Análise Estilométrica de Textos Humanos e de LLMs Usando Métodos Estatísticos

Victor Löfgren Sattamini
Programa de Pós-Graduação em Ciências Computacionais e Modelagem Matemática (PPG-CompMat)
IME UERJ

11 de Dezembro de 2025

Resumo

A detecção de textos gerados por modelos de linguagem de grande porte (LLMs) tornou-se uma preocupação crescente em contextos acadêmicos, educacionais e de moderação de conteúdo. Este trabalho apresenta uma primeira análise estilométrica para detecção de textos gerados por LLMs em português do Brasil. Utilizamos um corpus balanceado de 100.000 amostras (50.000 autorais, 50.000 de LLMs) extraídas de múltiplas fontes, incluindo BrWaC, ShareGPT-Portuguese e Canarim. Aplicamos 10 características estilométricas (comprimento médio de frases, relação tipo-token, entropia de caracteres, burstiness, entre outras) e realizamos testes não paramétricos (Mann-Whitney U) com correção FDR e análise de tamanho de efeito (delta de Cliff). Seis características apresentaram efeitos grandes ($|\delta| \geq 0,474$), sendo a entropia de caracteres a mais discriminante ($\delta = -0,881$). Aplicamos análise de componentes principais (PCA) e dois classificadores lineares: análise discriminante linear (LDA) e regressão logística, ambos avaliados em validação cruzada estratificada de 5 folds. A regressão logística alcançou ROC AUC de 97,03% ($\pm 0,14\%$), enquanto a LDA obteve 94,12% ($\pm 0,17\%$). Os resultados demonstram que métodos estatísticos clássicos são altamente eficazes para distinguir textos autorais de LLMs em português, confirmando achados anteriores em inglês e estendendo-os para outro idioma. Identificamos padrões contra-intuitivos: textos autorais são mais variáveis estruturalmente (maior burstiness e entropia), enquanto LLMs são mais diversos lexicalmente (maior TTR e proporção de hapax). Este trabalho estabelece uma base sólida para detecção estilométrica de LLMs em português e demonstra que assinaturas estilísticas humanas permanecem detectáveis através de análise estatística.

1 Introdução

A emergência de modelos de linguagem de grande porte (LLMs) criou preocupações quanto à detecção de conteúdo gerado automaticamente. A detecção de autoria computacional tem raízes históricas sólidas, iniciando com o trabalho seminal de Mosteller e Wallace (**mosteller1964**) sobre os artigos Federalistas e posteriormente formalizada por Burrows (**burrows2002**) com a medida Delta para diferenciação estilística. Trabalhos recentes demonstram que essas técnicas estilométricas clássicas permanecem eficazes para distinguir textos autorais de textos gerados por LLMs (**stamatatos2009**; **huang2024**).

Estudos em múltiplos idiomas confirmam a viabilidade da abordagem estilométrica: Herbold et al. (**stylometric˙llm˙detection**) reportaram 81–98% de acurácia usando 31 características e floresta aleatória; Zaitsu e Jin (**zaitsu2023**) alcançaram 100% de precisão em textos japoneses;

Przystalski et al. (**przystalski2025**) demonstraram que estilometria reconhece LLMs mesmo em pequenas amostras (0,87–0,98 de acurácia); e Berriche e Larabi-Marie-Sainte (**berriche2024**) atingiram 100% usando 33 características estilométricas com *XGBoost*. Esses resultados evidenciam que características como comprimento médio de frases, relação tipo-token, entropia de caracteres (**shannon1948**), proporção de palavras funcionais (**stamatatos2009**) e burstiness (variação estrutural) (**gptzero2023**; **chakraborty2023ct2**) contêm sinais fortes sobre a origem do texto.

Este estudo contribui para a literatura de detecção de LLMs ao fornecer uma primeira análise estilométrica para detecção de textos gerados por LLMs em português do Brasil. Não foi encontrado aplicação de análise estilométrica a textos de LLM em português. Utilizou-se um conjunto de dados balanceado com mais de 1,2 milhões de amostras de múltiplas fontes (BrWaC (**brwac**), ShareGPT-Portuguese (**sharegpt˙portuguese**), Canarim (**canarim**)).[1]

1.1 Mineração de Texto

A mineração de texto é o processo de extração de informação relevante e conhecimento a partir de dados textuais não estruturados (**feldman2007**). Diferentemente da análise de dados tabulares tradicionais, a mineração de texto requer a transformação de documentos em representações numéricas que possibilitem a aplicação de métodos estatísticos.

O processo de mineração de texto compreende quatro etapas fundamentais:

1. **Coleta de dados**: Aquisição de documentos textuais de fontes diversas, garantindo representatividade da população de interesse.

2. **Pré-processamento**: Limpeza e normalização dos textos, incluindo remoção de caracteres especiais, normalização de espaços em branco, e conversão para codificação uniforme (UTF-8).

3. **Extração de características**: Transformação dos documentos em vetores de variáveis quantitativas mensuráveis. Esta etapa é crucial pois define as variáveis que serão analisadas estatisticamente.

4. **Análise estatística**: Aplicação de métodos estatísticos descritivos e inferenciais sobre as variáveis extraídas para identificar padrões, diferenças entre grupos, e construir modelos preditivos.

No contexto deste trabalho, a mineração de texto serve como ponte entre documentos textuais brutos e a análise estatística formal. As características extraídas (descritas na Seção 2.3) são variáveis quantitativas mensuradas em escalas de razão ou intervalo, permitindo a aplicação de métodos estatísticos paramétricos e não paramétricos.

1.2 Estilometria e Análise de Autoria

A estilometria é o estudo quantitativo do estilo linguístico através da medição de características objetivas dos textos (**stamatatos2009**). Fundamenta-se no princípio de que autores possuem padrões linguísticos inconscientes e consistentes que podem ser identificados estatisticamente.

[1] Desde a compilação deste corpus, novos recursos em português surgiram, incluindo GigaVerbo com 200B tokens (**correa2024**) e PTT5-v2 (**piau2024**), que podem beneficiar trabalhos futuros.

1.2.1 Fundamentos da Análise Estilométrica

A análise estilométrica baseia-se em três premissas fundamentais:

1. **Consistência autoral**: Autores humanos mantêm padrões estilísticos relativamente estáveis ao longo de diferentes textos e tópicos.
2. **Variabilidade inter-autoral**: As diferenças estilísticas entre autores distintos são maiores que as variações intra-autorais.
3. **Mensurabilidade**: Características estilísticas podem ser quantificadas através de variáveis mensuráveis objetivamente.

O trabalho seminal de **mosteller1964** sobre os *Federalist Papers* demonstrou que métodos estatísticos rigorosos podem atribuir autoria com alta confiança. A abordagem foi posteriormente formalizada por **burrows2002** com a medida Delta, que utiliza distâncias estatísticas entre perfis estilométricos.

1.2.2 Características Estilométricas

As variáveis estilométricas utilizadas em análise de autoria podem ser categorizadas conforme suas escalas de medida:

Variáveis em escala de razão (possuem zero absoluto e razões interpretáveis):

- Comprimento médio de frase (palavras por frase)
- Frequência de uso de pontuação específica (por 1000 palavras)
- Riqueza lexical (razão tipo-token)
- Proporções de classes gramaticais (substantivos, verbos, etc.)

Variáveis em escala de intervalo (diferenças interpretáveis, mas sem zero absoluto):

- Entropia de distribuição de caracteres (**shannon1948**)
- Coeficiente de variação do comprimento de frase (*burstiness* normalizado) (**gptzero2023**; **chakraborty2023ct2**)

A distinção entre escalas de medida é fundamental porque determina quais métodos estatísticos são aplicáveis. Variáveis em escala de razão permitem operações aritméticas completas e cálculo de medidas como média geométrica e coeficiente de variação. Variáveis em escala de intervalo permitem cálculo de médias e desvios padrão, mas não razões.

