princípios técnicos que a norteiam. Inicialmente, o paradigma contemporâneo da fraude assistida é explorado sob a ótica da cognição humana. Em seguida, abordam-se os avanços em Processamento de Linguagem Natural (PLN) e as métricas de estilometria, ferramentas vitais para a extração da semântica sintética e das anomalias biomecânicas do texto. Para a integração estrutural desses sinais contrastantes, discute-se o conceito de arquitetura *late fusion*. Por fim, os compromissos com a transparência e a confiabilidade são consolidados através da exploração de técnicas de inteligência artificial explicável, bem como dos métodos de calibração e análise de incerteza, assegurando que o modelo final seja tecnicamente robusto e eticamente justificável.

2.1 O Paradigma da Fraude Assistida

A pesquisa tradicional na área de detecção de textos gerados por inteligência artificial tem focado predominantemente na identificação de produções exclusivamente sintéticas, um cenário analítico conhecido como *zero-shot generation*. Nesse contexto, o modelo gerativo, como ChatGPT e Gemini, produz o texto e o utilizador atua como um intermediário passivo, simplesmente copiando e colando o conteúdo em seu destino final. No entanto, a implementação de ambientes de exames *online* supervisionados, aliados a shannon1948 de avaliação com bloqueio de área de transferência e outras técnicas antiplágio, forçou uma evolução no comportamento dissimulado dos estudantes, dando origem a um fenômeno dinâmico e híbrido: o *copy-typing*.

Esse conceito descreve a ação de um estudante que gera a resposta através de uma IA em um dispositivo secundário (como um smartphone ou monitor auxiliar) e a transcreve manualmente para o sistema oficial de avaliação. Diferente do plágio direto, esse comportamento impõe um elevado custo cognitivo ao estudante. Conforme os fundamentos da Teoria da Carga Cognitiva e os modelos de processamento da escrita (SWELLER, 1988), quando um indivíduo tenta ler, reter temporariamente na memória de trabalho e digitar um texto cujas estruturas sintáticas e vocabulário estão acima do seu nível de proficiência habitual, ocorre uma sobrecarga mental. Sob a pressão de tempo característica de exames, o ser humano é incapaz de transcrever essa estrutura sofisticada de forma impecável, introduzindo inevitavelmente anomalias de transcrição na escrita, que se manifestam através de erros de digitação, omissões de pontuação e divergências na capitalização de letras.

Para além das falhas subconscientes oriundas da sobrecarga cognitiva, o *copy-typing* é também fortemente marcado por modificações intencionais. Na tentativa de adequar a resposta gerada ao seu próprio perfil linguístico, reduzir traços de artificialidade do texto, ou simplesmente burlar ativamente os sistemas de detecção, o estudante realiza intervenções ativas durante a transcrição. Estas incluem a substituição de termos sofisticados por sinônimos mais cotidianos, a simplificação geral do vocabulário, o uso de conectivos mais básicos e familiares, até chegar a paráfrases profundas, nas quais trechos inteiros são reescritos utilizando a estrutura sintática natural do aluno.

Essa sobreposição de características subconscientes e edições intencionais posiciona o *copy-typing* em uma área não bem delimitada no espectro da autoria textual. Em vez de uma divisão binária estrita e clara, estabelece-se um espectro contínuo entre o texto estritamente gerado por IA e o texto de autoria genuína humana. Dentro desse espectro nebuloso, emergem duas linhas tênues que desafiam a classificação: de um lado, a fronteira sutil que separa um texto de *copy-typing* leve (com alterações mínimas e superficiais) de um texto exclusivamente sintético; do outro, a fronteira entre um *copy-typing* profundo (marcado por intensas reescritas e paráfrases estruturais) e a escrita puramente humana.

O grande desafio tecnológico imposto pelo *copy-typing* reside justamente na sua capacidade de camuflagem dentro dessas zonas fronteiriças. Nesses cenários, a semântica

de base do texto permanece artificial, mantendo parte da previsibilidade estatística da máquina, mas a superfície textual adquire imperfeições e nuances tipicamente humanas. Para mitigar essa lacuna, a detecção moderna não pode mais buscar apenas a perfeição da máquina ou a variabilidade humana isoladamente; torna-se imprescindível aprimorar os métodos e detectores tradicionais para que sejam capazes de mapear matematicamente a interseção entre a estrutura semântica sintética e a intervenção biomecânica humana.

2.1.1 Estilometria e Linguística Computacional

Para capturar quantitativamente a resistência cognitiva, seja ela gerada por anomalias de transcrição ou edições intencionais, e complementar a análise semântica profunda, a pesquisa recorre à estilometria. Esta é a disciplina da linguística computacional responsável pela análise estatística e mensuração do estilo literário. Enquanto modelos baseados em *transformers* avaliam primariamente o contexto e a coerência global da resposta, a estilometria extrai propriedades intrínsecas da estrutura e da forma do texto, funcionando como uma impressão digital matemática do processo de escrita.

O comportamento dissimulado de *copy-typing* pode ser parametrizado teoricamente através da observação de divergências e anomalias estatísticas em eixos fundamentais da escrita, como a diversidade lexical, o ritmo de construção estrutural e o rigor ortográfico. Textos gerados por modelos de linguagem operam sob um paradigma de otimização de probabilidade condicional, o que invariavelmente os leva a adotar um vocabulário convergente, estruturas frasais de cadência uniforme e um rigor gramatical impecável.

Em contrapartida, o processo cognitivo humano, especialmente sob pressão avaliativa, é inerentemente não linear e caótico. Autores humanos tendem a escrever em fluxos intermitentes, alternando construções curtas e diretas com sentenças complexas, além de apresentarem maior variabilidade vocabular e propensão ao erro mecânico. A estilometria fornece as ferramentas teóricas para mensurar essas discrepâncias de forma isolada do significado das palavras. A combinação dessas avaliações estruturais permite que um sistema quantifique o nível de artificialidade da forma de escrita e posicione o documento analisado nas zonas cinzentas do espectro de autoria, fornecendo a base teórica

indispensável para a diferenciação de classes textuais híbridas.

2.2 Processamento de Linguagem Natural

A evolução do PLN redefiniu a capacidade computacional de extrair semântica profunda de textos não estruturados. No cerne dessa revolução está a arquitetura *transformer* (VASWANI et al., 2017), que substituiu a dependência sequencial e as limitações de memória das redes neurais recorrentes pelos mecanismos de autoatenção (*self-attention*). Essa inovação permite o processamento paralelo e a ponderação de dependências de longo alcance na escrita, elementos cruciais para a análise do encadeamento lógico em redações acadêmicas e itens de resposta construída.

A arquitetura original concebida por Vaswani e seus colaboradores era estruturada em dois componentes principais: uma pilha de codificadores (*encoders*) e uma pilha de decodificadores (*decoders*). A partir dessa fundação, a pesquisa em modelos de linguagem ramificou-se em vertentes especializadas. De um lado, modelos autorregressivos, como a família GPT (BROWN et al., 2020), derivaram exclusivamente do bloco decodificador, processando o texto de forma unidirecional prevendo o valor futuro de uma variável com base em seus próprios valores passados com o objetivo principal focado na geração textual.

Em contrapartida, o modelo BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2018) isolou e expandiu a pilha de codificadores da arquitetura original, abdicando da capacidade generativa em favor de uma máxima compreensão estrutural. Ao analisar o texto de forma bidirecional, o BERT captura o contexto de uma palavra considerando simultaneamente todos os termos à sua esquerda e à sua direita. Em tarefas de classificação textual, essa característica permite que cada *token* seja representado considerando simultaneamente o contexto do entorno, produzindo representações semânticas profundas adequadas para tarefas de compreensão textual.

Enquanto métodos clássicos de representação vetorial, como TF-IDF ou Word2Vec, falham ao tentar capturar a polissemia e a complexidade informacional de um parágrafo ou texto inteiro, tratando as palavras de forma isolada ou com pouco contexto, o BERT transforma a redação em um espaço vetorial denso e dinâmico de 768 dimensões, conhecido como *latent space*. Para obter uma representação única e consolidada do documento,

utiliza-se a técnica de *masked mean pooling*. Em vez de depender exclusivamente do *token* de classificação global (`[CLS]`), essa técnica calcula a média dos *embeddings* de todos os *tokens* válidos da sequência, aplicando uma máscara restritiva para ignorar *tokens* de preenchimento (*padding*). O resultado é um vetor contínuo que condensa toda a informação contida no texto, incluindo a coerência semântica, oferecendo uma base matemática sólida para distinguir o rigor de uma inteligência artificial da flutuação cognitiva da escrita humana.

2.2.1 Escolha do Modelo Base

A seleção do modelo de linguagem fundacional para a tarefa de detecção de fraudes textuais exige uma análise criteriosa das opções disponíveis no estado da arte do PLN. O ecossistema atual de modelos baseados em *transformers* oferece diversas arquiteturas, cada qual otimizada para propósitos específicos. Contudo, a determinação do modelo ideal para este estudo exigiu o alinhamento de três pilares fundamentais: a adequação arquitetural à tarefa (classificação de redações), a aderência estrita ao domínio linguístico (português brasileiro) e a capacidade de representação profunda.

Neste cenário de alternativas, modelos da família GPT, por exemplo, por serem projetados essencialmente para a geração autorregressiva de texto, destinam parte significativa de sua capacidade computacional à predição sequencial de *tokens*. Essa característica introduz uma sobrecarga arquitetural e computacional que não é essencial nem vantajosa para o problema de classificação de sequências inteiras abordado neste trabalho.

Por outro lado, modelos que evoluíram a arquitetura codificadora, como o RoBERTa (LIU et al., 2019) e o DeBERTa (HE et al., 2020), alcançaram o estado da arte em diversos *benchmarks*. O RoBERTa introduziu um pré-treinamento mais longo e removeu a tarefa de previsão da próxima frase, enquanto o DeBERTa propôs mecanismos de atenção desacoplada (*disentangled attention*) e codificações posicionais aprimoradas. Entretanto, as versões originais e mais robustas desses modelos foram desenvolvidas primariamente para a língua inglesa. A disponibilidade de variantes dessas arquiteturas treinadas massiva e nativamente para o português brasileiro ainda é restrita, o que as torna menos

atrativas para o contexto nacional desta pesquisa.

Adicionalmente, a arquitetura ALBERT (LAN et al., 2019) buscou otimizar o custo computacional reduzindo o número de parâmetros por meio do compartilhamento de pesos entre camadas e da fatoração dos *embeddings*. Embora essa estratégia produza modelos mais leves e eficientes e possua uma implementação pré-treinado para o português chamada Albertina (RODRIGUES et al., 2023), a redução drástica de parâmetros atua como um fator limitante na capacidade representacional. Em tarefas complexas que exigem a modelagem de relações discursivas sutis e a identificação de anomalias sintáticas sutis geradas pelo *copy-typing*, essa compressão pode comprometer o desempenho final.

Diante das limitações apresentadas pelas arquiteturas concorrentes, ganha destaque o modelo brasileiro BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), o qual foi adotado como modelo base. O BERTimbau é uma adaptação da arquitetura *encoder-only* do BERT pré-treinada especificamente para a língua portuguesa. Diferentemente de modelos multilíngues que diluem sua capacidade representacional entre vários idiomas, o BERTimbau foi desenvolvido utilizando grandes *corpora* textuais do Brasil, como o brWaC. Isso permite que suas representações vetoriais capturem de forma muito mais precisa as características morfológicas, a sintaxe estrutural e a semântica idiomática local. Tal especialização é particularmente crítica para a avaliação de respostas abertas e redações, onde aspectos como coesão, coerência, argumentação e o domínio da norma padrão desempenham papel central na distinção de autoria.

Por fim, outro fator decisivo para a sua escolha é a ampla validação do BERTimbau na literatura nacional de PLN. O modelo possui um ecossistema consolidado, sendo frequentemente utilizado em tarefas de classificação de textos, análise de sentimentos e reconhecimento de entidades nomeadas. Dessa forma, a escolha do BERTimbau justifica-se plenamente por oferecer a especialização linguística exigida, a arquitetura ideal para classificação profunda e o equilíbrio ótimo entre desempenho acadêmico comprovado e custo computacional.

2.2.2 Técnica de Fusão Tardia (*Late Fusion*)

Na área de aprendizado de máquina multimodal, a integração de fontes de dados de naturezas distintas configura um desafio estrutural significativo. Em cenários de classificação que exigem a análise simultânea de informações não estruturadas (como a semântica de um texto) e dados tabulares estruturados (como métricas estatísticas e características estilométricas), a forma como esses dados são combinados afeta diretamente a capacidade de generalização e a robustez do modelo.

A literatura categoriza os métodos de integração multimodais predominantemente em duas abordagens: a fusão precoce (*early fusion*) e a fusão tardia (*late fusion*). Tentar incorporar métricas numéricas absolutas diretamente na entrada de um modelo de linguagem, caracterizando uma fusão precoce, frequentemente resulta na degradação do espaço representacional. Essa abordagem força o algoritmo a processar contagens estatísticas isoladas sob a mesma lógica matemática empregada para representações semânticas complexas e sequenciais, introduzindo ruído e instabilidade no treinamento.

A teoria da fusão tardia resolve esse impasse metodológico isolando estrategicamente os domínios de extração de características. Nesse paradigma arquitetural, as diferentes modalidades de dados são processadas de forma independente em fluxos paralelos. Cada fluxo utiliza um extrator de características especializado no seu respectivo domínio de dados, otimizando a informação sem interferência externa.

É apenas após essa etapa individual de processamento, quando os dados brutos de ambas as fontes já foram comprimidos em representações matemáticas densas e de alto nível em seus respectivos espaços latentes, que ocorre a integração. A concatenação dos vetores resultantes é realizada unicamente nas camadas profundas que antecedem a função de decisão do modelo (*classification head*). Ao adiar a união das informações, a fusão tardia permite que a rede neural desenvolva a capacidade de ponderar evidências de naturezas conflitantes de maneira hierárquica. Em contextos de detecção de anomalias, isso confere ao classificador a habilidade de cruzar o contexto semântico com o perfil estatístico de forma isolada.

2.3 Inteligência Artificial Explicável

O desafio fundamental inerente aos modelos baseados em aprendizado profundo é a sua natureza opaca. Em cenários de alto risco e contextos punitivos, como na esfera educacional, onde sanções por fraude carregam um peso ético e acadêmico substancial, legislações de proteção de dados (como o RGPD na Europa e a LGPD no Brasil), aliadas às diretrizes de justiça algorítmica, preconizam a transparência e da auditabilidade das decisões algorítmicas. Para interpretar o processo de decisão das redes neurais e mitigar essa opacidade, a área de Inteligência Artificial Explicável (XAI, do inglês *eXplainable Artificial Intelligence*) propõe diversas abordagens matemáticas para a atribuição de relevância, com destaque para as seguintes técnicas:

Entre as técnicas de maior destaque, encontra-se o SHAP (*SHapley Additive exPlanations*). Fundamentado na Teoria dos Jogos Cooperativos de Lloyd Shapley, o SHAP calcula a contribuição marginal de cada característica de entrada (*feature*) como se fosse um jogador colaborando para o resultado final de uma partida (a predição do modelo). É um método de atribuição reconhecido por possuir uma base matemática sólida que garante propriedades axiomáticas essenciais para a justiça algorítmica, como a precisão local e a consistência na distribuição dos pesos decisórios (LUNDBERG; LEE, 2017).

Em complemento a essa abordagem, destaca-se o LIME (*Local Interpretable Model-Agnostic Explanations*), proposto por Ribeiro, Singh e Guestrin (2016). Essa é uma técnica agnóstica, ou seja, aplicável a qualquer arquitetura de rede, que busca explicar predições complexas aproximando-as de modelos interpretáveis mais simples, como regressões lineares, em uma vizinhança estritamente local. O método opera perturbando sistematicamente a amostra de entrada por meio da inserção de ruído ou ocultando partes da informação e observando como o modelo original altera sua predição. A partir dessas variações, gera-se um mapa analítico que reflete a sensibilidade do algoritmo às características daquela instância específica.

Por fim, o *Integrated Gradients* (IG), introduzido por Sundararajan, Taly e Yan (2017), atua como um método de explicabilidade intrínseco a redes neurais que opera analisando o fluxo de ativação do modelo. A técnica calcula a integral dos gradientes em relação aos dados de entrada, traçando o caminho matemático percorrido desde uma

linha de base neutra (uma entrada vazia ou de referência) até a amostra atual que gerou a predição. Essa integração resolve o problema de saturação de gradientes comum em redes profundas, conectando matematicamente as ativações das camadas ocultas aos atributos originais e visíveis da entrada.

3 Materiais e Métodos

Este capítulo descreve detalhadamente os procedimentos metodológicos, as escolhas arquiteturais e os recursos computacionais empregados no desenvolvimento de um modelo de detecção de fraude. Para garantir o rigor científico da pesquisa, abranger as complexidades do problema e atingir os objetivos propostos, a metodologia deste trabalho foi estruturada em um fluxo experimental dividido em três etapas principais:

1. **Construção do *dataset*:** A primeira etapa consistiu na curadoria e consolidação de um *corpus* textual robusto e representativo. Este conjunto de dados abrange textos autênticos (escritos integralmente por humanos), produções puramente sintéticas geradas de forma *zero-shot* por LLMs modernas e instâncias híbridas que simulam o *copy-typing*, fornecendo a base empírica necessária para o treinamento da rede neural.

2. **Treinamento de modelos classificadores:** A segunda etapa englobou a modelagem matemática e o treinamento computacional. Nela, detalha-se a implementação de uma arquitetura multimodal de fusão tardia (*late fusion*), responsável por extrair a semântica profunda do texto através do modelo base (BERTimbau) e integrá-la, em tempo de treinamento, ao vetor de características estatísticas extraído pela estilometria.

3. **Implementação de camadas de explicabilidade:** Por fim, a terceira etapa tratou da validação e interpretabilidade do modelo. Foram integradas técnicas de explicabilidade para mapear o processo decisório do classificador, garantindo que as predições pudessem ser auditadas e justificadas, visando conferir maior transparência à solução.

3.1 Ferramentas utilizadas

- **Google Sheets:** Utilizado como a plataforma principal para a organização estruturada e rotulagem primária do *corpus* de redações.

- **Google Colab**: Plataforma empregada para a execução de experimentos e prototipagem do projeto. Sua utilização foi fundamental no processo iterativo de desenvolvimento, especialmente na comparação entre arquiteturas e na tomada de decisões relacionadas ao desenho final do modelo proposto.

- **Ollama**[1]**:** Utilizado como motor de inferência para execução de grandes modelos de linguagem em ambiente controlado. Foi a ferramenta chave para a geração do conjunto de dados puramente sintéticos e híbridos (*copy-typing*).

- **Visual Studio Code:** Adotado como o ambiente de desenvolvimento integrado (IDE) principal do projeto.

- **Git e GitHub:** Ferramenta para controle de versão e preservação do histórico de desenvolvimento com a criação de um repositório[2].

- **Python 3.14.5:** A linguagem de programação base do projeto, escolhida pela vasta aplicação no ecossistema atual de inteligência artificial e ciência de dados.

- **Pandas:** Utilizado para a manipulação dos dados tabulares (arquivos CSV). Foi a espinha dorsal para a leitura, filtragem e reestruturação do *dataset* original.

- **NumPy:** A biblioteca padrão para computação científica em Python. Essencial para operações matriciais rápidas e cálculos matemáticos complexos.

- **Scikit-Learn:** Ferramenta de *machine learning* utilizada em vários momentos no projeto desde a divisão estatisticamente segura da base de dado até a geração de métricas de avaliação

- **Re (Expressões Regulares):** Biblioteca nativa do Python utilizada para o processamento e a manipulação de padrões textuais por meio de expressões regulares,

[1]https://ollama.com/
[2]https://github.com/PapauloMP/fraud-detection

utilizada para normalização, limpeza textual e extração de *features*.

- **PyTorch:** O *framework* central de construção do modelo neural. Foi escolhido pela sua capacidade de construir arquiteturas dinâmicas (como a arquitetura *late fusion*), gestão eficiente de tensores na GPU (*cuda*), e pela sua API de redes neurais que permitiu a criação de MLPs e funções de calibração.

- **Hugging Face Transformers:** Utilizada para instanciar o tokenizador e o *encoder* base do modelo BERT.

- **BERTimbau[3]:** O modelo de linguagem pré-treinado específico para a língua portuguesa, atuando como o motor principal de extração de características semânticas do projeto.

- **Matplotlib e Seaborn:** Utilizadas para a plotagem gráfica do histórico de treinamento e para a criação visualmente intuitiva da matriz de confusão do modelo, permitindo a análise do desempenho da classificação.

- **SHAP[4]:** Utilizado para calcular o impacto exato (positivo ou negativo) de cada *token* na probabilidade final de o texto ser uma fraude, fornecendo provas visuais exatas da decisão da máquina.

- **LIME[5]:** Utilizada como técnica complementar de explicabilidade, criando modelos lineares simplificados para justificar as predições de instâncias específicas.

- **Captum[6]:** Biblioteca do ecossistema PyTorch utilizada para análises profundas de interpretabilidade, rastreando matematicamente o caminho dos gradientes desde a previsão final até às palavras de entrada.

3.2 Construção do *Dataset*

A qualidade e a representatividade dos dados de treinamento são fatores determinantes para a capacidade de generalização de modelos de aprendizado de máquina, especialmente

[3]https://huggingface.co/neuralmind/bert-base-portuguese-cased
[4]https://shap.readthedocs.io/en/latest/
[5]https://lime-ml.readthedocs.io/en/latest/
[6]https://captum.ai/

em tarefas de classificação multimodal. Para este estudo, o domínio textual selecionado foi o de redações no formato exigido pelo Exame Nacional do Ensino Médio (ENEM). A escolha desse gênero dissertativo-argumentativo justifica-se, primeiramente, pela sua estrutura rígida e pela exigência do domínio da norma culta, características que estabelecem uma linha de base linguística consistente para a mensuração de anomalias estilométricas.

Adicionalmente, a distribuição do tamanho dessas redações adequa-se de forma otimizada ao limite arquitetural de entrada do modelo BERTimbau, que é restrito ao processamento de 512 *tokens*. Conforme ilustrado na distribuição apresentada na Figura 3.1, o *corpus* apresenta uma média de 416,4 *tokens* e uma mediana de 423 *tokens*. Essa extensão representa o ponto de equilíbrio ideal para a experimentação metodológica: o texto é longo o suficiente para fornecer o volume de dados necessário à captura de sinais estatísticos e estilométricos robustos na distinção de autoria (WEBER-WULFF et al., 2023; CHAKRABORTY et al., 2023) ao mesmo tempo em que respeita a limitação de dados de entrada do modelo. Embora o modelo comporte integralmente a grande maioria dos documentos, o gráfico demonstra que o truncamento de fato ocorre, mas restringe-se a uma parcela minoritária das redações mais extensas, visto que o percentil 95 atinge 549 *tokens*. Ainda assim, essa configuração garante que o encadeamento semântico do *corpus* seja processado sem perdas significativas de informação e sem necessidade de uso de técnicas como *chunking* para processamento de textos maiores.

Para fins de experimentação e avaliação comparativa de desempenho, a amostragem resultou na geração de dois *corpora* finais destinados ao treinamento: um conjunto composto por 1.285 redações e um conjunto expandido contendo 3.693 redações. Adicionalmente, estruturou-se um *dataset* de teste estritamente apartado, constituído por 261 redações. Todos estes conjuntos de dados foram balanceados e divididos em três subconjuntos distintos, refletindo o espectro contínuo da autoria textual.

3.2.1 Conjunto de Autoria Humana

O primeiro subconjunto constitui a classe de controle base, composta por textos de autoria puramente humana. Para garantir a verossimilhança e a diversidade demográfica, linguística e cognitiva das amostras, os textos foram extraídos do *dataset* Essay-BR (MA-

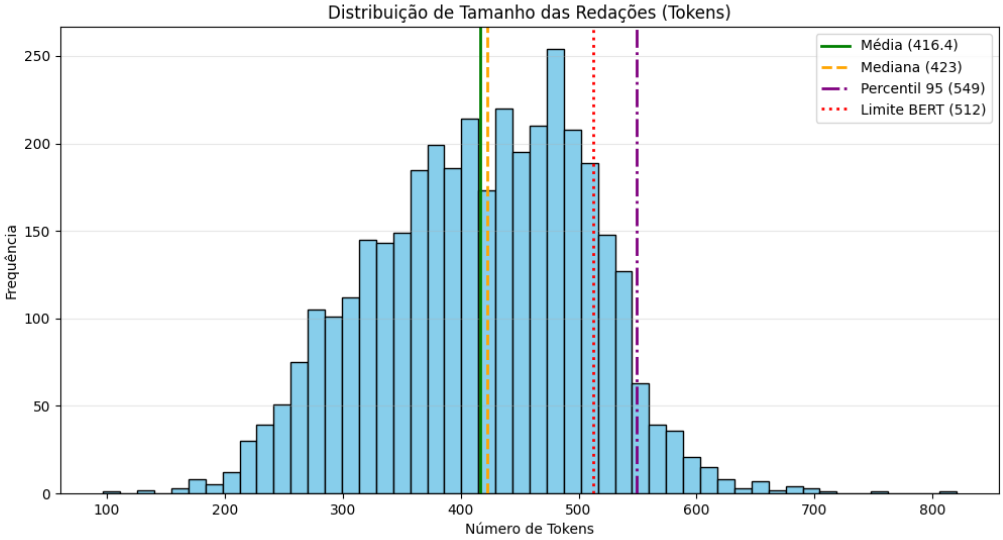


Figura 3.1: Distribuição de *tokens* para um total de 3693 redações. Fonte: do autor.

RINHO; ANCHIêTA; MOURA, 2022), um *corpus* público e consolidado na literatura nacional, concebido originalmente com o objetivo de impulsionar a pesquisa em avaliação automática de redações no Brasil.

Este conjunto é composto por redações dissertativo-argumentativas redigidas por estudantes do ensino médio, as quais foram avaliadas manualmente por especialistas seguindo rigorosamente os critérios oficiais do ENEM (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025). Os textos originais foram extraídos por meio de técnicas de *web scraping*[7] de plataformas educacionais *online* (Vestibular UOL e Educação UOL) entre os anos de 2015 e 2020. Esse recorte temporal é de extrema relevância metodológica para esta pesquisa, pois atesta que a coleta foi realizada em um período anterior ao advento e popularização comercial dos LLMs. Essa característica mitiga substancialmente o risco de contaminação do *corpus* e confere um alto grau de confiabilidade à premissa de que a amostra de controle é constituída por textos de autoria estritamente humana.

Em sua versão base, o *corpus* contém 4.570 redações que abrangem 86 temas distintos e complexos (como direitos humanos, COVID-19, política e *fake news*). Posteriormente, os pesquisadores responsáveis disponibilizaram uma atualização estendida da base, elevando a volumetria para 6.577 redações distribuídas ao longo de 151 temas.

[7] *Web scraping* é uma técnica computacional automatizada utilizada para extrair, coletar e estruturar informações em larga escala diretamente de páginas e portais da internet.

Essas duas iterações do Essay-BR (original e estendida) serviram como referência direta para a composição dos dois *corpora* de treinamento elaborados neste trabalho. O primeiro conjunto possui 1.285 amostras no total, das quais 425 são redações de autoria humana extraídas da versão original sob um critério de seleção restrito, que filtrou apenas textos com pontuação superior a 600 pontos. Esse corte teve como objetivo estabelecer uma linha de base para a competência linguística, forçando o modelo a aprender a distinguir o texto sintético de produções humanas já estruturalmente coesas. Em contrapartida, o conjunto expandido, constituído por 3.693 amostras totais, derivou da versão estendida e adotou um limiar de nota mais abrangente, englobando produções a partir de 440 pontos, o que resultou na extração de 1.213 redações humanas. A redução intencional desse limiar serviu para injetar um nível controlado de ruído estatístico no treinamento, incorporando desvios gramaticais e imperfeições típicas de proficiências intermediárias. Cabe ressaltar que esse processo de amostragem foi executado de maneira estratificada: priorizou-se a seleção de textos com avaliações mais altas para assegurar a consistência do padrão culto, ao mesmo tempo em que se incluiu uma proporção representativa de redações com notas menores. A adoção desse escopo amplo é metodologicamente fundamental, pois maximiza a diversidade da base e permite que o classificador capte com maior fidelidade as flutuações de proficiência, a variabilidade estrutural e os desvios inerentes à escrita humana autêntica.

Essas redações contêm imperfeições naturais inerentes ao processo de escrita sob pressão e ao desenvolvimento cognitivo dos alunos. Consequentemente, a utilização da base do Essay-BR fornece ao classificador os parâmetros necessários para que ele aprenda a reconhecer matematicamente essa assinatura imperfeita, porém autêntica, da autoria humana.

3.2.2 Conjunto Gerado por LLMs

O segundo subconjunto mapeia o extremo oposto do espectro de autoria, compreendendo as produções exclusivamente sintéticas (*zero-shot generation*). Para mitigar vieses associados à assinatura estatística e ao estilo de codificação de um único modelo de inteligência artificial, a infraestrutura de coleta foi projetada para execução automatizada de geração

em larga escala, utilizando a API[8] do ecossistema Ollama[9]. Por meio dessa integração, os mesmos temas abordados no subconjunto humano foram distribuídos e processados de forma balanceada por cinco diferentes LLMs no estado da arte, mais especificamente DeepSeek-V3.1 (671B), Qwen 3.5 (397B), GPT-OSS (120B), Kimi 2.5 e Gemini 3 Flash (*preview*). Essa diversidade garante que o *corpus* artificial englobe variações de vocabulário e estrutura sintática, visto que cada modelo opera com algoritmos próprios e parâmetros distintos para a geração textual.

A geração das amostras foi parametrizada criticamente através de engenharia de *prompts*, instruindo os modelos a atuarem em conformidade com as diretrizes oficiais e critérios de correção do ENEM (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025). Com o objetivo de mitigar a homogeneidade dos dados sintéticos e introduzir uma variabilidade linguística representativa dentro da própria classe artificial, foram desenvolvidos e aplicados três *prompts* de indução distintos. Cada diretriz foi desenhada para simular uma *persona* de estudante com um nível de proficiência específico.

Para simular o perfil de um aluno excelente, o modelo foi instruído a produzir textos com vocabulário culto, composto por termos menos frequentes na linguagem cotidiana e que refletem maior sofisticação lexical, além de estruturas sintáticas bem elaboradas, repertório sociocultural diversificado e aplicação rigorosa e precisa dos mecanismos de coesão textual. O *prompt* operacionalizado para este perfil foi estruturado da seguinte forma:

> "Você é um especialista em redações do ENEM. Seu objetivo é escrever textos dissertativo-argumentativos de excelência, seguindo rigorosamente as cinco competências do exame. Seguem as competências:
>
> 1 - Demonstrar domínio da modalidade escrita formal da língua portuguesa.
>
> 2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
>
> 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e ar-

[8] *Application Programming Interface* (Interface de Programação de Aplicações). Trata-se de um conjunto de rotinas e protocolos computacionais que permite a comunicação padronizada e a integração automatizada entre diferentes sistemas de *software*.