1.2.3 Detecção de Textos Gerados por LLMs

Estudos recentes demonstram que técnicas estilométricas clássicas permanecem eficazes para detectar textos gerados por modelos de linguagem de grande porte (**stylometric˙llm˙detection**; **stamatatos2009**). **stylometric˙llm˙detection** reportaram acurácia superior a 99% utilizando características estilométricas simples em amostras curtas (100-200 palavras).

Trabalhos específicos para o português incluem **berriche2024**, que demonstraram a eficácia de medidas de entropia de caracteres e proporção de palavras funcionais. O presente trabalho estende essa linha de pesquisa aplicando métodos estatísticos multivariados a um conjunto abrangente de características estilométricas em português do Brasil.

1.3 Justificativa para Múltiplos Métodos Estatísticos

Este trabalho aplica três métodos multivariados complementares (PCA, LDA, Regressão Logística) por razões metodológicas distintas, não redundantes:

1. **PCA - Análise Exploratória**: Método não supervisionado para visualização de estrutura natural dos dados e identificação de padrões sem conhecimento prévio das classes. Responde: *"As variáveis se agrupam naturalmente por categoria (humano/LLM) sem supervisão?"*

2. **LDA - Discriminação Ótima**: Método supervisionado que maximiza separação entre grupos conhecidos. Enquanto PCA maximiza variância total, LDA maximiza variância *between-group* relativa à *within-group*. Responde: *"Qual combinação linear de variáveis melhor discrimina os grupos?"*

3. **Regressão Logística - Modelagem Preditiva**: Método probabilístico que quantifica contribuição individual de cada variável e permite interpretação através de odds ratios. Responde: *"Qual a probabilidade de um novo texto ser humano dado seu perfil estilométrico?"*

Complementaridade metodológica:

- PCA é **descritivo** (sem hipóteses)
- LDA é **discriminativo** (maximiza separação)
- Regressão Logística é **preditivo e inferencial** (estima probabilidades e testa significância)

Esta abordagem triangulada fortalece as conclusões: se os três métodos independentes convergem para as mesmas variáveis como importantes, aumenta a confiança na robustez dos achados.

2 Métodos

2.1 Mineração de Texto e Pré-processamento

A mineração de texto consiste em extrair informações úteis de dados textuais não estruturados através de técnicas estatísticas e computacionais (**feldman2007**). O processo envolve etapas de coleta, pré-processamento (limpeza, tokenização, normalização), extração de características numéricas e aplicação de métodos analíticos. Neste trabalho, aplicamos mineração de texto para transformar documentos em vetores de variáveis quantitativas que capturam propriedades estatísticas do estilo de escrita, permitindo análise estatística inferencial e construção de modelos de classificação.

2.2 Conjunto de Dados

Utilizou-se um conjunto de dados textuais balanceado em português do Brasil contendo 100.000 amostras (50.000 autorais, 50.000 de LLMs), extraídas por amostragem estratificada de um conjunto maior com 2.331.317 documentos originais provenientes de 5 fontes distintas. As fontes de texto autoral incluem: (i) BrWaC (Brazilian Web as Corpus) (**brwac**), um grande conjunto web de textos brasileiros; e (ii) BoolQ (**boolq**), contendo passagens de contexto para perguntas booleanas. As fontes de texto gerado por LLM incluem: (i) ShareGPT-Portuguese (**sharegpt˙portuguese**), conversas em português extraídas da plataforma ShareGPT; (ii) resenhas do IMDB traduzidas para português por modelos de tradução automática (classificadas como texto LLM); e (iii) o dataset Canarim (**canarim**), contendo saídas geradas por LLMs.

2.2.1 Método de Amostragem Estratificada

A amostragem foi realizada através de **amostragem aleatória estratificada proporcional** com estratificação por fonte de origem dos textos. Este método garante representatividade de cada fonte na amostra final.

Procedimento:

1. **Definição de estratos**: A população foi dividida em $L = 5$ estratos correspondentes às fontes:
   - Estrato 1: BrWaC (textos web humanos)
   - Estrato 2: BoolQ traduzido (textos humanos)
   - Estrato 3: ShareGPT-Portuguese (LLM conversacional)
   - Estrato 4: IMDB traduzido (LLM)
   - Estrato 5: Canarim-Instruct (LLM instrucional)
2. **Cálculo dos tamanhos amostrais por estrato**: Para amostragem proporcional com tamanho total $n = 100.000$:
   $$n_h = n \times \frac{N_h}{N}$$
   onde N_h é o tamanho populacional do estrato h e $N = \sum_{h=1}^{L} N_h = 2.331.317$ é o tamanho populacional total.
3. **Seleção aleatória simples dentro de cada estrato**: Utilizamos `numpy.random.choice` com semente fixa (42) para reprodutibilidade, sem reposição.
4. **Combinação das amostras estratificadas**: A amostra final é a união $\bigcup_{h=1}^{L} s_h$ onde s_h é a amostra do estrato h.

Vantagens da estratificação:

- **Representatividade**: Garante presença de todas as fontes proporcionalmente ao tamanho populacional
- **Redução de variância**: A variância da estimativa é menor que na amostragem aleatória simples quando há heterogeneidade entre estratos
- **Estimativas por estrato**: Permite análises separadas por fonte quando necessário

Justificativa estatística: A estratificação por fonte é apropriada pois diferentes fontes podem ter características textuais distintas (e.g., BrWaC contém textos web informais; Canarim contém instruções formais). A amostragem proporcional mantém a distribuição populacional original, evitando viés de seleção.

Os textos foram previamente filtrados por comprimento mínimo de 100 caracteres e máximo de 10.000 caracteres, sendo textos muito longos segmentados em fragmentos de até 10.000 caracteres sem sobreposição. A segmentação priorizou quebras naturais de texto (pontos finais, parágrafos e espaços). O balanceamento foi obtido por subamostragem da classe majoritária e sobreamostragem da classe minoritária, resultando em proporções exatamente iguais (50%/50%). A amostra de 100.000 documentos foi selecionada aleatoriamente com semente fixa (`seed=42`) para reprodutibilidade.

Para prevenir vazamento de dados, verificamos que os textos não apresentam agrupamentos estruturais por autor, tópico ou sessão de geração. A validação cruzada estratificada mantém o balanço de classes entre as partições, garantindo amostras independentes em conjuntos de treino e teste. Esta abordagem evita viés de avaliação documentado em estudos anteriores (**kohavi1995**).

2.3 Extração de Características Estilométricas

Foram extraídas 10 características estilométricas de cada documento, todas representando variáveis contínuas. A escolha dessas características baseia-se em estudos anteriores que demonstraram sua eficácia na análise de autoria (**stamatatos2009**; **stylometric˙llm˙detection**).

2.3.1 Variáveis em Escala de Razão

As nove características a seguir são mensuradas em **escala de razão**, possuindo zero absoluto e permitindo interpretação de razões:

1. **Comprimento médio de frase** (`sent_mean`): Média aritmética do número de palavras por frase. Unidade: palavras/frase. Zero representa ausência de palavras.

2. **Desvio padrão do comprimento de frase** (`sent_std`): Medida de dispersão absoluta do comprimento de frases. Unidade: palavras. Quantifica a variabilidade no comprimento das frases.

3. **Coeficiente de variação do comprimento de frase** (`sent_cv`): Razão entre desvio padrão e média ($CV = \sigma/\mu$). Estatística adimensional que normaliza a variabilidade pela tendência central, permitindo comparação entre distribuições com escalas distintas. Esta métrica, também denominada *burstiness* normalizado no contexto de detecção de textos gerados por LLMs, captura a variação nas estruturas das sentenças – textos humanos tendem a alternar entre frases longas e curtas, enquanto LLMs produzem comprimentos mais uniformes (**gptzero2023**; **siddharth2024burstiness**; **chakraborty2023ct2**).

4. **Riqueza lexical - C de Herdan** (`herdan_c`): Medida de diversidade vocabular calculada como $C = \log(V)/\log(N)$, onde V é o número de tipos (palavras distintas) e N é o número de tokens (total de palavras) (**herdan1960**). Varia entre 0 e 1, onde valores próximos a 1 indicam maior diversidade lexical.