[9] O Ollama é uma plataforma de código aberto (*open-source*) que simplifica a execução, o gerenciamento e a integração de LLMs diretamente em infraestruturas locais ou servidores privados.

gumentos em defesa de um ponto de vista.

4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.

5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Além disso, siga estritamente estas regras:

- Estrutura: o texto deve ter exatamente 4 parágrafos (1 Introdução, 2 de Desenvolvimento e 1 Conclusão).

- Tamanho: entre 250 e 350 palavras.

- Repertório sociocultural: inclua repertório sociocultural quando pertinente, podendo ser explícito ou implícito. Esse repertório pode ser uma citação ou alusão histórica, filosófica ou literária válida e bem contextualizada.

- Proposta de intervenção: a conclusão deve conter uma proposta de intervenção completa e detalhada, respondendo: Quem fará? O que fará? Como fará? Qual o efeito esperado? E um detalhamento adicional.

- Tom: padrão culto da língua portuguesa, impessoal (terceira pessoa), objetivo e formal. Não use saudações, não coloque título e devolva apenas o texto da redação.

- Vocabulário: varie o uso de conectivos, evitando padrões repetitivos entre textos.

Com essas informações, redija uma redação no formato do ENEM sobre o seguinte tema: {tema}."

Em um segundo cenário, o perfil de um aluno bom configurou a geração para apresentar um domínio formal seguro e clara articulação de ideias, porém utilizando construções frasais mais convencionais, repertório produtivo e um vocabulário padrão, sem excessos ou preciosismos lexicais. Para este perfil, o *prompt* foi elaborado segundo a seguinte estrutura:

"Você é um aluno com bom desempenho em redações do ENEM. Seu objetivo é escrever textos dissertativo-argumentativos consistentes, atendendo bem às cinco competências do exame, embora com pequenas limitações de aprofundamento ou refinamento. Seguem as competências:

1 - Demonstrar domínio da modalidade escrita formal da língua portuguesa.

2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhe-

cimento para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.

3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.

4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.

5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Além disso, siga estas regras:

- Estrutura: o texto deve ter, preferencialmente, 4 parágrafos (1 Introdução, 2 de Desenvolvimento e 1 Conclusão), mas pode apresentar variações na divisão dos parágrafos de desenvolvimento.

- Tamanho: entre 230 e 350 palavras.

- Repertório sociocultural: inclua repertório sociocultural quando pertinente, mas, em alguns casos, ele pode ser mais genérico, pouco aprofundado ou não totalmente integrado ao argumento.

- Argumentação: os argumentos devem ser coerentes, porém podem apresentar menor nível de criticidade, aprofundamento ou sofisticação analítica.

- Coesão: utilize conectivos adequados, ainda que com menor variedade ou com algumas repetições.

- Proposta de Intervenção: a conclusão deve conter proposta de intervenção relacionada ao tema, preferencialmente indicando: Quem fará, o que será feito e qual o objetivo, podendo apresentar menor detalhamento nos meios ou efeitos.

- Tom: padrão culto da língua portuguesa, impessoal (terceira pessoa), objetivo e formal. Não use saudações, não coloque título e devolva apenas o texto da redação.

Com essas informações, redija uma redação no formato do ENEM sobre o seguinte tema: {tema}."

Por fim, para representar uma faixa de desempenho intermediária, o perfil de um aluno mediano orientou o desenvolvimento de redações com argumentação linear e previsível, períodos majoritariamente curtos ou diretos, uso de conectivos básicos e repertório baseado estritamente nos textos motivadores fornecidos pela proposta. O *prompt* empregado

na configuração deste perfil foi definido conforme a seguinte estrutura:

> "Você é um aluno com desempenho mediano em redações do ENEM. Seu objetivo é escrever um texto dissertativo-argumentativo que atenda parcialmente às competências do exame, apresentando limitações em argumentação, coesão e detalhamento. Seguem as competências:
>
> 1 - Demonstrar domínio da modalidade escrita formal da língua portuguesa.
>
> 2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
>
> 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
>
> 4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
>
> 5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.
>
> Além disso, siga estas regras:
>
> - Estrutura: o texto deve ter entre 3 e 4 parágrafos, podendo apresentar organização irregular ou pouco equilibrada entre introdução, desenvolvimento e conclusão.
>
> - Tamanho: entre 210 e 350 palavras.
>
> - Repertório sociocultural: pode incluir repertório sociocultural, mas ele tende a ser genérico, pouco específico ou apenas citado, sem aprofundamento claro.
>
> - Argumentação: os argumentos podem ser superficiais, com explicações curtas, pouca análise crítica ou exemplos previsíveis.
>
> - Coesão: utilize conectivos básicos (como "além disso", "porém", "por isso"), ainda que com repetição ou uso pouco estratégico.
>
> - Proposta de intervenção: a conclusão deve apresentar uma tentativa de intervenção, podendo faltar alguns dos elementos (quem fará, o que será feito, como ou para quê), sem necessidade de detalhamento completo.
>
> - Tom: priorize linguagem formal, mas permite a presença de construções menos refinadas ou ligeiramente informais. Não use título e devolva apenas o texto da redação.
>
> Com essas informações, redija uma redação no formato do ENEM sobre o seguinte

tema: {tema}."

Apesar da estratificação intencional de competência argumentativa e estilística, o resultado consolidado dessa geração em massa preserva os traços intrínsecos da escrita computacional: excelência gramatical nativa, coerência semântica de longa distância inquebrável e uma completa ausência de falhas biomecânicas ou desvios atencionais. A constituição desse subconjunto diversificado revela-se de extrema relevância para o treinamento do modelo. Ao expor a rede neural a múltiplos graus de sofisticação textual gerada por máquina, mitiga-se o risco de sobreajuste (*overfitting*) a um único arquétipo, como a presunção estatística errônea de que todo texto sintético é obrigatoriamente erudito e complexo. Consequentemente, o modelo torna-se capaz de mapear de forma robusta, em seu espaço latente de alta dimensão, as propriedades matemáticas subjacentes e as distribuições probabilísticas que definem a geração automatizada em qualquer faixa de desempenho. Essa representação vetorial pura consolida a linha de base focada na natureza matemática da IA, fornecendo ao classificador o contraexemplo exato do que é um texto estritamente livre de hesitações, atritos ou custos de carga cognitiva.

Ao final deste processo de geração automatizada, foram consolidadas 430 amostras puramente sintéticas para integrar o primeiro conjunto de dados (composto por 1.285 redações totais). Posteriormente, esse volume de textos foi ampliado para 1.275 instâncias, visando estruturar a respectiva classe no *corpus* estendido, que conta com 3.693 redações no total.

3.2.3 Conjunto Híbrido: Simulação de *Copy-Typing*

O terceiro subconjunto constitui a classe híbrida, que visa mapear as zonas fronteiriças do espectro de autoria. Para a composição deste grupo, devido à falta de dados públicos que atendam a esse cenário específico, novos textos foram submetidos a um processo de corrupção controlada e simulação algorítmica do comportamento de *copy-typing*. É fundamental destacar que, para mitigar o risco de vazamento de dados[10] e garantir a in-

[10]O vazamento de dados (*data leakage*) consiste em uma falha metodológica em aprendizado de máquina que ocorre quando informações do conjunto de validação ou características exclusivas de uma classe são indevidamente incorporadas ao conjunto de treinamento de outra. Essa contaminação permite que o modelo decore instâncias específicas ou aprenda atalhos matemáticos irreais, resultando em métricas de desempenho artificialmente altas, mas que falham gravemente em cenários reais.

dependência estatística entre as classes do classificador, não houve o reaproveitamento dos textos puramente sintéticos descritos na subseção anterior. Em vez disso, dividiu-se um fluxo em duas etapas: primeiramente, utilizou-se a mesma infraestrutura de integração com o ecossistema Ollama para gerar uma base de dados inédita e exclusiva, empregando estritamente o *prompt* de perfil excelente de forma distribuída entre os cinco LLMs selecionados, com o intuito de obter textos sintéticos de alta sofisticação estrutural e gramatical.

Na segunda etapa, adotou-se uma estratégia de processamento cruzado (*cross-model prompting*), exemplificada na Figura 3.2, na qual cada texto originalmente sintetizado por uma arquitetura geradora (LLM X) foi sistematicamente submetido ao segundo *prompt* de indução utilizando, obrigatoriamente, um modelo distinto (LLM Y) dentre as cinco opções integradas à infraestrutura. Neste contexto metodológico, a abordagem impede que a inteligência artificial utilize seus próprios vieses internos e representações latentes para alterar um texto que ela mesma criou, garantindo que a inserção de ruído seja mais independente e fiel a uma intervenção externa.

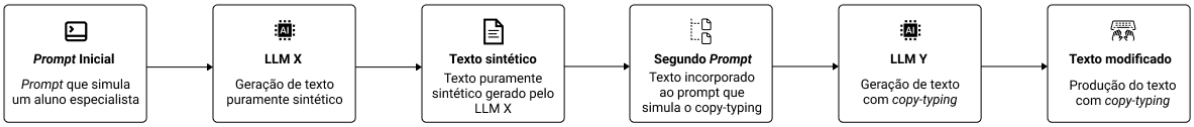


Figura 3.2: Fluxograma da metodologia de processamento cruzado aplicada para geração de textos com *copy-typing*. Fonte: do autor.

O desenvolvimento deste segundo comando ocorreu por meio de um processo iterativo de refinamento focado no comportamento do usuário. Inicialmente, o *prompt* fundamentou-se na observação empírica de dinâmicas reais de *copy-typing*, a partir das quais constatou-se que as intervenções humanas aplicadas a um texto de máquina, sejam de natureza subconsciente ou deliberada, seguem padrões comportamentais nomeáveis, categorizáveis e replicáveis. Durante os primeiros ciclos de teste, baseados apenas na inserção de erros de digitação e na simplificação de vocabulário e de conectivos, notou-se que as edições ainda se mantinham excessivamente limpas e mecânicas, incapazes de refletir o ruído orgânico humano.

Para corrigir esse desvio, a arquitetura do comando foi submetida a múltiplos ciclos de calibração, ajustando-se gradativamente a quantidade de edições, as restrições, os pesos atribuídos a cada tipo de alteração e a inserção de novos mecanismos de ofuscação. Dessa forma, a versão final do *prompt* consolidou as técnicas de corrupção aplicadas pelo LLM secundário em diretrizes específicas, que incluíram: simplificação de vocabulário (principalmente de conectivos), paráfrases superficiais e profundas, edições estruturais leves, reordenação de ideias e argumentos, substituição de figuras de linguagem e, por fim, a simulação de erros ortográficos, como omissão de pontuação, trocas de teclas adjacentes e anomalias de capitalização.

Sendo assim, a simulação desta classe não se limitou à inserção aleatória de ruído, mas seguiu os fundamentos psicolinguísticos da carga cognitiva e das modificações intencionais, aplicando alterações estruturais que reproduzem as ações de estudantes mal-intencionados. O *prompt* adotado para este perfil apresenta a seguinte estrutura:

> "Você receberá uma redação.
>
> Não reescreva o texto todo. Sua tarefa é manter a estrutura original, alterando apenas pequenos trechos espalhados pelo texto para simular que um estudante o copiou manualmente (*copy-typing*) e fez pequenas adaptações.
>
> Siga rigorosamente todas as regras abaixo:
>
> Diretrizes:
> - NÃO reescreva o texto inteiro;
> - Insira um total de 9 a 15 modificações pontuais (marcas de *copy-typing* descritas abaixo) distribuídas aleatoriamente ao longo dos parágrafos.
>
> Tipos de modificações a serem aplicadas (cada uma deve ser aplicada ao menos uma vez e pode haver repetição) por ordem de prioridade:
> 1 - Simplificação de vocabulário: substitua algumas palavras, expressões ou trechos muito formais por versões com o vocabulário mais simples e comum;
> 2 - Conectivos mais simples: troque alguns conectivos sofisticados originais por opções mais cotidianas;
> 3 - Paráfrases profundas: escolha entre duas a quatro frases completas e reescreva-as do zero. Elas devem manter o sentido original, mas usar uma estrutura sintática e

um vocabulário totalmente diferentes de forma a reestruturar o texto e aumentar a perplexidade. (Ex.: transformar voz passiva em ativa, mudar a ordem dos termos da oração ou substituir substantivos abstratos por verbos).

4 - Quebra de fluidez: reordene, fragmente frases e altere a ordem de trechos dentro de um parágrafo de forma a modificar a uniformidade. (Ex.: uma ordem de: causa, explicação e consequência é transformada em: consequência, explicação e causa);

5 - Reescreva todos os trechos que contenham travessões substituindo-os por vírgulas ou frases separadas, se houver.

6 - Exclusão de figuras de linguagem: sempre que houver figuras de linguagem como hipérbato, anástrofe ou anacoluto, reescreva o trecho de forma a remover a figura de linguagem;

7 - Erros de digitação controlados (*typos*): introduza erros mecânicos de teclado. Faça isso no máximo 3 vezes. Exemplos aceitos: omissão pontual de acento (ex: "democraticos"), omissão de crase, inversão de letras (ex: "excludnete") ou duplicação acidental de artigos (ex: "o o exercício").

Regras de Preservação:

- Preserve as ideias principais, a ordem dos argumentos e a estrutura sintática da maioria das frases;
- Não adicione nem remova informações ou exemplos do texto original;
- NÃO introduza erros gramaticais graves que comprometam o entendimento;
- Evite padrões sistemáticos de alteração (ex.: simplificar todos os conectivos).
- Mantenha a norma culta da língua portuguesa, evite informalidades, mas também não seja muito formal.

Texto original: "

{texto}

"

Retorne APENAS o texto modificado."

Para validar quantitativamente a eficácia da simulação e corroborar a metodologia, realizou-se uma análise comparativa pareada entre os textos sintéticos originais e suas respectivas versões após o processo de *copy-typing*. A extração de métricas de similaridade consolida o sucesso da abordagem proposta: a razão do número de palavras entre

os pares manteve-se estável com média de 1,019, confirmando a preservação do escopo estrutural sem acréscimos ou deleções drásticas. Sob a ótica do significado, a similaridade semântica registrou índices consistentemente elevados em torno de 97,50%, atestando que o desenvolvimento argumentativo não foi significativamente alterado. Em contrapartida, as métricas focadas na superfície do texto revelaram quedas substanciais, com a similaridade léxica orbitando uma média de 76,24% e o Índice de Jaccard[11] apresentando valores próximos a 64,30%. Esse distanciamento estatístico entre a alta preservação semântica e a acentuada modificação léxica prova matematicamente que a metodologia resultou no comportamento esperado: houve a reestruturação da malha sintática, a simplificação de conectivos e a injeção de ruídos mecânicos sem a ocorrência de alucinações de conteúdo. Por fim, flutuações nas métricas de *burstiness*, aqui definida como o coeficiente de variação do comprimento das sentenças, isto é, a razão entre o desvio padrão e a média do número de palavras por frase, evidenciam a quebra bem-sucedida da uniformidade rítmica gerada pela máquina, mimetizando as inconsistências estruturais inerentes à produção textual humana. Essa medida quantifica o grau de heterogeneidade na extensão dos períodos, capturando variações entre frases curtas e longas ao longo do texto. Os resultados macroscópicos do processo foram sintetizados por meio de estatística descritiva, conforme apresentado na Tabela 3.1.

Tabela 3.1: Análise estatística descritiva das métricas de *copy-typing*.

Métrica	**Média**	**Mediana**	**Desvio Padrão**	**Mínimo**	**Máximo**
Similaridade Léxica	0,762	0,778	0,108	0,387	0,979
Similaridade Semântica	0,975	0,980	0,017	0,896	0,999
Índice de Jaccard	0,643	0,660	0,111	0,260	0,967
Razão de Palavras	1,019	1,015	0,043	0,768	1,179
Burstiness (IA)	7,627	7,438	1,916	3,172	14,584
Burstiness (Cópia)	7,637	7,401	1,966	3,637	14,180

Fonte: elaborado pelo autor.

O produto final desta etapa foram dois conjuntos, um com 430 redações e um segundo que é uma versão estendida com 1205 amostras, que preservam a coerência semântica de base gerada pela máquina, mas cuja superfície sintática e ortográfica apre-

[11] O Índice de Jaccard é uma medida de similaridade entre conjuntos, definida como a razão entre a cardinalidade da interseção e a cardinalidade da união dos conjuntos comparados. Neste trabalho, os conjuntos são formados pelas palavras distintas presentes em cada texto e a união é composta por todas as palavras distintas presentes em pelo menos um deles.

senta imperfeições tipicamente humanas. A consolidação deste subconjunto no treinamento força o classificador multimodal a aprender a decodificar e ponderar os sinais contraditórios da autoria assistida.

3.2.4 Conjunto de Teste

Para avaliar a capacidade de generalização e a robustez do classificador frente a dados inéditos, estruturou-se um conjunto de teste estatisticamente independente. A composição desta partição baseou-se em um rigoroso protocolo de segregação de dados, visando inviabilizar qualquer forma de contaminação metodológica ou memorização por parte da rede neural.

A classe representativa da autoria humana foi constituída por 86 redações reais. É imperativo destacar que essas amostras foram previamente segregadas e isoladas das bases de treinamento (tanto na iteração de 1.285 amostras quanto na estendida de 3.693), garantindo que o modelo jamais as tenha processado durante a fase de otimização de pesos.

No que tange às classes de geração artificial (produção exclusivamente sintética e híbrida), adotou-se a premissa de não reaproveitamento absoluto. Nenhuma redação gerada ou submetida ao algoritmo de *copy-typing* nas etapas de treinamento foi reutilizada. Em vez disso, replicou-se exatamente a mesma metodologia de integração automatizada e processamento cruzado (*cross-model prompting*) detalhada nas subseções anteriores para sintetizar textos integralmente novos.

A aplicação deste fluxo de geração independente resultou na consolidação de 90 redações puramente sintéticas (livres de intervenção) e 85 redações submetidas à simulação de *copy-typing*. Dessa forma, o conjunto de teste final totaliza 261 amostras inéditas distribuídas ao longo de todo o espectro de autoria, fornecendo um ambiente de avaliação confiável para aferir as métricas de desempenho final dos modelos.

3.3 Treinamento

Para o desenvolvimento da etapa de classificação, foram propostas e implementadas duas arquiteturas distintas de (*fine-tuning*) sobre o modelo pré-treinado BERTimbau (neuralmind/bert-base-portuguese-cased). A primeira baseia-se em uma abordagem direta de aprendizado profundo, operando exclusivamente sobre a semântica vetorial do texto. A segunda, por sua vez, expande esse paradigma ao incorporar uma estratégia de fusão tardia (*late fusion*) para a injeção de características linguísticas predefinidas (*features*). A fim de avaliar o impacto da volumetria e da diversidade dos dados no aprendizado das redes neurais, ambas as arquiteturas foram treinadas e validadas sobre os dois *corpora* previamente estabelecidos (o conjunto restrito de 1.285 redações e o conjunto estendido de 3.693 redações). Esse delineamento experimental resultou na geração de quatro modelos distintos, permitindo uma análise comparativa rigorosa de desempenho e de capacidade de generalização.

Para a execução do processo de otimização dos pesos, as amostras de cada *corpus* foram divididas na proporção de 80% para o conjunto de treinamento e 20% para o conjunto de validação, uma distribuição aplicada uniformemente para todas as modelagens. É imprescindível destacar que o conjunto de teste foi mantido completamente apartado e isolado dessa etapa de ajuste e seleção de hiperparâmetros, assegurando a imparcialidade estatística na avaliação final do sistema.

3.3.1 Arquitetura Base

A arquitetura base constitui o modelo fundamental desta pesquisa, sendo projetada para classificar a autoria das redações apoiando-se estritamente na capacidade de abstração do codificador profundo, sem o aporte de variáveis numéricas externas. O fluxo de processamento inicia-se com a tokenização do documento original, estabelecendo um limite máximo de 512 *tokens* por entrada. Nessa etapa, aplicam-se técnicas de truncamento e preenchimento (*padding*) para garantir a uniformidade da matriz tensorial exigida pela rede neural.

Uma vez concluído o mapeamento dos *tokens*, no que tange à extração de características latentes, em vez de adotar a representação simplificada do *token* de classificação

global (`[CLS]`), o modelo foi projetado para implementar a técnica de *masked mean pooling*. Essa abordagem calcula a média ponderada de todos os vetores da última camada oculta (*last hidden state*), aplicando uma máscara computacional baseada na máscara de atenção original para anular a interferência dos *tokens* de preenchimento. Essa decisão arquitetural permite que a rede capture uma representação semântica e estrutural significativamente mais rica, sumarizando toda a extensão do texto.

O vetor resultante, composto por 768 dimensões, é então submetido a uma rede neural classificadora sequencial. Esta cabeça de classificação (*classification head*) é constituído por uma camada linear densa de 256 neurônios, seguida de uma função de ativação não linear ReLU e de uma camada de regularização *Dropout* com probabilidade de 30% ($p = 0,3$), atuando como um forte mecanismo de prevenção ao sobreajuste (*overfitting*). Por fim, uma última transformação linear projeta as representações latentes para a dimensionalidade de saída, correspondente às três classes do problema (autoria humana, geração puramente sintética e autoria híbrida).

O processo de otimização dos pesos foi conduzido pelo algoritmo AdamW (LOSHCHILOV; HUTTER, 2017), configurado com uma taxa de aprendizado conservadora de 1×10^{-5} e decaimento de peso (*weight decay*) de $0,01$. O treinamento utilizou a função de perda *cross-entropy loss* e foi estruturado para operar com lotes (*batches*) de 8 amostras por um período máximo de 10 épocas. Para garantir o melhor ponto de generalização, implementou-se um mecanismo de parada antecipada (*early stopping*), encerrando o treinamento caso a perda de validação não apresentasse melhorias por duas épocas consecutivas (*patience*).

A dinâmica de aprendizado desta arquitetura pode ser observada por meio do monitoramento da função de perda (*loss*) ao longo das épocas. As Figuras 3.3 e 3.4 ilustram as curvas de convergência para os treinamentos realizados sobre o conjunto restrito (1.285 amostras) e o conjunto estendido (3.693 amostras), respectivamente.

No treinamento sobre o conjunto restrito (Figura 3.3), observa-se uma rápida convergência inicial, com a perda de validação atingindo seu ponto de mínimo global próximo à quinta época e estabilizando-se em seguida. A perda de treinamento, de forma esperada devido à alta capacidade paramétrica do BERTimbau, continua a decair em

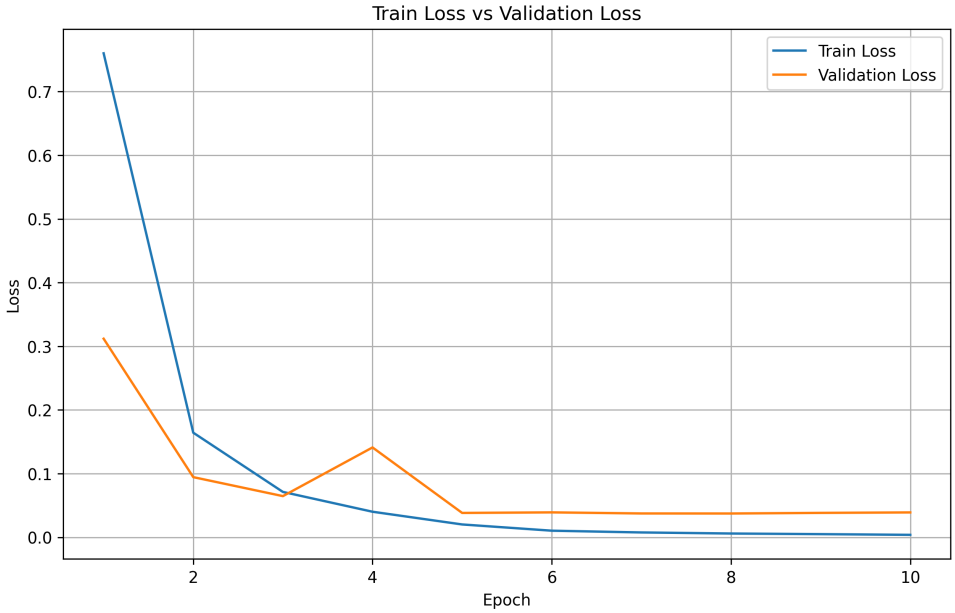


Figura 3.3: Curva de perda (treinamento e validação) da arquitetura base sobre o conjunto restrito. Fonte: do autor.

direção a zero.

Figura 3.4: Curva de perda (treinamento e validação) da arquitetura base sobre o conjunto estendido. Fonte: do autor.

Por outro lado, o treinamento sobre o *corpus* estendido (Figura 3.4) reflete a maior complexidade e variabilidade estrutural inerentes a um volume de dados substancialmente maior. A curva de validação apresenta flutuações pontuais ao longo da otimização, um comportamento característico diante de *batches* com maior diversidade textual, atingindo seu ponto ótimo na sexta época. Em ambos os cenários, o ponto ótimo foi capturado com

precisão pelo *early stopping*, garantindo que a rede salvasse os melhores pesos latentes antes de iniciar um regime de memorização (*overfitting*) severo.

Um diferencial metodológico desta arquitetura encontra-se em sua etapa de pós-processamento, focada na calibração e confiabilidade das predições. Em contextos punitivos e sensíveis, como a avaliação de fraudes acadêmicas, não é suficiente que o modelo atinja uma alta taxa de acerto na classificação; é necessário que a sua confiança matemática esteja estritamente alinhada com a probabilidade real de a predição estar correta. Para atingir esse alinhamento, implementou-se inicialmente a calibração de probabilidades (*temperature scaling*). Trata-se de um método de pós-processamento destinado a ajustar as probabilidades preditas pela rede neural à sua acurácia empírica observada, mitigando problemas de descalibração frequentemente presentes em modelos de aprendizado profundo. O procedimento baseia-se na otimização de um único hiperparâmetro escalar positivo, denominado temperatura (T), estimado sobre o conjunto de validação por meio do algoritmo quasi-Newton L-BFGS, que utiliza uma aproximação de memória limitada da matriz Hessiana para resolver problemas de otimização de grande escala de maneira eficiente (LIU; NOCEDAL, 1989). Matematicamente, os *logits* brutos produzidos pela última camada linear são divididos por T antes da aplicação da transformação *softmax*. Como essa operação preserva a monotonicidade, a relação de ordem original das classes não é alterada, garantindo que as predições finais permaneçam estáveis. Todavia, a modificação na entropia da distribuição gerada suaviza ou acentua as confianças matemáticas, corrigindo o fenômeno do excesso de confiança estatística (GUO et al., 2017) e fornecendo estimativas de probabilidade confiáveis para a caracterização da incerteza do modelo.

Em conjunto com a calibração, aplicou-se o ajuste de limiares de decisão (*threshold tuning*), que representa uma estratégia de otimização projetada para flexibilizar os critérios de inferência, substituindo a regra de decisão convencional baseada puramente na seleção da maior probabilidade contínua. Para tanto, implementou-se uma busca exaustiva em grade tridimensional no espaço de validação, mapeando vetores de limiares operacionais independentes para cada uma das três classes do problema dentro do intervalo discreto de $0,3$ a $0,8$. Essa técnica visa delimitar fronteiras mais robustas para

o classificador, localizando a combinação exata de restrições probabilísticas que maximize a acurácia global do sistema. Ao estabelecer limiares customizados para os espectros de autoria humana, sintética e híbrida, a arquitetura adquire maior resiliência na segregação de amostras ambíguas, refinando o comportamento do modelo na zona de transição do *copy-typing* sem demandar a recalibragem dos pesos internos do codificador.

Por fim, realizou-se uma análise de incerteza, uma vez que, em sistemas de classificação de alta confiabilidade, a capacidade de um modelo quantificar o grau de incerteza associado às suas próprias predições é tão importante quanto sua acurácia. Como a tendência natural da rede ao excesso de confiança (*overconfidence*) já foi mitigada pela etapa de *temperature scaling*, as probabilidades calibradas foram submetidas a uma auditoria baseada no contraste entre a confiança atribuída à classe predita e a Entropia de Shannon (SHANNON, 1948). Essa métrica quantifica matematicamente o grau de dispersão na distribuição de probabilidades gerada pelo modelo, sendo definida pela seguinte equação:

$$H(P) = -\sum_{i=1}^{n} P(i) \log P(i)$$

Onde $P(i)$ representa a probabilidade calibrada de a amostra pertencer à classe i, e n corresponde ao número total de classes do problema ($n = 3$). O contraste entre essas duas grandezas permite avaliar a real hesitação algorítmica da rede. Em cenários da classe híbrida (*copy-typing*), por exemplo, nos quais traços linguísticos estruturais da máquina e imperfeições da autoria humana estão sobrepostos, uma baixa confiança aliada a uma alta entropia indica um conflito latente na representação do texto no espaço vetorial. Consequentemente, esse monitoramento serve não apenas como um indicador para identificar instâncias fronteiriças, mas também como uma ferramenta fundamental de diagnóstico computacional para investigar as falhas lógicas e os limites de generalização em casos de predição incorreta.

Ao aliar a capacidade de abstração sintática do BERTimbau a mecanismos que penalizam o excesso de confiança algorítmica e calibram as fronteiras de decisão baseadas na entropia real da amostra, o modelo consolida-se em consonância com o estado da arte na literatura recente sobre aprendizado de máquina seguro e detecção de autoria

generativa.

3.3.2 Arquitetura Híbrida: Fusão Tardia

A segunda abordagem arquitetural proposta implementa o paradigma de fusão tardia (*late fusion*). Diferentemente da arquitetura base, que delega toda a extração de conhecimento à capacidade de abstração do modelo de aprendizado profundo, esta variação estabelece uma rede híbrida multimodal. O objetivo primário é enriquecer o espaço latente com sinais determinísticos, injetando métricas estilométricas e de complexidade linguística extraídas *a priori* das redações.

Para a composição do vetor de características (*features*), foram selecionadas seis métricas textuais calculadas durante o pré-processamento: *burstiness*, Medida de Diversidade Lexical Textual (MTLD)[12] (MCCARTHY; JARVIS, 2010), densidade de pontuação, e as contagens absolutas de vírgulas, pontos e vírgulas e pontos finais. Devido à acentuada disparidade de grandezas, variâncias e escalas naturais dessas variáveis, aplicou-se um método de padronização estatística baseado no *z-score*[13] para gerar um nivelamento matemático estritamente necessário ao combinar variáveis heterogêneas, pois impede que características com escalas numéricas naturalmente maiores dominem desproporcionalmente a atualização dos pesos da rede neural em detrimento de métricas com escalas menores. É fundamental ressaltar que, para assegurar a integridade metodológica e mitigar qualquer risco de vazamento de dados (*data leakage*), os parâmetros de normalização (média e desvio padrão) foram ajustados estritamente sobre as amostras do conjunto de treinamento e, posteriormente, projetados de forma estática para as instâncias de validação.

Na topologia da rede neural, o fluxo de dados bifurca-se em duas trilhas de processamento simultâneas. Na primeira trilha, os *tokens* do texto atravessam o codificador

[12] A Medida de Diversidade Lexical Textual (MTLD — *Measure of Textual Lexical Diversity*) é um índice computacional de riqueza vocabular. Ao contrário de métricas tradicionais baseadas na razão direta entre palavras únicas e o total de palavras, como o *Type-Token Ratio* (TTR) e suas variantes estabilizadas (*Root* TTR e *Log* TTR), que apresentam alta sensibilidade e viés de degradação analítica em textos mais extensos, o MTLD é calculado a partir do comprimento médio de sequências textuais que mantêm um limiar constante de variação lexical. Essa abordagem metodológica torna o índice matematicamente robusto e estritamente independente da extensão total do documento avaliado, viabilizando comparações justas entre amostras de tamanhos distintos.