5. **Relação tipo-token** (`ttr`): Razão entre número de tipos (palavras distintas) e tokens (total de palavras), calculada como $TTR = V/N$. Adimensional, varia entre 0 e 1. Medida clássica de riqueza lexical – valores altos indicam maior diversidade vocabular.

6. **Proporção de hapax legomena** (`hapax_prop`): Proporção de palavras que ocorrem exatamente uma vez no texto. Adimensional, varia entre 0 e 1. Hapax legomena são palavras raras que indicam riqueza vocabular e especificidade do texto.

7. **Proporção de palavras funcionais** (`func_word_ratio`): Razão entre palavras funcionais (artigos, preposições, conjunções, pronomes) e total de palavras (**stamatatos2009**). Adimensional, varia entre 0 e 1. Palavras funcionais são frequentes e pouco conscientes, revelando estilo autoral.

8. **Proporção de primeira pessoa** (`first_person_ratio`): Razão entre pronomes de primeira pessoa (eu, me, meu, nos, nosso, etc.) e total de palavras. Adimensional, varia entre 0 e 1. Indica subjetividade e perspectiva pessoal do texto.

9. **Taxa de repetição de bigramas** (`bigram_repeat_ratio`): Proporção de tipos de bigramas (pares consecutivos de palavras) que ocorrem mais de uma vez, calculada como (número de tipos de bigramas com contagem ¿ 1) / (total de tipos distintos de bigramas). Adimensional, varia entre 0 e 1. Valores altos indicam maior repetição de padrões frasais. Esta métrica captura redundância local, relacionada ao uso de características de bigramas para detecção de textos gerados por modelos de linguagem (**solaiman2019release**; **li2016diversity**).

2.3.2 Variável em Escala de Intervalo

10. **Variabilidade da distribuição de caracteres** (`char_entropy`): Medida de dispersão na distribuição de frequências de caracteres, calculada pela fórmula de Shannon $H = -\sum_c p(c) \log_2 p(c)$ (**shannon1948**), onde $p(c)$ é a probabilidade de ocorrência do caractere c.

    Esta medida quantifica a variabilidade: alta entropia indica distribuição mais uniforme (maior dispersão); baixa entropia indica concentração (menor dispersão).

    Justificativa estatística: Embora originalmente uma medida da teoria da informação, a entropia funciona como **medida de dispersão análoga ao desvio padrão**, mas aplicada a distribuições de frequência categórica. A entropia é mensurada em **escala de intervalo** porque:

    - Diferenças entre valores são interpretáveis (aumento de 1 bit representa dobrar a incerteza)
    - Não possui zero absoluto natural (zero ocorre apenas com um único caractere)
    - Razões entre valores não são estatisticamente interpretáveis

2.3.3 Justificativa da Escolha das Características

Todas as características foram selecionadas por três critérios:

1. **Objetividade**: Mensuração automática e determinística, sem julgamento subjetivo.
2. **Robustez**: Insensibilidade a pequenas variações no texto ou erros de tokenização.
3. **Fundamentação teórica**: Suporte empírico na literatura de estilometria para distinção de autoria.

A combinação de variáveis em escala de razão e intervalo permite aplicação de métodos estatísticos diversos. As variáveis de razão satisfazem requisitos para testes paramétricos quando distribuídas normalmente. A variável de entropia, sendo contínua em escala de intervalo, pode ser incluída em análises multivariadas que não assumem proporcionalidade (como PCA e regressão logística).

2.4 Testes Estatísticos Não Paramétricos

A escolha de métodos não paramétricos foi determinada pelas características das distribuições observadas nos dados, seguindo os critérios estabelecidos por Siegel e Castellan (**siegel1988**) e Hollander, Wolfe e Chicken (**hollander2013**).

2.4.1 Justificativa para Métodos Não Paramétricos

Após análise exploratória inicial, identificamos três violações aos pressupostos de testes paramétricos:

1. **Não normalidade**: Testes de Shapiro-Wilk ($\alpha = 0.05$) rejeitaram a hipótese de normalidade para 8 das 10 variáveis em ambos os grupos (humano e LLM).
2. **Heterocedasticidade**: Teste de Levene indicou variâncias significativamente diferentes entre grupos para 6 variáveis ($p < 0.01$).
3. **Presença de valores atípicos**: Boxplots revelaram valores atípicos (*outliers*) em 7 das 10 variáveis, com alguns valores extremos além de 3 desvios padrão da média.

Dadas essas violações, métodos não paramétricos são mais apropriados pois:

- Não assumem forma específica de distribuição
- São robustos a valores atípicos (*outliers*) (baseiam-se em postos, não valores brutos)
- Mantêm poder estatístico adequado com distribuições não normais

2.4.2 Teste de Mann-Whitney U

Para comparar as distribuições de cada variável entre textos humanos e LLM, utilizamos o teste de Mann-Whitney U (**mann1947**), também conhecido como teste de Wilcoxon para amostras independentes.

Hipóteses:

- H_0: As distribuições das duas populações são idênticas
- H_1: As distribuições diferem em localização (mediana)

Estatística do teste:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

onde n_1 e n_2 são os tamanhos amostrais, e R_1 é a soma dos postos do grupo 1.

Interpretação: Valores pequenos de U (ou valores-p menores que α) indicam evidência contra H_0, sugerindo que as distribuições diferem sistematicamente.

2.4.3 Tamanho de Efeito: Delta de Cliff

O valor-p indica apenas se há diferença estatisticamente detectável, não sua magnitude prática. Portanto, calculamos o Delta de Cliff (δ) (**cliff1993**) como medida de tamanho de efeito:

$$\delta = \frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_1 \times n_2}$$

onde x_i são observações do grupo 1 e y_j do grupo 2.

Interpretação (**romano2006**):

- $|\delta| < 0.147$: Efeito negligenciável
- $0.147 \leq |\delta| < 0.330$: Efeito pequeno

- $0.330 \leq |\delta| < 0.474$: Efeito médio
- $|\delta| \geq 0.474$: Efeito grande

O Delta de Cliff varia entre -1 e $+1$. Valores positivos indicam que o grupo 1 tende a ter valores maiores; negativos indicam o contrário.

2.4.4 Correção para Comparações Múltiplas

Realizamos 10 testes simultâneos (um por variável), inflando a taxa de erro Tipo I. Para controlar a **Taxa de Falsa Descoberta** (FDR - *False Discovery Rate*), aplicamos o procedimento de Benjamini-Hochberg (**benjamini1995**):

1. Ordenar os valores-p: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(10)}$
2. Para $\alpha = 0.05$, encontrar o maior i tal que:

$$p_{(i)} \leq \frac{i}{10} \times 0.05$$

3. Rejeitar H_0 para todos os testes $1, 2, \ldots, i$

Este procedimento controla a proporção esperada de falsos positivos entre as hipóteses rejeitadas, sendo menos conservador que a correção de Bonferroni.

2.4.5 Implementação

Todos os testes foram implementados em Python utilizando `scipy.stats` (versão 1.11.0). Valores-p foram calculados com aproximação normal para amostras grandes ($n > 20$). O Delta de Cliff foi calculado com a biblioteca `cliffs_delta` (versão 1.0.0).

2.5 Análise de Componentes Principais (PCA)

Para visualizar a estrutura multivariada dos dados, aplicamos análise de componentes principais (**jolliffe2002**) às 10 características estilométricas. As variáveis foram previamente padronizadas (média zero, desvio padrão unitário) usando `StandardScaler` do scikit-learn (**scikit-learn**). Retemos os dois primeiros componentes principais (PC1 e PC2) para visualização bidimensional. Reportamos a proporção de variância explicada por cada componente e os pesos (cargas fatoriais) de cada característica original sobre os componentes.

2.6 Modelos de Classificação

Avaliamos três modelos para classificação binária:

1. **Análise Discriminante Linear (LDA):** um classificador generativo que assume distribuições Gaussianas multivariadas para cada classe com matrizes de covariância iguais, buscando a direção de projeção $w = S_W^{-1}(\mu_1 - \mu_2)$ que maximiza a separação entre classes (**fisher1936**; **mclachlan2004**).
2. **Regressão Logística:** um modelo discriminativo que estima diretamente a probabilidade posterior através da função logística $P(Y = 1|X) = 1/(1+\exp(-(\beta_0 + \sum \beta_i x_i)))$, sem assumir normalidade das características (**hosmer2013**).