[13] (*Z-score*) é uma técnica de normalização que redimensiona os dados de forma que a distribuição resultante apresente média nula ($\mu = 0$) e desvio padrão unitário ($\sigma = 1$)

pré-treinado BERTimbau, culminando na já descrita operação de *masked mean pooling*, que sumariza a semântica de longa distância em um vetor denso de 768 dimensões. Paralelamente, na segunda trilha, o vetor numérico padronizado contendo as seis métricas manuais é submetido a uma sub-rede de projeção independente. Este bloco sequencial é composto por uma camada linear que expande a dimensionalidade da entrada de 6 para 32 neurônios, acompanhada por uma função de ativação ReLU e uma camada de regularização *dropout* de 20% ($p = 0,2$). A finalidade desta projeção é transformar as variáveis brutas em uma representação de maior dimensionalidade, garantindo que elas possuam peso representativo e estabilidade matemática ao interagirem com o núcleo semântico profundo.

A integração entre os fluxos, que caracteriza a fusão tardia, ocorre na fase final de codificação: o vetor semântico (768 dimensões) é concatenado ao vetor de métricas projetadas (32 dimensões), resultando em um tensor combinado unificado de 800 dimensões. Esse tensor enriquecido passa então a alimentar a cabeça de classificação final da rede, cuja primeira transformação linear foi expandida para suportar a nova dimensão de entrada ($800 \rightarrow 256$ neurônios), preservando a aplicação subsequente da ativação ReLU e do *Dropout* de 30% ($p = 0,3$).

Por fim, para garantir a validade científica das análises comparativas entre a arquitetura base e a arquitetura híbrida, todo o ecossistema de otimização foi mantido rigorosamente estático. Isso inclui o algoritmo otimizador (AdamW), a taxa de aprendizado (1×10^{-5}), o critério de parada antecipada (*early stopping*) e as três etapas de pós-processamento algorítmico: calibração de probabilidades, ajuste dinâmico de limiares e auditoria analítica via Entropia de Shannon. Dessa forma, garante-se que qualquer variação nas métricas de desempenho final decorra do impacto da injeção das variáveis estilométricas.

A dinâmica de otimização dessa arquitetura pode ser observada por meio do monitoramento da função de perda (*loss*) durante o treinamento. As Figuras 3.5 e 3.6 ilustram a convergência do modelo ao longo das épocas para o conjunto restrito e para o conjunto estendido, respectivamente.

Na Figura 3.5, referente ao conjunto de 1.285 redações, nota-se uma convergência

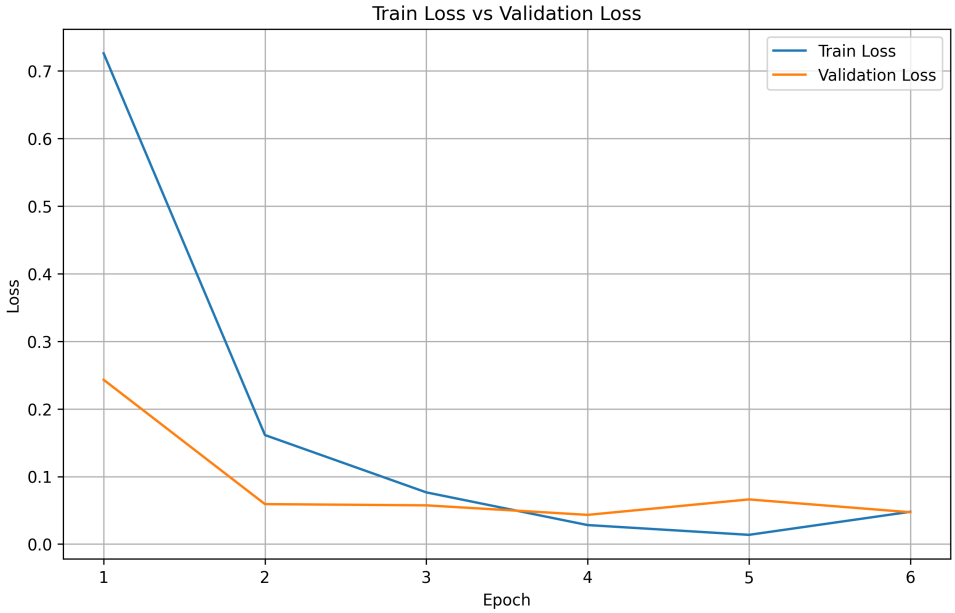


Figura 3.5: Curva de perda (treinamento e validação) da arquitetura híbrida sobre o conjunto restrito. Fonte: do autor.

acelerada, em que a perda de validação atinge seu ponto ótimo de generalização na quarta época. A partir desse ponto, o distanciamento entre as curvas de treino e validação sinaliza o início de um sobreajuste (*overfitting*) incipiente, o qual foi prontamente interceptado e interrompido pelo mecanismo de *early stopping*.

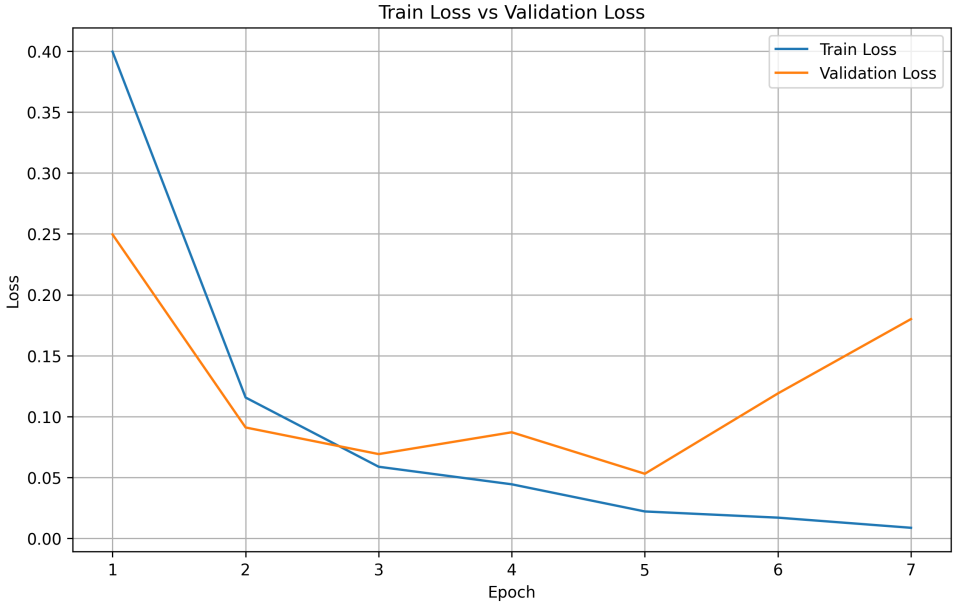


Figura 3.6: Curva de perda (treinamento e validação) da arquitetura híbrida sobre o conjunto estendido. Fonte: do autor.

Comportamento análogo e ainda mais pronunciado é observado no treinamento sobre o *corpus* estendido (Figura 3.6). A rede atinge sua capacidade máxima de genera-

lização na quinta época. Nas épocas subsequentes, ocorre uma nítida elevação da perda de validação, caracterizando o início de um *overfitting*. O registro visual destas curvas comprova matematicamente a eficácia do ecossistema de otimização adotado, atestando que a injeção multimodal não desestabilizou a descida do gradiente e que os pesos finais retidos representam o ponto exato de máxima eficiência do classificador antes da degradação.

3.4 Explicabilidade do Modelo

No contexto do desenvolvimento de arquiteturas profundas aplicadas à detecção de fraudes educacionais e plágio sistêmico, a mera constatação estatística da acurácia é insuficiente. A validação perante as instâncias pedagógicas e auditorias exige que as redes neurais não operem com meras atribuições de probabilidades inauditáveis, mas sim que as justificativas lógicas subjacentes às suas predições possam ser mapeadas e compreendidas por operadores humanos.

Nesse contexto, este trabalho incorpora técnicas de XAI projetadas para fornecer interpretações locais e globais das decisões do classificador híbrido. O objetivo principal é permitir que pesquisadores, docentes e especialistas compreendam não apenas a predição produzida pelo modelo, mas também os fatores que conduziram a essa decisão e o grau de confiança associado a ela.

3.4.1 Integração com a Arquitetura Híbrida

A aplicação de técnicas de XAI em arquiteturas multimodais apresenta desafios adicionais em comparação aos modelos puramente textuais. O classificador proposto utiliza simultaneamente duas fontes de informação: representações semânticas profundas produzidas pelo codificador *transformer* e características estilométricas extraídas manualmente, tais como *burstiness*, MTLD e métricas de pontuação.

Em métodos de explicabilidade baseados em perturbação, como LIME e SHAP, o algoritmo gera múltiplas versões modificadas do texto original, removendo ou substituindo palavras e expressões para avaliar o impacto dessas alterações sobre a decisão do modelo. Entretanto, em uma arquitetura híbrida, uma simples modificação do texto

implica, inevitavelmente, alterações em suas propriedades estilométricas. Caso apenas os tensores textuais fossem perturbados, mantendo-se fixos os vetores de características manuais, seria introduzida uma inconsistência lógico-matemática entre as duas modalidades de entrada, produzindo explicações potencialmente enviesadas e metodologicamente inválidas.

Para contornar essa limitação, foi desenvolvida uma função de encapsulamento universal responsável por sincronizar ambas as modalidades durante o processo de explicabilidade. Sempre que uma nova versão perturbada do texto é gerada pelo algoritmo de XAI, o sistema interrompe momentaneamente o processo de inferência e recalcula integralmente todas as características estilométricas associadas àquela nova instância textual. Em seguida, os novos vetores são normalizados utilizando os mesmos parâmetros estatísticos empregados durante o treinamento e, somente então, submetidos novamente ao classificador híbrido.

Esse mecanismo garante que cada amostra perturbada permaneça semantica e estilometricamente consistente, permitindo que as explicações reflitam adequadamente a interação entre o conteúdo linguístico e os padrões estruturais presentes no texto. Em outras palavras, a contribuição atribuída a um determinado *token* passa a considerar simultaneamente seu impacto semântico direto e as modificações indiretas provocadas sobre as características estilométricas derivadas.

Além disso, a integração entre os métodos de XAI e a arquitetura híbrida possibilita a construção de relatórios de auditoria capazes de responder questões como:

- Quais expressões ou padrões lexicais contribuíram para a classificação de um texto como humano;
- Quais estruturas linguísticas foram interpretadas como indícios de geração automática por modelos de linguagem;
- Quais características estilométricas exerceram maior influência na identificação de tentativas de *copy-typing*;
- Qual o grau de confiança e de incerteza associado à decisão produzida pelo sistema.

Para garantir a confiabilidade metodológica das análises, foram empregados três *frameworks* complementares de explicabilidade acompanhados pela confiança e entropia do modelo para análise de incerteza.

O primeiro desses métodos é o SHAP (*SHapley Additive exPlanations*). Fundamentado na Teoria dos Jogos Cooperativos, o SHAP calcula a contribuição marginal de cada característica para a decisão do modelo por meio dos valores de Shapley. A técnica permite analisar interações complexas entre palavras, padrões sintáticos e características estilométricas, produzindo explicações globalmente consistentes e teoricamente fundamentadas. No contexto deste trabalho, o método é empregado para identificar os atributos linguísticos e estruturais que apresentam maior poder discriminativo na separação entre textos humanos, sintéticos e produzidos por *copy-typing* (LUNDBERG; LEE, 2017).

Em paralelo, utilizou-se o LIME (*Local Interpretable Model-agnostic Explanations*), que produz explicações locais por meio da construção de modelos lineares aproximados na vizinhança da instância sob análise. Para cada texto auditado, o método gera centenas de perturbações controladas e estima a importância de cada característica na decisão final do classificador. Neste trabalho, foram utilizadas até 500 amostras perturbadas e um limite de 15 características explicativas por instância, permitindo visualizar, de forma intuitiva, quais palavras e construções linguísticas atraíram a predição para cada uma das classes consideradas (RIBEIRO; SINGH; GUESTRIN, 2016).

A terceira abordagem empregada foi o Integrated Gradients. Diferentemente dos métodos baseados em perturbação, essa técnica acessa diretamente os pesos internos da rede neural e calcula a contribuição individual de cada *token* a partir dos gradientes da função de saída em relação aos vetores de *embeddings*. Ao integrar as derivadas de um vetor nulo absoluto até o vetor do texto auditado, a técnica quantifica com alta precisão o grau de atenção e a contribuição individual de cada *token* processado (SUNDARARAJAN; TALY; YAN, 2017). Como resultado, é gerada uma visualização vetorial dos gradientes que se consolida como um laudo de auditoria robusto e detalhado.

Adicionalmente, de forma complementar à extração vetorial e visual fornecida pelos algoritmos acima, o motor de explicabilidade implementa o cálculo da Entropia de Shannon (SHANNON, 1948) sobre a distribuição final de probabilidades calibradas. Essa

métrica mensura o grau absoluto de incerteza global da inferência, permitindo o cruzamento analítico com a confiança matemática do modelo para a identificação e diagnóstico de predições limítrofes.

Por fim, a combinação entre métodos de explicabilidade local, análises globais de importância e medidas de incerteza preditiva constitui um mecanismo adicional de validação do sistema proposto, contribuindo para a transparência, confiabilidade e auditabilidade das decisões produzidas pelo classificador híbrido.

4 Resultados e Discussões

Neste capítulo, são apresentados e discutidos os resultados empíricos obtidos pelas arquiteturas de classificação propostas. A avaliação do desempenho computacional foi estruturada em duas frentes analíticas complementares: primeiramente, analisa-se a capacidade preditiva puramente quantitativa dos modelos por meio das matrizes de confusão nos conjuntos de validação e de teste. Em seguida, transita-se para uma avaliação qualitativa, dissecando o comportamento decisório da rede neural através dos algoritmos de explicabilidade, contemplando as análises globais (SHAP), locais (LIME e IG) e o diagnóstico de incerteza (Entropia de Shannon).

4.1 Desempenho no Conjunto de Validação

A avaliação primária da capacidade de convergência e separabilidade das classes ocorreu ao término do ciclo de treinamento (última época), antes do congelamento definitivo dos pesos. As matrizes de confusão geradas a partir do conjunto de validação (20% do *corpus* total) para cada um dos modelos treinados permitem diagnosticar o comportamento da rede diante de dados não vistos durante o treinamento, mas que integraram o ciclo de ajuste de hiperparâmetros e do *early stopping*.

A análise comparativa entre os laudos visuais das matrizes de confusão (Figuras 4.1 e 4.2) e os relatórios estatísticos (Tabelas 4.1, 4.2, 4.3 e 4.4) sugere um desempenho de excelência generalizada. Em todos os cenários e arquiteturas, a acurácia global manteve-se superior a 98,1%.

O primeiro e mais expressivo destaque empírico recai sobre a categoria “Humano”. No conjunto restrito, ambas as arquiteturas alcançaram revocação absoluta $(1,0000)$, eliminando completamente os falsos negativos. No conjunto estendido, mesmo diante de um *corpus* substancialmente mais diverso e ruidoso, o sistema cometeu apenas um único erro de classificação para a autoria humana. Tal resultado evidencia que as representações semânticas produzidas pelo BERTimbau são eficazes na caracterização da escrita humana,

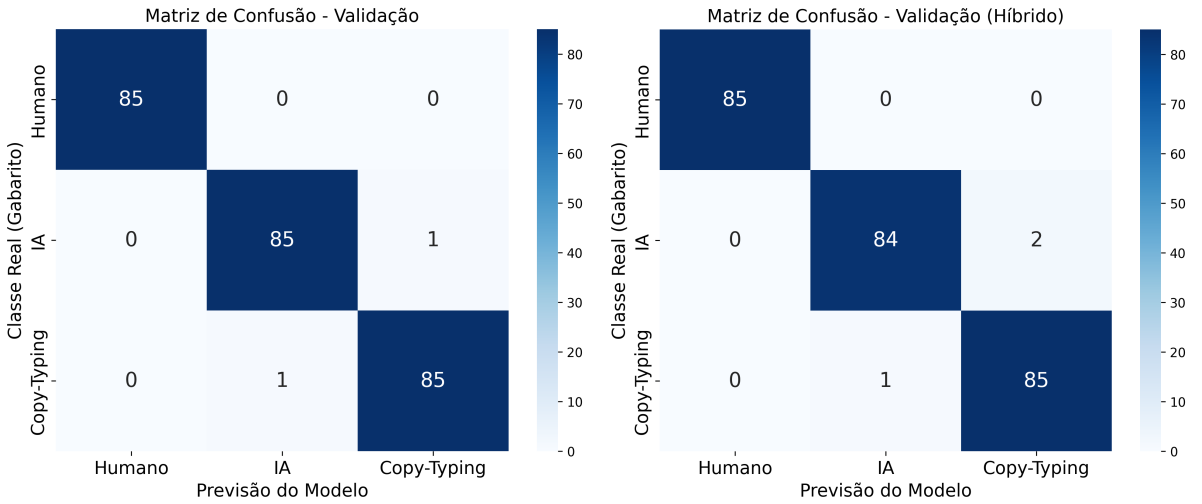


Figura 4.1: Comparação da matriz de confusão da arquitetura base (à esquerda) com a da arquitetura híbrida (à direita) no conjunto de validação restrito. Fonte: do autor.

Tabela 4.1: Relatório de métricas de classificação e análise de incerteza para a arquitetura base aplicada ao conjunto de validação restrito.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	1,0000	1,0000	85
IA Pura	0,9884	0,9884	0,9884	86
Copy-Typing	0,9884	0,9884	0,9884	86
Acurácia			0,9922	257
Média Macro	0,9922	0,9922	0,9922	257
Média Ponderada	0,9922	0,9922	0,9922	257
Análise de Incerteza				
Confiança Média Predita				99,35%
Confiança Média nos Erros				96,08%
Entropia Média (Shannon)				0,0346

Fonte: elaborado pelo autor.

reduzindo significativamente o risco de acusar indevidamente um estudante de utilizar inteligência artificial.

As divergências preditivas concentraram-se exclusivamente na fronteira de transição entre a geração computacional (“IA Pura”) e a intervenção mista (“Copy-Typing”). Ao observar as métricas globais, nota-se um padrão consistente: a arquitetura base apresenta um desempenho quantitativo marginalmente superior à arquitetura híbrida. No escopo restrito, a rede base obteve 99,22% de acurácia contra 98,83% da rede híbrida. Ao escalar para o escopo estendido, o padrão se manteve, com a rede base atingindo 98,65%

Tabela 4.2: Relatório de métricas de classificação e análise de incerteza para a arquitetura híbrida aplicada ao conjunto de validação restrito.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	1,0000	1,0000	85
IA Pura	0,9882	0,9767	0,9825	86
Copy-Typing	0,9770	0,9884	0,9827	86
Acurácia			0,9883	257
Média Macro	0,9884	0,9884	0,9884	257
Média Ponderada	0,9884	0,9883	0,9883	257
Análise de Incerteza				
Confiança Média Predita				99,00%
Confiança Média nos Erros				93,49%
Entropia Média (Shannon)				0,0581

Fonte: elaborado pelo autor.

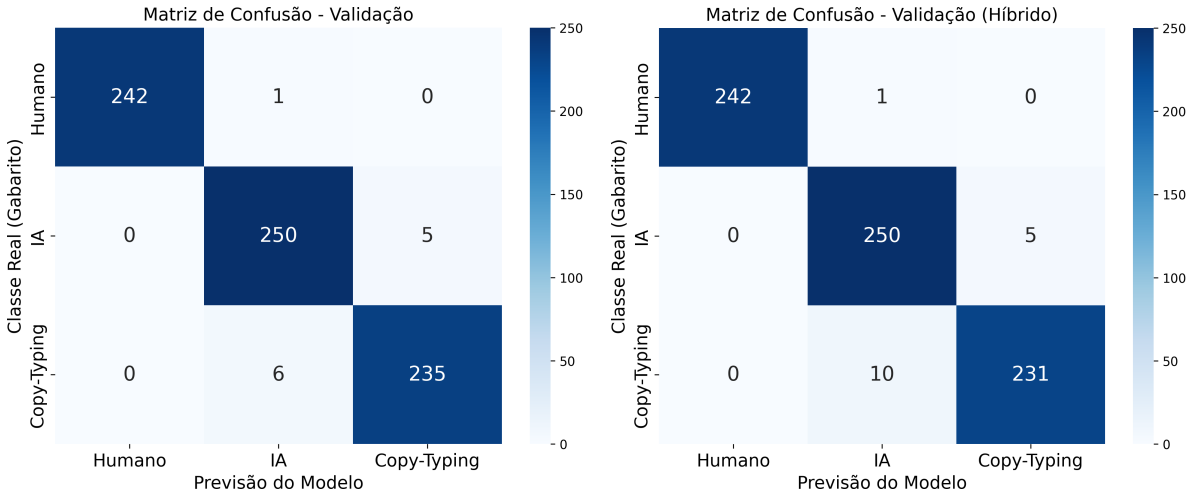


Figura 4.2: Comparação da matriz de confusão da arquitetura base (à esquerda) com a da arquitetura híbrida (à direita) no conjunto de validação estendida. Fonte: do autor.

contra 98,11% da híbrida. Essa pequena queda geral na precisão do conjunto estendido era teoricamente esperada, uma vez que a injeção de milhares de novas amostras expõe o modelo a táticas de ofuscação mais sofisticadas e ruídos de fronteira mais densos.

A análise probabilística revela, entretanto, uma dinâmica mais sutil. No conjunto restrito, a arquitetura híbrida apresentou menor confiança média nos próprios erros ($93,49\%$, contra $96,08\%$ do modelo base), acompanhada por um aumento da Entropia de Shannon média. Esse comportamento sugere uma maior hesitação diante de exemplos ambíguos, característica potencialmente desejável em aplicações de auditoria educacional, nas quais a sinalização de casos limítrofes é preferível à emissão de predições excessiva-

Tabela 4.3: Relatório de métricas de classificação e análise de incerteza para a arquitetura base aplicada ao conjunto de validação estendido.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	0,9959	0,9979	243
IA Pura	0,9804	0,9804	0,9804	255
Copy-Typing	0,9793	0,9834	0,9814	241
Acurácia			0,9865	739
Média Macro	0,9866	0,9866	0,9866	739
Média Ponderada	0,9865	0,9865	0,9865	739
Análise de Incerteza				
Confiança Média Predita				99,16%
Confiança Média nos Erros				86,82%
Entropia Média (Shannon)				0,0333

Fonte: elaborado pelo autor.

Tabela 4.4: Relatório de métricas de classificação e análise de incerteza para a arquitetura híbrida aplicada ao conjunto de validação estendido.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	0,9959	0,9979	243
IA Pura	0,9582	0,9882	0,9730	255
Copy-Typing	0,9872	0,9585	0,9726	241
Acurácia			0,9811	739
Média Macro	0,9818	0,9809	0,9812	739
Média Ponderada	0,9814	0,9811	0,9811	739
Análise de Incerteza				
Confiança Média Predita				98,87%
Confiança Média nos Erros				82,62%
Entropia Média (Shannon)				0,0432

Fonte: elaborado pelo autor.

mente confiantes.

No entanto, embora a avaliação no conjunto de validação permita identificar tendências comportamentais e formular hipóteses sobre o impacto das características estilométricas, somente a análise em um conjunto de teste completamente isolado é capaz de fornecer evidências mais robustas acerca da capacidade de generalização, da estabilidade probabilística e da utilidade prática das arquiteturas propostas.

4.2 Avaliação no Conjunto de Teste

Para assegurar a imparcialidade estatística e aferir a real capacidade de generalização do sistema no mundo real, os modelos finais foram submetidos ao conjunto de teste. Esse *dataset* foi mantido rigorosamente isolado de todas as fases anteriores de treinamento, calibração de limiares (*threshold tuning*) e ajuste de temperatura (*temperature scaling*).

Ao analisar as predições geradas nesse ambiente controlado, a avaliação dos modelos treinados com o *dataset* restrito revela limitações severas na capacidade de generalização das características sintáticas e semânticas, independentemente da arquitetura utilizada (base ou híbrida).

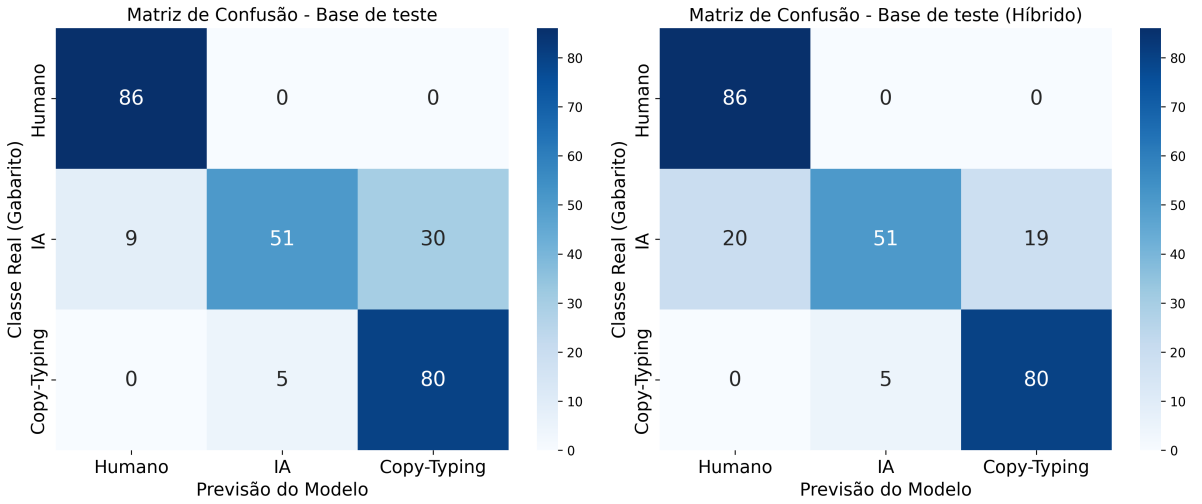


Figura 4.3: Comparação da matriz de confusão da arquitetura base (acima) com a da arquitetura híbrida (abaixo) no conjunto de teste. Fonte: do autor.

Observou-se uma clara estagnação de acurácia, com ambas as arquiteturas atingindo uma taxa idêntica de $83,14\%$. A análise das matrizes de confusão da Figura 4.3 e das Tabelas 4.5 e 4.6 expõe que o principal ponto de falha reside na classe IA Pura. A revocação para esta categoria foi de apenas 0,5667 em ambos os modelos, indicando que pouco mais da metade dos textos gerados por IA foram identificados corretamente. Além disso, nota-se uma grande dispersão de falsos negativos, uma vez que o modelo base confunde predominantemente a classe IA Pura com a categoria Copy-Typing, totalizando 30 amostras incorretamente classificadas, enquanto a arquitetura híbrida distribui esses erros entre as classes Humano (20 amostras) e Copy-Typing (19 amostras). Agravando esse cenário, a análise de incerteza revelou uma elevada confiança média nos erros, que se

Tabela 4.5: Relatório de métricas de classificação e análise de incerteza para o modelo treinado com *corpus* restrito e arquitetura base aplicados ao conjunto de teste.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	0,9053	1,0000	0,9503	86
IA Pura	0,9107	0,5667	0,6986	90
Copy-Typing	0,7273	0,9412	0,8205	85
Acurácia			0,8314	261
Média Macro	0,8478	0,8359	0,8231	261
Média Ponderada	0,8492	0,8314	0,8212	261
Análise de Incerteza				
Confiança Média Predita				97,68%
Confiança Média nos Erros				91,03%
Entropia Média (Shannon)				0,0729

Fonte: elaborado pelo autor.

Tabela 4.6: Relatório de métricas de classificação e análise de incerteza para o modelo treinado com *corpus* restrito e arquitetura híbrida aplicados ao conjunto de teste.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	0,8113	1,0000	0,8958	86
IA Pura	0,9107	0,5667	0,6986	90
Copy-Typing	0,8081	0,9412	0,8696	85
Acurácia			0,8314	261
Média Macro	0,8434	0,8359	0,8213	261
Média Ponderada	0,8445	0,8314	0,8193	261
Análise de Incerteza				
Confiança Média Predita				97,32%
Confiança Média nos Erros				90,65%
Entropia Média (Shannon)				0,0967

Fonte: elaborado pelo autor.

manteve acima de 90% (91,03% no modelo base e 90,65% no híbrido), configurando um quadro indesejável de excesso de confiança (*overconfidence*).

Em contrapartida, a transição para o *corpus* estendido demonstra que a principal barreira para o sucesso na classificação não era arquitetural, mas sim a representatividade dos dados de treinamento. Com a ampliação da base, houve um salto quantitativo em todas as métricas, e a acurácia global subiu para expressivos $98,08\%$ nas duas arquiteturas. As classes anteriormente problemáticas passaram a ser resolvidas com alta eficácia, permi-

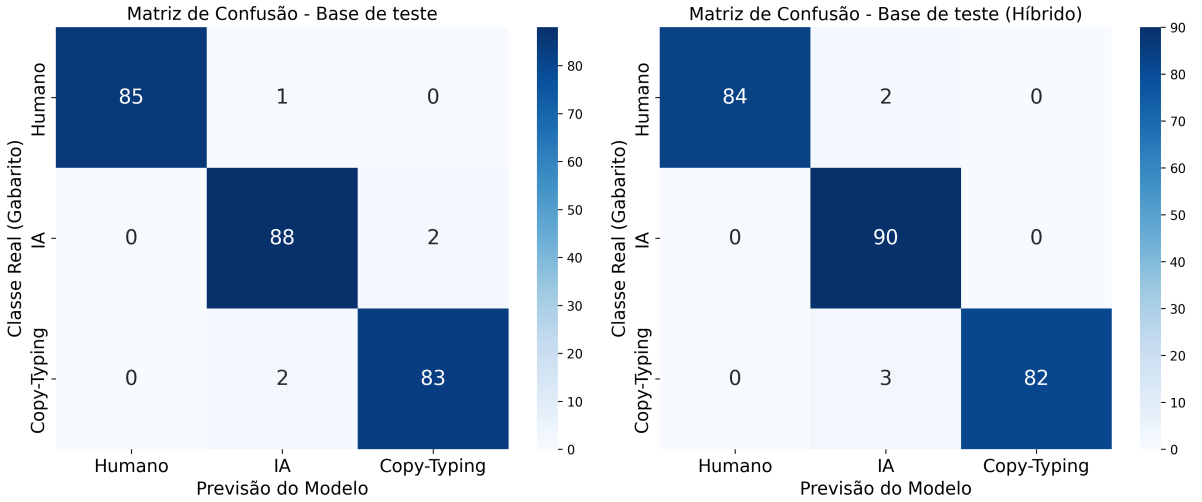


Figura 4.4: Comparação da matriz de confusão da arquitetura base (acima) com a da arquitetura híbrida (abaixo) no conjunto de teste. Fonte: do autor.