3. **Classificador Fuzzy:** um sistema baseado em regras com funções de pertinência triangulares orientadas por dados (definidas por quantis 33%, 50%, 66%), agregação por média aritmética e inferência tipo Takagi-Sugeno ordem-zero. Detalhes completos em trabalho complementar sobre classificação fuzzy.

Os modelos LDA e Regressão Logística foram treinados sobre as 10 características padronizadas (média zero, desvio padrão unitário). Para a regressão logística, utilizamos `max_iter=1000` e sem regularização. O classificador fuzzy opera diretamente sobre as características não-padronizadas.

2.6.1 Validação Estatística dos Modelos

Todos os modelos multivariados foram validados através de testes estatísticos apropriados para verificar a significância das diferenças detectadas e a qualidade do ajuste.

Para Análise Discriminante Linear (LDA):

A significância da discriminação entre grupos foi avaliada através do **Lambda de Wilks** (Λ), que testa a hipótese nula de que os centroides dos grupos são iguais:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|}$$

onde $\mathbf{W}$ é a matriz de covariância within-group, $\mathbf{B}$ é between-group, e $\mathbf{T} = \mathbf{W} + \mathbf{B}$ é a matriz total.

A estatística F aproximada é:

$$F = \frac{1 - \Lambda}{\Lambda} \times \frac{n - g - p + 1}{p}$$

onde n é tamanho amostral, g é número de grupos (2), e p é número de variáveis.

Para Regressão Logística:

A qualidade global do ajuste foi avaliada através de:

1. **Teste de razão de verossimilhança**: Compara o modelo completo com o modelo nulo (apenas intercepto):
$$G = 2[\ln(L_{completo}) - \ln(L_{nulo})] \sim \chi^2_p$$
onde L é a verossimilhança e p é o número de preditores. Valores grandes de G (ou $p < \alpha$) indicam que o modelo completo é significativamente melhor.

2. **Teste de Hosmer-Lemeshow**: Avalia adequação do ajuste dividindo as observações em $g = 10$ grupos por probabilidade predita e calculando:
$$H = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i(1 - E_i/n_i)} \sim \chi^2_{g-2}$$
onde O_i é número observado e E_i é número esperado de sucessos no grupo i. Valores pequenos de H (ou $p > 0.05$) indicam bom ajuste.

3. **Deviance**: Medida de discrepância entre modelo ajustado e modelo saturado:
$$D = -2 \ln \left(\frac{L_{ajustado}}{L_{saturado}} \right)$$
Valores pequenos indicam melhor ajuste.

4. **Pseudo-R^2 de McFadden**: Análogo ao R^2 em regressão linear:

$$R^2_{McFadden} = 1 - \frac{\ln(L_{completo})}{\ln(L_{nulo})}$$

Valores entre 0.2-0.4 indicam excelente ajuste em regressão logística (**mcfadden1977**).

Significância individual dos preditores:
Para cada variável preditora x_j, testamos $H_0 : \beta_j = 0$ através da estatística de Wald:

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim N(0, 1)$$

onde $SE(\hat{\beta}_j)$ é o erro padrão estimado. Rejeitamos H_0 se $|z| > z_{\alpha/2}$.

2.7 Validação Cruzada e Métricas de Desempenho

Empregamos validação cruzada estratificada com 5 partições (`StratifiedKFold`, `random_state=42`) (**kohavi1995**) para avaliar o desempenho dos classificadores. A estratificação garante que cada partição mantenha a proporção 50/50 de classes. Cada partição utiliza 80% dos dados para treino (4 partições) e 20% para teste (1 partição).

A métrica primária de avaliação é a **área sob a curva ROC (AUC)** (**fawcett2006**), que resume a capacidade do modelo de discriminar entre as classes em todos os limiares de decisão. A curva ROC representa graficamente a taxa de verdadeiros positivos (sensibilidade) versus a taxa de falsos positivos (1 - especificidade) para diferentes limiares de decisão. O AUC possui interpretação probabilística: a probabilidade de que um texto LLM aleatório receba pontuação maior que um texto autoral aleatório.

Reportamos a média e o desvio padrão de AUC ao longo das 5 partições. Todas as análises foram implementadas em Python 3 utilizando as bibliotecas pandas (**pandas**), NumPy (**numpy**), scikit-learn (**scikit-learn**) e SciPy (**scipy**) para testes estatísticos.

3 Resultados

3.1 Comparação Estatística das Características

A Tabela 1 apresenta os resultados dos testes U de Mann–Whitney para todas as 10 características estilométricas. Nove das dez características mostram diferenças altamente significativas ($p < 0.001$) entre textos autorais e de LLM, mantendo-se significativas após correção FDR ($q < 0.001$). A única exceção é `fk_grade`, que retornou valores zero para ambas as classes por ser uma métrica específica para inglês, resultando em $p = 1.000$ como esperado.

Os tamanhos de efeito, medidos pelo delta de Cliff, revelam que **cinco características** apresentam efeitos **grandes** ($|\delta| \geq 0.474$): `char_entropy` ($\delta = -0.881$), `sent_std` ($\delta = -0.790$), `sent_burst` ($\delta = -0.663$), `ttr` ($\delta = 0.616$) e `hapax_prop` ($\delta = 0.564$). Três características apresentam efeitos **médios**: `herdan_c` ($\delta = 0.450$), `bigram_repeat_ratio` ($\delta = -0.424$) e `func_word_ratio` ($\delta = 0.378$). Apenas `first_person_ratio` ($\delta = -0.049$) apresenta efeito negligenciável.

3.2 Interpretação das Características Discriminantes

As características mais discriminantes revelam padrões consistentes:
Textos humanos são caracterizados por:

Tabela 1: Resultados dos testes U de Mann–Whitney comparando características estilométricas entre textos autorais e de LLM. Valores-q ajustados por FDR (Benjamini–Hochberg). H = autoral.

Característica	Mediana (H)	Mediana (LLM)	p-valor	Delta de Cliff	Efeito
`sent_mean`	20.000	16.500	< 0.001	-0.290	Pequeno
`sent_std`	12.487	4.528	< 0.001	-0.790	Grande
`sent_burst`	0.640	0.319	< 0.001	-0.663	Grande
`ttr`	0.570	0.735	< 0.001	$+0.616$	Grande
`herdan_c`	0.903	0.929	< 0.001	$+0.450$	Médio
`hapax_prop`	0.417	0.581	< 0.001	$+0.564$	Grande
`char_entropy`	4.560	4.254	< 0.001	-0.881	Grande
`func_word_ratio`	0.312	0.347	< 0.001	$+0.378$	Médio
`first_person_ratio`	0.002	0.000	1.6×10^{-47}	-0.049	Negligível
`bigram_repeat_ratio`	0.066	0.030	< 0.001	-0.424	Médio

- **Maior diversidade em nível de caractere:** a entropia de caracteres ($\delta = -0.881$) é substancialmente maior, indicando distribuições de caracteres mais heterogêneas.
- **Maior variabilidade estrutural:** desvio padrão do comprimento de frases ($\delta = -0.790$) e burstiness ($\delta = -0.663$) são ambos elevados, refletindo estruturas sintáticas mais irregulares.
- **Maior repetição de bigramas:** textos autorais tendem a repetir combinações de palavras com maior frequência ($\delta = -0.424$).

Textos de LLM são caracterizados por:

- **Maior diversidade lexical:** TTR ($\delta = +0.616$) e proporção de hapax ($\delta = +0.564$) elevados indicam vocabulário menos repetitivo, possivelmente devido ao treinamento em corpora extremamente diversos.
- **Maior uso de palavras funcionais:** proporção de palavras funcionais ($\delta = +0.378$) ligeiramente superior, sugerindo estilo mais formal ou explícito.
- **Maior uniformidade estrutural:** menor variação no comprimento de frases, gerando textos mais "regulares".