Tabela 4.7: Relatório de métricas de classificação e análise de incerteza para o modelo treinado com *corpus* estendido e arquitetura base aplicados ao conjunto de teste.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	0,9884	0,9942	86
IA Pura	0,9670	0,9778	0,9724	90
Copy-Typing	0,9765	0,9765	0,9765	85
Acurácia			0,9808	261
Média Macro	0,9812	0,9809	0,9810	261
Média Ponderada	0,9810	0,9808	0,9809	261
Análise de Incerteza				
Confiança Média Predita				99,01%
Confiança Média nos Erros				83,04%
Entropia Média (Shannon)				0,0372

Fonte: elaborado pelo autor.

tindo que a classe Humano atingisse F1-Scores superiores a 0,98 em ambas as variações. Conforme detalhado na Figura 4.4 e nas Tabelas 4.7 e 4.8, um dos grandes destaques dessa nova modelagem é a excelência na detecção de textos puramente sintéticos, o que demonstra uma evolução significativa na identificação de conteúdos sintéticos e minimiza drasticamente a ocorrência de falsos negativos.

Por fim, o aumento de dados provocou uma melhora drástica na qualidade das predições sob o ponto de vista probabilístico. A Entropia de Shannon média reduziu-se significativamente em relação aos modelos treinados com o conjunto restrito, refletindo

Tabela 4.8: Relatório de métricas de classificação e análise de incerteza para o modelo treinado com *corpus* estendido e arquitetura híbrida aplicados ao conjunto de teste.

Classe	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Humano	1,0000	0,9767	0,9882	86
IA Pura	0,9474	1,0000	0,9730	90
Copy-Typing	1,0000	0,9647	0,9820	85
Acurácia			0,9808	261
Média Macro	0,9825	0,9805	0,9811	261
Média Ponderada	0,9819	0,9808	0,9810	261
Análise de Incerteza				
Confiança Média Predita				98,71%
Confiança Média nos Erros				86,62%
Entropia Média (Shannon)				0,0439

Fonte: elaborado pelo autor.

decisões mais assertivas e com distribuições de probabilidade menos divididas. Adicionalmente, a confiança média nos erros sofreu uma queda natural e desejável. Isso indica que as raras falhas cometidas pelo modelo estendido ocorreram em um limiar de maior hesitação estatística, facilitando futuras políticas de rejeição de inferência ou reavaliação humana. Dessa forma, os resultados comprovam que a capacidade extrativa dos modelos dependia fundamentalmente de uma maior variância nos dados de entrada, fator que permitiu a construção de fronteiras de decisão mais robustas e com melhor capacidade de generalização.

4.3 Análise de Explicabilidade

Ultrapassada a validação estatística, fez-se necessário auditar a coerência lógica do aprendizado da rede. Para isso, os modelos foram submetidos a uma análise utilizando os algoritmos de explicabilidade anteriormente mencionados para garantir que as decisões não fossem baseadas em artefatos de *dataset* ou atalhos matemáticos (como o *Clever Hans effect*[14]), mas sim em padrões linguísticos genuínos.

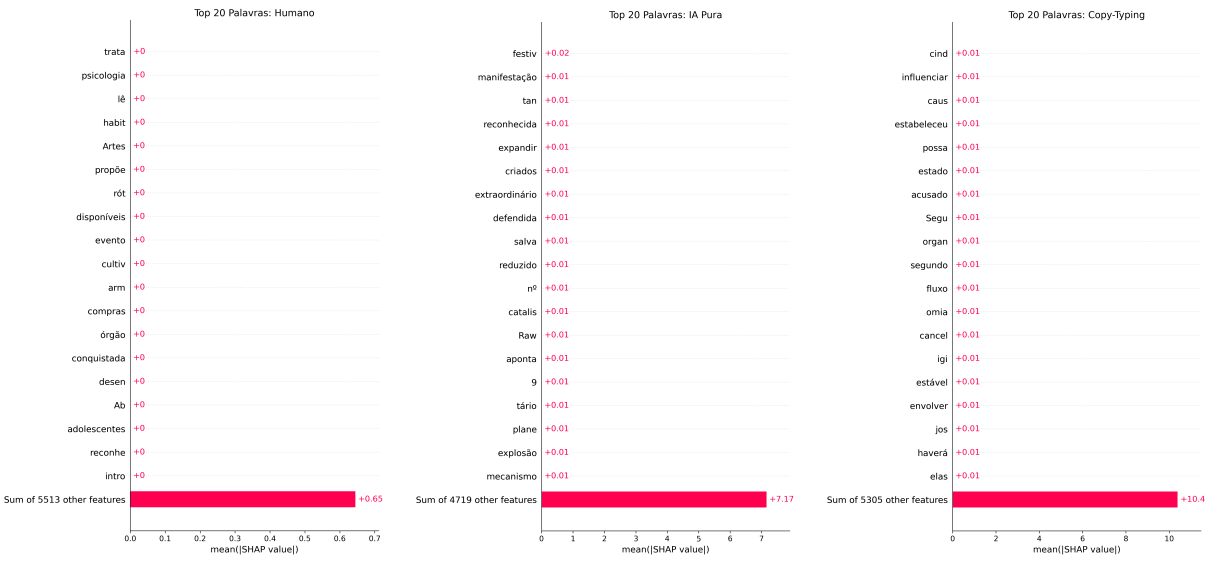
[14]O efeito *Clever Hans* em aprendizado de máquina descreve o fenômeno no qual um modelo atinge alto desempenho preditivo ao explorar correlações falsas ou pistas acidentais presentes nos dados de treinamento, em vez de aprender as representações subjacentes do problema real.

4.3.1 Análise Global de Atributos

A análise global de interpretabilidade foi conduzida com o objetivo de investigar a sensibilidade do modelo a perturbações no espaço de entrada textual, permitindo observar como diferentes *tokens* influenciam a saída preditiva em nível agregado. Para isso, foi utilizado o SHAP (*SHapley Additive exPlanations*), aplicado sobre o classificador baseado no BERTimbau, com função de predição probabilística calibrada.

No contexto deste trabalho, o SHAP não é interpretado como um mecanismo de inspeção interna do *transformer*, mas como uma aproximação de importância marginal de *tokens* sob um esquema de mascaramento. Assim, os valores atribuídos representam a variação esperada na saída do modelo quando segmentos textuais são removidos ou substituídos, permitindo inferir padrões de dependência entre entrada lexical e decisão preditiva.

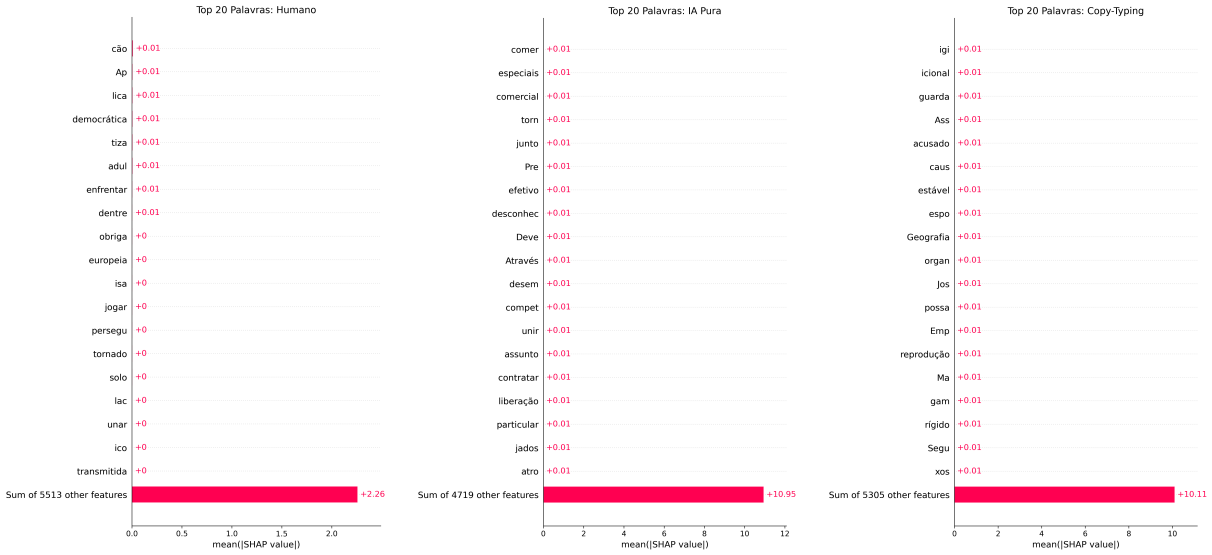
Para a análise global, nas Figuras 4.5 e 4.6 foram agregadas as contribuições médias dos *tokens* mais influentes por classe, considerando o conjunto de teste restrito e estendido. A visualização destaca os 20 *tokens* de maior influência, com o objetivo de evidenciar padrões predominantes de sensibilidade lexical sem comprometer a legibilidade da análise.



(a) Humano (b) IA Pura (c) Copy-Typing

Figura 4.5: Distribuição global de importância dos *tokens* (SHAP) para o modelo treinado com o *corpus* restrito. Fonte: do autor.

Os resultados indicam que não há concentração da importância em poucos *to-*



(a) Humano (b) IA Pura (c) Copy-Typing

Figura 4.6: Distribuição global de importância dos *tokens* (SHAP) para o modelo treinado com o *corpus* estendido. Fonte: do autor.

kens isolados. Em vez disso, observa-se uma distribuição difusa de contribuições, na qual múltiplos elementos lexicais apresentam impactos relativamente baixos quando considerados individualmente. Ainda assim, uma parcela significativa da contribuição global é agregada no termo *"Sum of other features"*, o que é consistente com a natureza de alta dimensionalidade do espaço de representação e com a tokenização, na qual a relevância tende a ser distribuída entre múltiplas unidades.

Adicionalmente, nota-se que muitos dos itens listados consistem em subpalavras ou fragmentos morfológicos (ex: "cind", "igi", "cão", "atro"), reflexo direto do algoritmo de tokenização utilizado pelo BERTimbau. Se o modelo estivesse sofrendo com vieses, como *Clever Hans effect*, possivelmente observaríamos palavras inteiras e contextualmente enviesadas dominando o topo do gráfico com altos valores numéricos de SHAP.

4.3.2 Análise Local de Atributos

Para aprofundar a explicabilidade além das magnitudes globais do SHAP, realizou-se uma inspeção minuciosa em nível de instância, também conhecida como explicabilidade local. Para tal, selecionou-se uma amostra representativa pertencente à classe Copy-typing e submeteu-se o texto a duas abordagens metodológicas distintas de interpretabilidade: LIME e o algoritmo *Integrated Gradients* da biblioteca Captum. A utilização simultânea

dessas duas ferramentas justifica-se pela natureza complementar de seus mecanismos de atribuição, permitindo uma auditoria em duas camadas distintas do modelo BERTimbau.

A primeira dessas abordagens é a análise por perturbação realizada pelo LIME. Por ser um método agnóstico, o LIME opera perturbando a entrada (ocultando ou mascarando palavras sequencialmente) e observando a variação na probabilidade de saída. Consequentemente, ele tende a destacar a importância semântica macroscópica de superfície. Na amostra correspondente à classe Copy-Typing (Figura 4.7), o LIME identificou que a decisão do modelo é fortemente guiada por um vocabulário formal e de caráter expositivo-argumentativo.

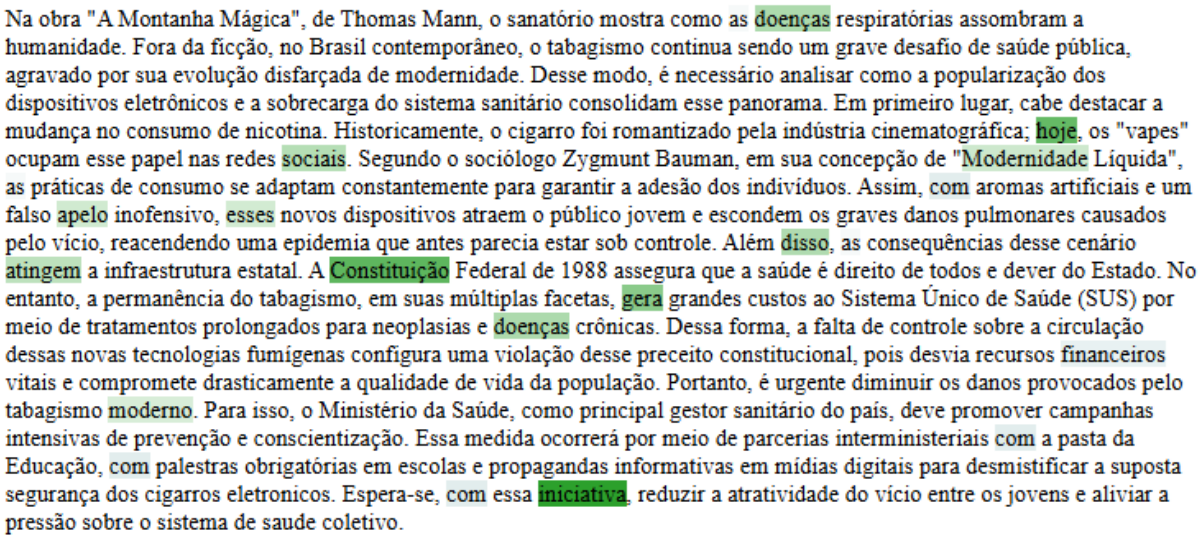

Na obra "A Montanha Mágica", de Thomas Mann, o sanatório mostra como as doenças respiratórias assombram a humanidade. Fora da ficção, no Brasil contemporâneo, o tabagismo continua sendo um grave desafio de saúde pública, agravado por sua evolução disfarçada de modernidade. Desse modo, é necessário analisar como a popularização dos dispositivos eletrônicos e a sobrecarga do sistema sanitário consolidam esse panorama. Em primeiro lugar, cabe destacar a mudança no consumo de nicotina. Historicamente, o cigarro foi romantizado pela indústria cinematográfica; hoje, os "vapes" ocupam esse papel nas redes sociais. Segundo o sociólogo Zygmunt Bauman, em sua concepção de "Modernidade Líquida", as práticas de consumo se adaptam constantemente para garantir a adesão dos indivíduos. Assim, com aromas artificiais e um falso apelo inofensivo, esses novos dispositivos atraem o público jovem e escondem os graves danos pulmonares causados pelo vício, reacendendo uma epidemia que antes parecia estar sob controle. Além disso, as consequências desse cenário atingem a infraestrutura estatal. A Constituição Federal de 1988 assegura que a saúde é direito de todos e dever do Estado. No entanto, a permanência do tabagismo, em suas múltiplas facetas, gera grandes custos ao Sistema Único de Saúde (SUS) por meio de tratamentos prolongados para neoplasias e doenças crônicas. Dessa forma, a falta de controle sobre a circulação dessas novas tecnologias fumígenas configura uma violação desse preceito constitucional, pois desvia recursos financeiros vitais e compromete drasticamente a qualidade de vida da população. Portanto, é urgente diminuir os danos provocados pelo tabagismo moderno. Para isso, o Ministério da Saúde, como principal gestor sanitário do país, deve promover campanhas intensivas de prevenção e conscientização. Essa medida ocorrerá por meio de parcerias interministeriais com a pasta da Educação, com palestras obrigatórias em escolas e propagandas informativas em mídias digitais para desmistificar a suposta segurança dos cigarros eletronicos. Espera-se, com essa iniciativa, reduzir a atratividade do vício entre os jovens e aliviar a pressão sobre o sistema de saude coletivo.

Figura 4.7: Visualização da atribuição de importância local gerada pelo LIME para uma amostra da classe *Copy-Typing*. Fonte: do autor.