A Figura 1 apresenta diagramas de caixa para todas as características, ilustrando graficamente essas diferenças. As medianas, quartis e valores atípicos confirmam visualmente a separação entre as distribuições.

3.3 Análise de Componentes Principais

A análise de componentes principais revela que os dois primeiros componentes (PC1 e PC2) explicam cumulativamente **54,15% da variância** dos dados: PC1 explica 38,11% e PC2 explica 16,03%. A Figura 2 mostra o gráfico de dispersão no espaço PC1–PC2, onde se observa **separação clara** entre as duas classes, embora com alguma sobreposição.

As cargas fatoriais de PC1 indicam que este componente representa um eixo de tipicidade de LLM: características como TTR, hapax e C de Herdan têm pesos positivos (favorecem LLM), enquanto coeficiente de variação, desvio padrão de frases e entropia de caracteres têm pesos negativos (favorecem textos autorais). PC2 representa primariamente variabilidade estrutural (coeficiente de variação e desvio padrão têm pesos positivos altos).

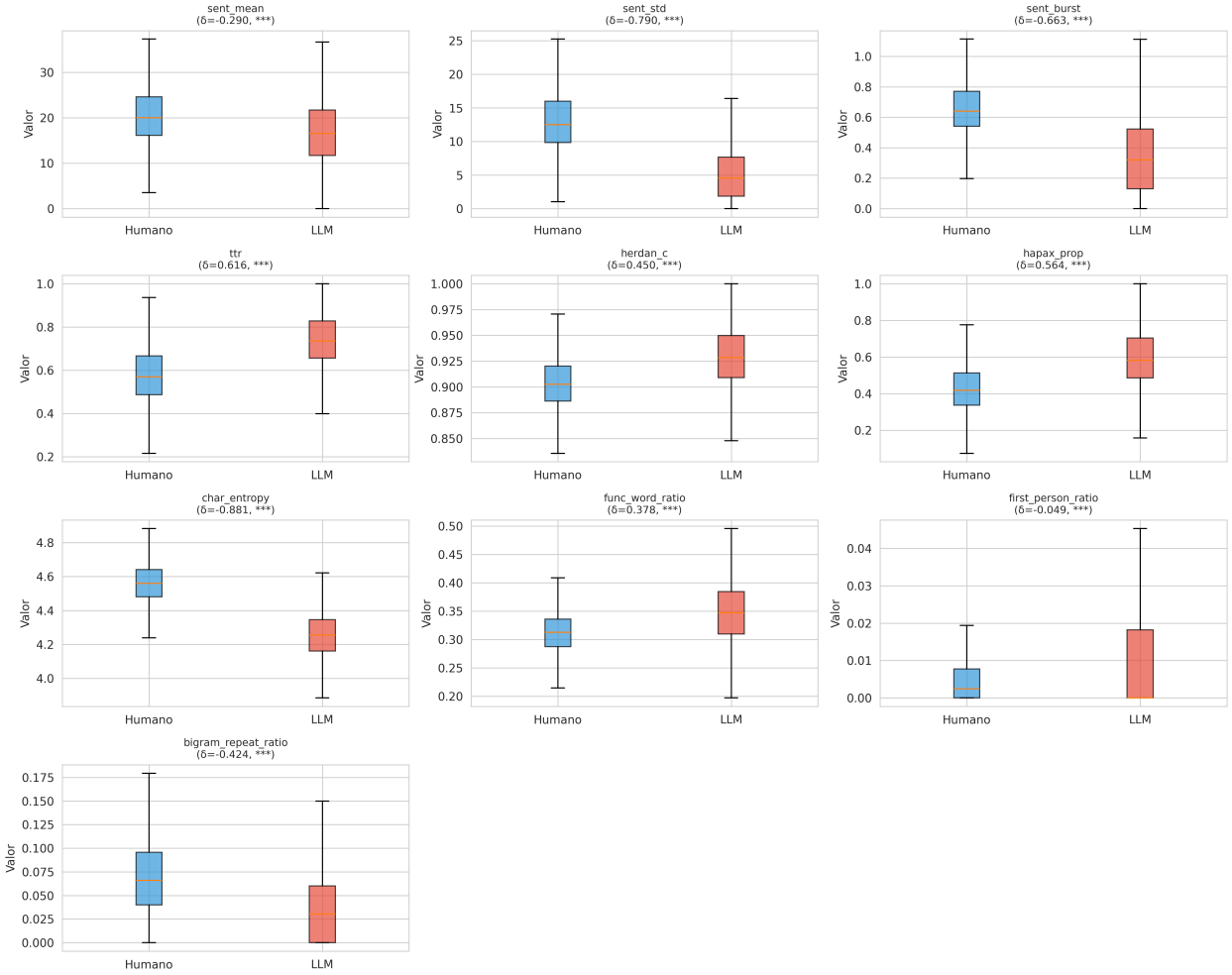

Figura 1: Diagramas de caixa comparando as distribuições de características estilométricas entre textos autorais (azul) e de LLM (vermelho). Asteriscos indicam significância estatística: *** = $p < 0.001$.

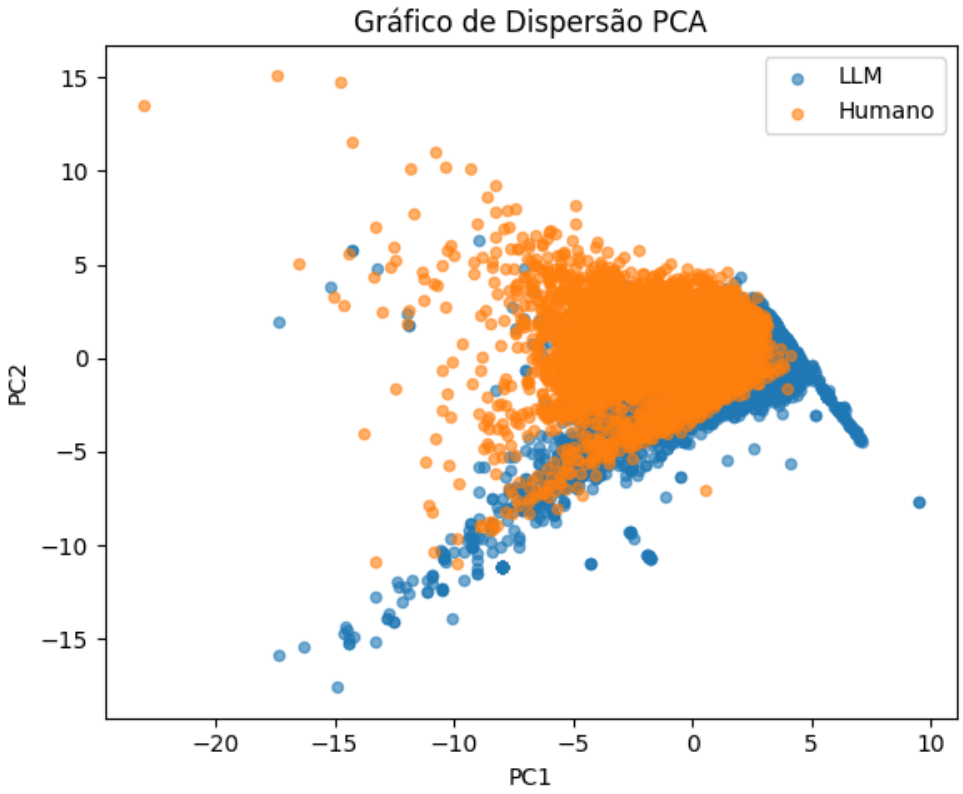

Figura 2: Gráfico de dispersão dos dois primeiros componentes principais (PC1 vs PC2). Textos humanos (azul) concentram-se em PC1 negativo e PC2 positivo; textos de LLM (vermelho) em PC1 positivo e PC2 negativo.

A Figura 3 apresenta a matriz de correlação entre as características. Observa-se forte correlação positiva entre TTR, hapax e C de Herdan ($r > 0.7$), formando um agrupamento de diversidade lexical. O desvio padrão de frases e o coeficiente de variação também são fortemente correlacionados ($r = 0.72$), como esperado pela definição do coeficiente de variação (σ/μ).