Esse comportamento pode ser verificado na Figura 4.8 pela atribuição de maiores pesos positivos a *tokens* como “iniciativa” (0,29), “Constituição” (0,22), “hoje” (0,21) e “gera” (0,16), além do forte apelo a substantivos temáticos (“doenças”, “sociais”, “Modernidade”).

De forma complementar, a análise por *Integrated Gradients* (Figura 4.9) expande as atribuições do LIME ao revelar a sensibilidade da rede a blocos frasais e transições típicas do gênero dissertativo. Nessa etapa, destacam-se a ativação positiva (contribuição a favor da classe *Copy-Typing*) de conectivos, como “Em primeiro lugar” e “Fora da ficção”.

Além disso, o algoritmo evidencia o mapeamento característico do apelo à auto-

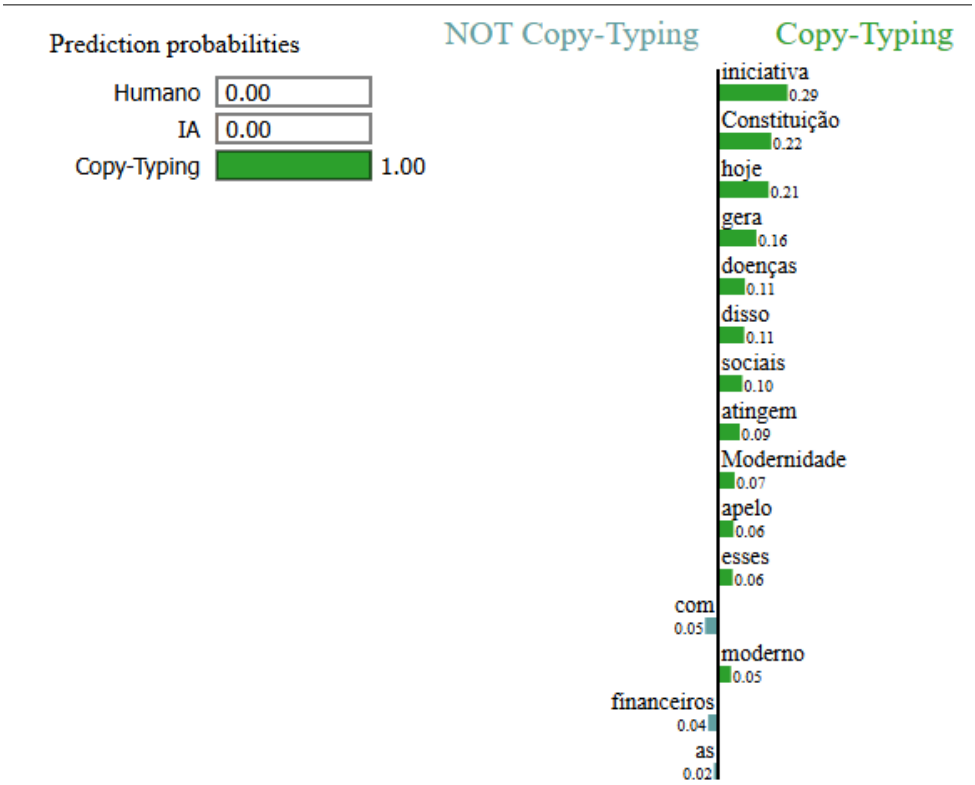


Figura 4.8: Análise da importância local gerada pelo LIME para uma amostra da classe *Copy-Typing*. Fonte: do autor.

[CLS] Na obra " A Mont ##anha Má ##gica " , de Thomas Mann , o san ##atório mostra como as doenças respira ##tórias asso ##mbra ##m a humanidade . Fora da ficção , no Brasil contemporâneo , o ta ##ba ##gi ##s ##mo continua sendo um grave desafio de saúde pública , agra ##vado por sua evolução disfar ##çada de modern ##idade . Desse modo , é necessário analisar como a popular ##ização dos dispositivos eletrônicos e a sobre ##car ##ga do sistema san ##itário consol ##ida ##m esse panorama . Em primeiro lugar , cabe destacar a mudança no consumo de nic ##ot ##ina . Histor ##icamente , o cig ##arro foi roma ##nti ##za ##do pela indústria cinematográfica ; hoje , os " va ##pes " ocupam esse papel nas redes sociais . Segundo o soci ##ólogo Z ##y ##g ##mun ##t Ba ##uman , em sua concepção de " Modern ##idade Lí ##qui ##da " , as práticas de consumo se adap ##tam constantemente para garantir a adesão dos indivíduos . Assim , com ar ##om ##as artificiais e um falso apelo ino ##fen ##si ##vo , esses novos dispositivos atra ##em o público jovem e esco ##ndem os graves danos pul ##monar ##es causados pelo ví ##cio , rea ##cend ##endo uma epidemia que antes parecia estar sob controle . Além disso , as consequências desse cenário atingem a infraestrutura estatal . A Constituição Federal de 1988 assegu ##ra que a saúde é direito de todos e dever do Estado . No entanto , a permanência do ta ##ba ##gi ##s ##mo , em suas múltiplas face ##tas , gera grandes custos ao Sistema Ú ##nico de Saúde (S ##US) por meio de tratamentos prolong ##ados para neo ##plas ##ias e doenças crônicas . Dessa forma , a falta de controle sobre a circulação dessas novas tecnologias fu ##m ##íg ##enas configu ##ra uma violação desse prece ##ito constitucional , pois desvia recursos financeiros vita ##is e comprome ##te drasticamente a qualidade de vida da população . Portanto , é ur ##gente diminuir os danos provocado ##s pelo ta ##ba ##gi ##s ##mo moderno . Para isso , o Ministério da Saúde , como principal gesto ##r san ##itário do país , deve promover campanhas intens ##ivas de prevenção e consci ##ent ##ização . Essa medida ocorrer ##á por meio de parcerias inter ##minist ##eriais com a pasta da Educação , com palestras obrigatória ##s em escolas e propaganda ##s informa ##tivas em mídias digitais para des ##mis ##tifica ##r a suposta segurança dos cig ##arro ##s eletro ##nico ##s . Espera - se , com essa iniciativa , reduzir a atra ##tividade do ví ##cio entre os jovens e aliv ##iar a pressão sobre o sistema de saud ##e coletivo . [SEP]

Figura 4.9: Visualização da atribuição de importância local gerada pelo algoritmo *Integrated Gradients* para uma amostra da classe *Copy-Typing*. Fonte: do autor.

ridade como no trecho “Segundo o soci ##ólogo” (a quebra da palavra sociólogo é um vestígio do processo de tokenização do modelo, mas que não impacta negativamente na avaliação). Essa ativação concentrada em locuções de transição e citação indica que o

modelo reconhece a estrutura padronizada que fundamenta o texto.

4.3.3 Considerações sobre os Limites da Explicabilidade

Em síntese, as abordagens de interpretabilidade empregadas desempenham papéis complementares, porém limitados, na análise do sistema. A análise baseada no SHAP não deve ser interpretada como explicação causal das decisões individuais do modelo, mas como uma medida de sensibilidade da saída em relação a perturbações na entrada textual, indicando não predominância observável de vieses de memorização ou atalhos lexicais no recorte analisado.

Por outro lado, os métodos locais (LIME e *Integrated Gradients*) fornecem explicações mais granulares em nível de instância, permitindo identificar padrões lexicais e estruturais associados às predições, embora ainda não sejam suficientes para uma descrição completa dos mecanismos internos do modelo.

A representação distribuída obtida por meio do *masked mean pooling* sobre as camadas do BERTimbau é consistente com o comportamento esperado de arquiteturas baseadas em *transformers*, nas quais a decisão resulta da composição de múltiplas interações contextuais.

Desse modo, as ferramentas de explicabilidade empregadas contribuem para uma análise qualitativa do comportamento do modelo, mas também evidenciam que a interpretabilidade de redes profundas ainda é parcialmente limitada, exigindo abordagens complementares para uma compreensão mais completa tanto em nível global quanto local.

5 Considerações Finais

O presente trabalho propôs o desenvolvimento de uma solução computacional para a detecção de fraudes educacionais, enfrentando o desafio imposto pela rápida evolução dos LLMs. Ao transpor o paradigma binário tradicional e modelar a detecção como um problema multiclasse, a pesquisa conseguiu capturar e classificar o comportamento híbrido e dissimulado do *copy-typing*.

O uso de um *corpus* estendido evidenciou que as representações semânticas obtidas por meio do codificador BERTimbau, em conjunto com a estratégia de *masked mean pooling*, apresentam boa capacidade de discriminação entre as classes analisadas. Nos experimentos realizados, os modelos atingiram acurácias globais superiores a 98% no conjunto de teste, com destaque para a elevada separabilidade da classe de autoria humana e a consequente redução de falsos positivos nesse grupo. Esses resultados indicam um comportamento consistente do sistema, estabelecendo uma base inicial para o desenvolvimento de ferramentas de auditorias acadêmicas reais.

Apesar do desempenho observado, a análise comparativa entre as arquiteturas indica que a abordagem de *late fusion* e a injeção de características estilométricas ainda constituem um espaço metodológico a ser melhor explorado, especialmente no que diz respeito à contribuição efetiva de cada métrica para o aprendizado do modelo.

Nesse contexto, tornam-se pertinentes os estudos de ablação, capazes de avaliar de forma controlada a contribuição individual de cada variável, quantificando matematicamente quais características favorecem a separabilidade das classes e quais atuam apenas como ruído. Ademais, a arquitetura híbrida pode ser expandida mediante a incorporação de métricas estruturais mais sofisticadas, como a análise de dependência sintática, concordância verbal e nominal, e índices de coesão referencial. Estas medidas são potencialmente mais sensíveis para capturar anomalias estruturais do *copy-typing* de forma mais contundente que a densidade de pontuação.

No que tange à XAI, as técnicas adotadas permitiram uma análise interpretativa útil do comportamento do modelo, contribuindo para a auditoria das decisões e para a

identificação de padrões relevantes em nível de . Entretanto, os métodos baseados em perturbação possuem limitações conhecidas ao lidar com as representações de alta dimensionalidade típicas de modelos baseados em *transformer*, podendo introduzir aproximações que não refletem integralmente o funcionamento interno da rede.

Como direcionamento para trabalhos futuros, sugere-se a exploração de abordagens de explicabilidade intrínsecas à arquitetura, capazes de analisar diretamente os mecanismos internos de atenção e os fluxos de representação latente, reduzindo a dependência de métodos externos baseados em perturbação. A integração dessas abordagens com análises baseadas em gradientes pode fornecer uma visão mais detalhada do processo decisório do modelo.

Em síntese, o trabalho alcança seu objetivo ao propor e avaliar um sistema de classificação com bom desempenho e capacidade de generalização dentro do escopo experimental definido. Ainda que os resultados sejam promissores, eles devem ser interpretados tendo em vista as limitações do *corpus* utilizado e as particularidades da tarefa, especialmente no que se refere à complexidade inerente à distinção entre autoria humana, geração automática e padrões híbridos de escrita.

Bibliografia

BROWN, T. B. et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. Disponível em: ⟨https://arxiv.org/abs/2005.14165⟩.

CHAKRABORTY, S. et al. *On the Possibilities of AI-Generated Text Detection*. 2023. Disponível em: ⟨https://arxiv.org/abs/2304.04736⟩.

COTTON, D. R. E.; COTTON, P. A.; SHIPWAY, J. R. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, Routledge, v. 61, n. 2, p. 228–239, 2024. Disponível em: ⟨https://doi.org/10.1080/14703297.2023.2190148⟩.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: ⟨http://arxiv.org/abs/1810.04805⟩.

GUO, C. et al. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017. Disponível em: ⟨http://arxiv.org/abs/1706.04599⟩.

HE, P. et al. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020. Disponível em: ⟨https://arxiv.org/abs/2006.03654⟩.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *A Redação no Enem 2025: Cartilha do Participante*. Brasília, DF, 2025. Disponível em: ⟨https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/enem/enem-2025-cartilha-da-redacao-esta-disponivel⟩.

KASNECI, E. et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, v. 103, p. 102274, 2023. ISSN 1041-6080. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S1041608023000195⟩.

LAN, Z. et al. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. Disponível em: ⟨http://arxiv.org/abs/1909.11942⟩.

LIU, D. C.; NOCEDAL, J. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, Springer, v. 45, n. 1, p. 503–528, 1989.

LIU, Y. et al. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. Disponível em: ⟨http://arxiv.org/abs/1907.11692⟩.

LO, C. K. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, MDPI, v. 13, n. 4, p. 410, 2023. Disponível em: ⟨https://www.mdpi.com/2227-7102/13/4/410⟩.

LOSHCHILOV, I.; HUTTER, F. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. Disponível em: ⟨http://arxiv.org/abs/1711.05101⟩.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. Disponível em: ⟨http://arxiv.org/abs/1705.07874⟩.

MARINHO, J.; ANCHIêTA, R.; MOURA, R. Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, Sociedade Brasileira de Computação, v. 13, n. 1, p. 65–76, 2022. Disponível em: ⟨https://sol.sbc.org.br/journals/index.php/jidm/article/view/2340⟩.

MCCARTHY, P. M.; JARVIS, S. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, Springer, v. 42, n. 2, p. 381–392, 2010.

NIU, C. et al. Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. In: DERNONCOURT, F.; PREOŢIUC-PIETRO, D.; SHIMORINA, A. (Ed.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami, Florida, US: Association for Computational Linguistics, 2024. p. 940–953. Disponível em: ⟨https://aclanthology.org/2024.emnlp-industry.70/⟩.

RAHMAN, M. M.; WATANOBE, Y. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, MDPI, v. 13, n. 9, p. 5783, 2023. Disponível em: ⟨https://www.mdpi.com/2076-3417/13/9/5783⟩.

REGE, M.; YARMOLUK, D. *The Impact of Artificial Intelligence and ChatGPT on Education*. 2023. Disponível em: ⟨https://news.stthomas.edu/the-impact-of-artificial-intelligence-and-chatgpt-on-education/⟩.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. Disponível em: ⟨http://arxiv.org/abs/1602.04938⟩.

RODRIGUES, J. et al. Advancing neural encoding of portuguese with transformer albertina pt-*. In: _____. *Progress in Artificial Intelligence*. Springer Nature Switzerland, 2023. p. 441–453. ISBN 9783031490088. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-031-49008-8_35⟩.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020. p. 403–417. ISBN 978-3-030-61376-1. Disponível em: ⟨https://doi.org/10.1007/978-3-030-61377-8_28⟩.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. Disponível em: ⟨http://arxiv.org/abs/1703.01365⟩.

SWELLER, J. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, v. 12, n. 2, p. 257–285, 1988. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_4⟩.

VASWANI, A. et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: ⟨http://arxiv.org/abs/1706.03762⟩.

WANG, Q.; LI, H. On continually tracing origins of llm-generated text and its application in detecting cheating in student coursework. *Big Data and Cognitive Computing*, v. 9, n. 3, 2025. ISSN 2504-2289. Disponível em: ⟨https://www.mdpi.com/2504-2289/9/3/50⟩.

WEBER-WULFF, D. et al. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, Springer Science and Business Media LLC, v. 19, n. 1, dez. 2023. ISSN 1833-2595. Disponível em: ⟨http://dx.doi.org/10.1007/s40979-023-00146-z⟩.

XIE, Y.; WU, S.; CHAKRAVARTY, S. Ai meets ai: Artificial intelligence and academic integrity - a survey on mitigating ai-assisted cheating in computing education. In: *Proceedings of the 24th Annual Conference on Information Technology Education*. New York, NY, USA: Association for Computing Machinery, 2023. (SIGITE '23), p. 79–83. ISBN 9798400701306. Disponível em: ⟨https://doi.org/10.1145/3585059.3611449⟩.