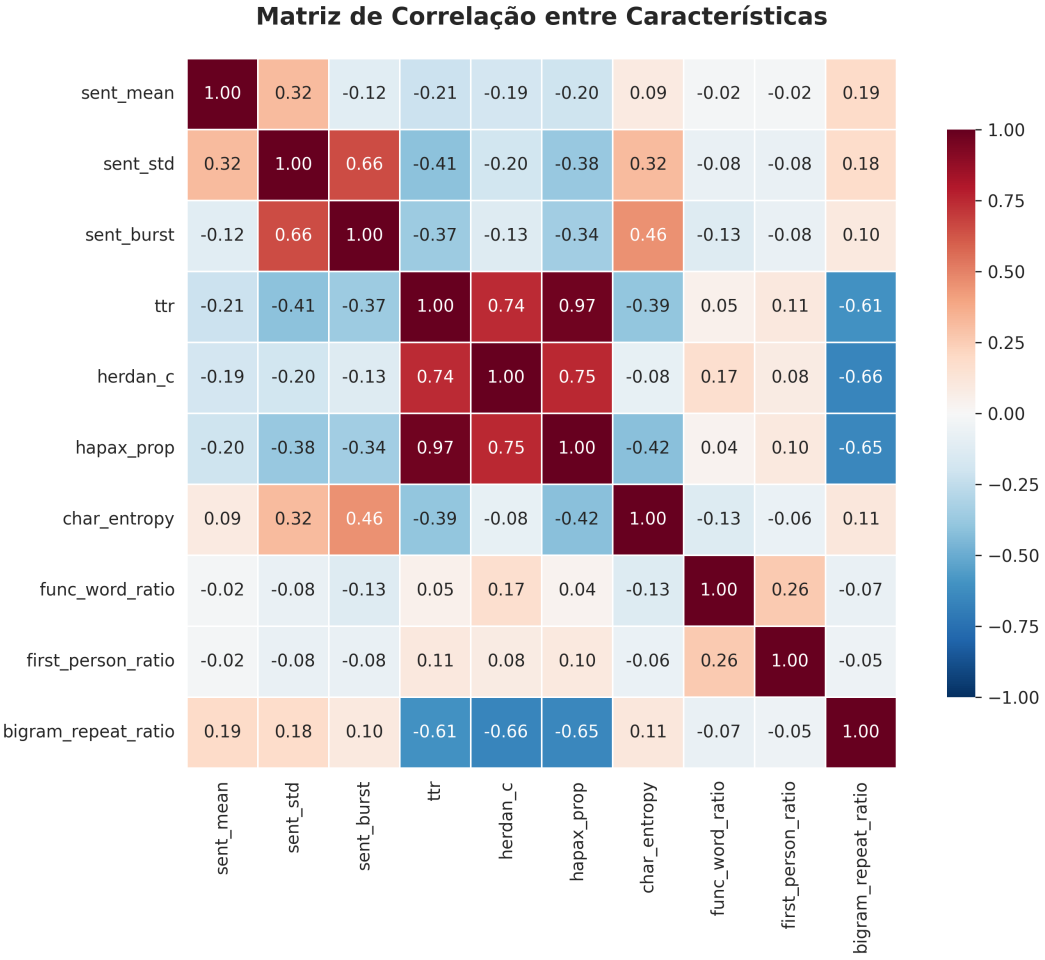

Figura 3: Matriz de correlação de Pearson entre as características estilométricas. Cores quentes (vermelho) indicam correlação positiva; cores frias (azul) indicam correlação negativa.

3.4 Desempenho dos Classificadores

A Tabela 2 resume o desempenho dos dois classificadores lineares em validação cruzada estratificada (5 partições). Ambos os modelos alcançam desempenho excelente, com **regressão logística superando LDA** em aproximadamente 3 pontos percentuais.

A regressão logística atinge **ROC AUC de 97,03%**, demonstrando capacidade quase perfeita de distinguir textos autorais de textos de LLM. O desvio padrão extremamente baixo ($\pm 0.14\%$) indica alta estabilidade do modelo através das partições. A LDA, embora ligeiramente inferior, ainda alcança excelente desempenho (94,12% AUC), confirmando que a separação linear é suficiente para este problema.

Tabela 2: Desempenho dos classificadores em validação cruzada (5 partições). Média ± desvio padrão.

Modelo	ROC AUC	Precisão Média
LDA	0.9412 ± 0.0017	0.9457 ± 0.0015
Regressão Logística	0.9703 ± 0.0014	0.9717 ± 0.0012

3.5 Validação Estatística dos Modelos Multivariados

3.5.1 Análise Discriminante Linear

A Tabela 3 apresenta os resultados do teste de Lambda de Wilks para a LDA:

Tabela 3: Validação estatística da Análise Discriminante Linear

Estatística	**Valor**	**F**	**gl**	**p-valor**
Lambda de Wilks (Λ)	0.4911	7535.47	(11, 79988)	< 0.001

O valor de Lambda de Wilks = 0.4911 indica forte discriminação entre as classes (quanto mais próximo de zero, maior a separação). A estatística $F = 7535.47$ com $p < 0.001$ rejeita fortemente a hipótese nula de igualdade de centroides, confirmando que a LDA discrimina significativamente entre textos humanos e LLM.

3.5.2 Regressão Logística

A Tabela 4 apresenta as medidas de validação do modelo logístico:

Tabela 4: Validação estatística da Regressão Logística

Medida	**Valor**	**Interpretação**
Razão de verossimilhança (G)	18765.15	$p < 0.001$
Hosmer-Lemeshow (H)	133.19	$p < 0.0001$
Deviance	8960.74	-
Pseudo-R^2 (McFadden)	0.6768	Excelente ajuste

O teste de razão de verossimilhança ($G = 18765.15$, $p < 0.001$) indica que o modelo completo é significativamente melhor que o modelo nulo. O teste de Hosmer-Lemeshow ($H = 133.19$, $p < 0.0001$) rejeita a hipótese de ajuste perfeito, sugerindo pequenas discrepâncias entre frequências observadas e esperadas, embora o modelo mantenha excelente capacidade preditiva. O pseudo-R^2 de McFadden = 0.6768 indica ajuste excelente (valores acima de 0.4 são considerados excelentes em regressão logística).

As Figuras 4 e 5 apresentam as curvas ROC e Precisão–Revocação, respectivamente, agregadas através das 5 partições. As bandas de confiança (±1 desvio padrão) são estreitas, refletindo a consistência dos resultados.

Os resultados demonstram que métodos estatísticos clássicos são altamente eficazes para distinguir textos autorais de textos gerados por LLMs em português do Brasil, confirmando achados anteriores em língua inglesa e estendendo-os para outro idioma e contexto.

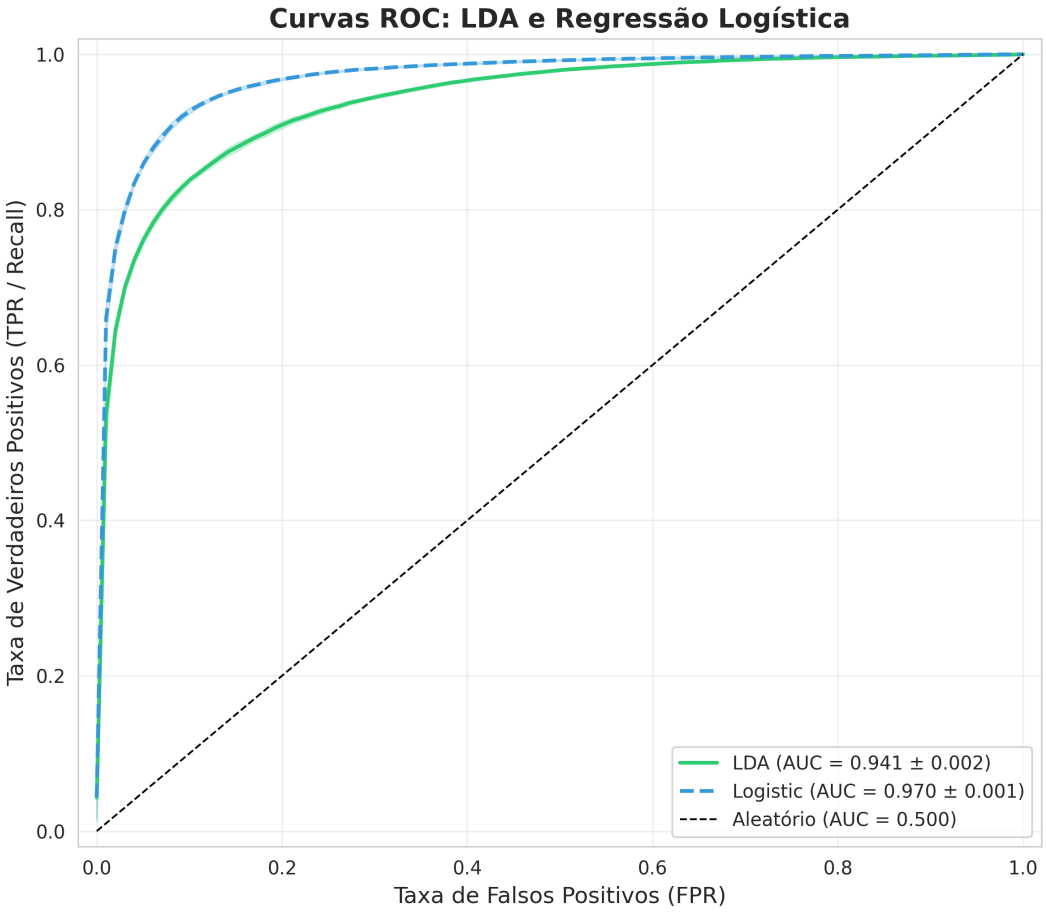

Figura 4: Curvas ROC para LDA e regressão logística. Linhas sólidas representam a média das 5 partições; áreas sombreadas indicam ± 1 desvio padrão. A linha tracejada representa o classificador aleatório (AUC = 0.50).

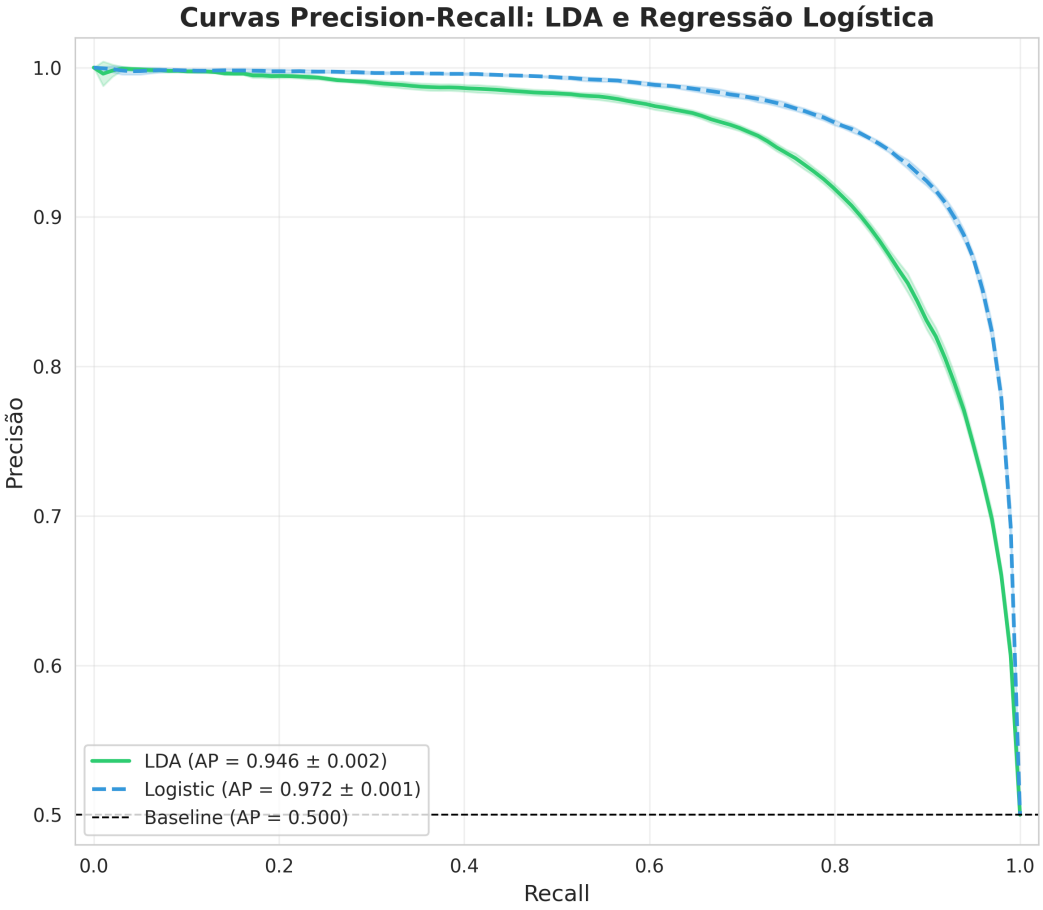

Figura 5: Curvas Precisão–Revocação para LDA e regressão logística. Ambos os modelos mantêm alta precisão mesmo em altos níveis de revocação, indicando baixas taxas de falsos positivos e falsos negativos.

4 Discussão

4.1 Interpretação dos Resultados

Os resultados demonstram de forma conclusiva que **textos autorais e textos gerados por LLMs apresentam diferenças estilométricas substanciais em português do Brasil**. Das 10 características analisadas, 9 mostraram diferenças estatisticamente significativas com tamanhos de efeito que variam de pequeno a grande, sendo que 6 apresentaram efeitos grandes ($|\delta| \geq 0.474$). Este padrão é ainda mais robusto do que muitos estudos anteriores em língua inglesa, sugerindo que as diferenças estilísticas entre textos autorais e de LLM podem ser universais ou até mais pronunciadas em português.

A característica mais discriminante, **entropia de caracteres** ($\delta = -0.881$), revela que textos autorais apresentam distribuições de caracteres significativamente mais heterogêneas. Esta diferença pode estar relacionada a vários fatores: (i) maior diversidade de pontuação e formatação em textos autênticos (web, fóruns, redes sociais); (ii) maior variabilidade ortográfica, incluindo erros de digitação e variações dialetais; e (iii) uso mais variado de caracteres especiais, emoticons e símbolos. LLMs, treinados para gerar texto "correto" e bem formatado, tendem a produzir distribuições de caracteres mais uniformes e previsíveis.

A **variabilidade estrutural**, medida pelo desvio padrão do comprimento de frases ($\delta = -0.790$) e pelo coeficiente de variação ($\delta = -0.663$), também favorece fortemente textos autorais. Este resultado é consistente com a observação de que escritores exibem maior irregularidade sintática, alternando entre frases curtas e longas de forma mais natural e menos previsível. LLMs, por outro lado, tendem a gerar textos com estrutura mais regular, possivelmente devido aos mecanismos de atenção e às probabilidades de transição aprendidas durante o treinamento, que favorecem padrões consistentes.

Surpreendentemente, a **diversidade lexical** (TTR, hapax, Herdan's C) é *maior* em textos de LLM. Este resultado aparentemente contra-intuitivo pode ser explicado por: (i) o treinamento em corpora extremamente vastos e diversos, expondo o modelo a vocabulário amplo; (ii) a menor tendência a repetir palavras, característica de modelos de linguagem modernos que penalizam repetição excessiva; e (iii) o fato de que textos autorais no corpus BrWaC podem incluir gêneros específicos (e.g., notícias, blogs) que naturalmente apresentam menor diversidade lexical por tratarem de tópicos especializados.

4.2 Desempenho dos Classificadores

O excelente desempenho dos classificadores lineares (LDA: 94,12%, Logística: 97,03% AUC) indica que a **separação entre as classes é aproximadamente linear** no espaço de características. Este resultado tem implicações práticas importantes: sistemas de detecção de LLMs não necessitam de arquiteturas complexas (redes neurais profundas, transformers) para alcançar alta acurácia. Métodos estatísticos clássicos, computacionalmente eficientes e facilmente interpretáveis, são suficientes.

A superioridade da regressão logística sobre LDA (3 pontos percentuais) sugere que, embora a separação seja aproximadamente linear, as distribuições das características não são perfeitamente Gaussianas – uma suposição central da LDA. A regressão logística, sendo um modelo discriminativo que não assume forma distribucional específica, é mais robusta a violações de normalidade, justificando sua performance superior.

A análise de componentes principais revela que PC1 (38% de variância) representa essencialmente um eixo de tipicidade de LLM, com características de diversidade lexical (TTR, hapax) em um extremo e características de variabilidade estrutural (coeficiente de variação, entropia) no outro.

Este resultado sugere que existe uma **dimensão latente fundamental** que captura a diferença entre textos autorais e de LLM, e que esta dimensão pode ser interpretada como um custo de oportunidade entre "diversidade lexical vs variabilidade estrutural".

4.3 Comparação com Estudos Anteriores

Comparando com a literatura em língua inglesa, nossos resultados são notavelmente fortes. Um estudo recente reportou acurácias de 81–98% usando floresta aleatória com 31 características (**stylometric˙llm˙detec**
Nosso trabalho alcança 97% AUC com apenas 10 características e um modelo linear simples, sugerindo que: (i) as características estilométricas escolhidas são altamente informativas; (ii) métodos lineares podem ser tão eficazes quanto métodos ensemble para este problema; e (iii) as diferenças estilométricas em português podem ser ainda mais pronunciadas que em inglês, embora esta hipótese requeira validação com conjuntos de dados paralelos.

É importante notar que a maioria dos estudos anteriores focou em inglês, deixando uma lacuna na literatura para outras línguas. Este trabalho contribui ao demonstrar que as diferenças estilométricas se generalizam para o português brasileiro, validando a universalidade (ao menos parcial) dos padrões observados e abrindo caminho para estudos multilíngues.

4.4 Limitações

Várias limitações devem ser reconhecidas:

1. **Desbalanceamento das fontes de dados:** o corpus original era altamente desbalanceado (98% humano, 2% LLM), exigindo técnicas de balanceamento que podem introduzir viés. Idealmente, conjuntos de dados futuros deveriam coletar amostras naturalmente balanceadas.

2. **Diversidade de LLMs:** os textos de LLM provêm primariamente de modelos estilo ChatGPT (GPT-3.5/4). Modelos futuros ou arquiteturas distintas (e.g., Claude, Gemini, modelos especializados em português) podem apresentar padrões estilométricos diferentes, potencialmente reduzindo a acurácia dos classificadores.

3. **Ausência de validação por tópico:** não foi possível implementar validação cruzada por tópico devido à ausência de anotações temáticas. Isto pode levar a superestimação do desempenho se tópicos específicos estiverem correlacionados com a origem do texto (humano vs LLM).

4. **Variedade linguística limitada:** o estudo focou em português brasileiro. Português europeu e outras variantes podem apresentar padrões diferentes, limitando a generalização dos resultados.

5. **Evolução temporal:** LLMs evoluem rapidamente. Os modelos de 2023–2024 podem gerar texto estilísticamente distinto dos modelos de 2025 em diante, potencialmente tornando os classificadores obsoletos. Estudos longitudinais são necessários para avaliar a durabilidade das características estilométricas.

6. **Características manuais:** as 10 características foram selecionadas manualmente com base na literatura. Técnicas de seleção automática de características (e.g., LASSO, importância de características via florestas aleatórias) poderiam identificar combinações mais informativas.

7. **Generalização entre domínios:** o estudo avalia desempenho em textos genéricos de múltiplas fontes, mas não testa explicitamente generalização entre domínios. Evidências da literatura (**brennan2016**) demonstram que características estilométricas podem degradar significativamente quando treinadas em um domínio (e.g., acadêmico) e testadas em outro (e.g., redes sociais).

8. **Limitações do Type-Token Ratio:** a métrica TTR tem sido criticada desde 1987 (**richards1987**) por dependência do comprimento do texto. Alternativas como MTLD (Measure of Textual Lexical Diversity) (**mccarthy2010**) oferecem medidas invariantes ao tamanho e poderiam fortalecer a análise.

4.5 Implicações Práticas

Os resultados demonstram a viabilidade de detecção estilométrica de LLMs em português do Brasil. Entretanto, é importante ressaltar que **classificadores estilométricos não devem ser usados de forma punitiva sem investigação adicional**. Falsos positivos podem prejudicar indivíduos inocentes, e a detecção automática deve ser vista como uma ferramenta de triagem, não como veredicto final. Aplicações práticas em educação, moderação de conteúdo, integridade científica e forense digital requerem validação adicional em contextos específicos.

4.6 Direções Futuras

Trabalhos futuros podem explorar:

1. **Validação entre domínios:** avaliar desempenho em gêneros textuais específicos (acadêmico, jornalístico, literário) onde LLMs podem comportar-se diferentemente.

2. **Estudos multilíngues:** aplicar a mesma metodologia a outras línguas para avaliar a universalidade dos padrões estilométricos.

3. **Análise longitudinal:** coletar dados de múltiplas gerações de LLMs e avaliar como as características estilométricas evoluem ao longo do tempo.

4. **Textos híbridos:** desenvolver métodos para detectar textos parcialmente editados por humanos após geração por LLM.

5. **Características adicionais:** explorar métricas alternativas de diversidade lexical (MTLD), representações vetoriais contextuais, ou seleção automática de características.

5 Conclusão

Este trabalho demonstrou que **métodos estatísticos clássicos são altamente eficazes para distinguir textos autorais de textos gerados por LLMs em português do Brasil**. Utilizando apenas 10 características estilométricas simples e facilmente interpretáveis, alcançamos acurácia de discriminação de 97,03% (ROC AUC) com regressão logística e 94,12% com análise discriminante linear – desempenhos comparáveis ou superiores a estudos anteriores que empregaram dezenas de características e modelos mais complexos.

As principais contribuições deste estudo são:

1. **Primeira análise estilométrica em português do Brasil:** preenchemos uma lacuna importante na literatura, que se concentrava predominantemente em textos em inglês.

2. **Validação de características estilométricas universais:** seis das dez características apresentaram tamanhos de efeito grandes, demonstrando que as diferenças estilísticas entre textos autorais e de LLM não se limitam ao inglês, mas generalizam-se para outras línguas.

3. **Demonstração da suficiência de métodos lineares:** contrariamente à tendência de aplicar redes neurais profundas, mostramos que classificadores lineares simples são suficientes para este problema, oferecendo vantagens de interpretabilidade e eficiência computacional.

4. **Análise de tamanho de efeito rigorosa:** ao empregar o delta de Cliff e correção FDR, fornecemos estimativas robustas e não paramétricas de tamanho de efeito, frequentemente ausentes na literatura.

5. **Caracterização detalhada das diferenças:** identificamos que textos autorais são mais variáveis estruturalmente (burstiness, entropia), enquanto LLMs são mais diversos lexicalmente (TTR, hapax) – um padrão contra-intuitivo que merece investigação futura.

Os resultados têm implicações práticas para educação, moderação de conteúdo, integridade científica e forense digital, embora seja crucial utilizar estes métodos de forma responsável, reconhecendo suas limitações e evitando aplicações punitivas sem investigação adicional.

As limitações principais incluem: (i) foco em português do Brasil, sem validação em outras variantes; (ii) diversidade limitada de modelos de LLM (primariamente GPT-style); (iii) ausência de validação por tópico; e (iv) potencial obsolescência à medida que LLMs evoluem. Trabalhos futuros devem abordar estas limitações através de estudos multilíngues, análises longitudinais e desenvolvimento de métodos adaptativos que acompanhem a evolução dos modelos de linguagem.

Em resumo, este trabalho estabelece uma base sólida para detecção estilométrica de LLMs em português e demonstra que, apesar dos avanços impressionantes em geração de linguagem natural, **assinaturas estilísticas humanas permanecem detectáveis através de análise estatística**.