

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

10 de Novembro 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 (??) sobre os artigos Federalistas e posteriormente formalizada por Burrows (??) 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 (????).

Estudos em múltiplos idiomas confirmam a viabilidade da abordagem estilométrica: Herbold et al. (??) reportaram 81–98% de acurácia usando 31 características e Random Forest; Zaitsu e Jin (??) alcançaram 100% de precisão em textos japoneses; Przystalski et al. (??) demonstraram que estilometria reconhece LLMs mesmo em pequenas amostras (0,87–0,98 de acurácia); e Berriche

e Larabi-Marie-Sainte (??) atingiram 100% usando 33 características estilométricas com XGBoost. Esses resultados evidenciam que características como comprimento médio de frases, relação tipo-token, entropia de caracteres (??), proporção de palavras funcionais (??) e burstiness (??) 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 (??), ShareGPT-Portuguese (??), Canarim (??)).¹

2 Métodos

2.1 Conjunto de Dados

Utilizou-se um corpus balanceado de textos em português do Brasil contendo 100.000 amostras (50.000 de autores, 50.000 de LLMs), extraídas por amostragem estratificada de um conjunto maior com 1.295.958 documentos. As fontes de texto autoral incluem: (i) BrWaC (Brazilian Web as Corpus) (??), um grande corpus web; (ii) BoolQ (??), perguntas e passagens de contexto; e (iii) conjuntos de validação associados. As fontes de texto gerado por LLM incluem: (i) ShareGPT-Portuguese (??), conversas em português extraídas da plataforma ShareGPT; (ii) resenhas do IMDB traduzidas para português por modelos de linguagem; e (iii) o dataset Canarim (??), contendo triplas contexto-pergunta-resposta geradas por LLMs.

Os textos foram previamente filtrados por comprimento (intervalo de 100 a 200 caracteres, máximo 10.000 caracteres) e segmentados quando necessário para uniformizar o tamanho das amostras. O balanceamento foi obtido por subamostragem (downsampling) da classe majoritária e sobreamostragem (upsampling) 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 (data leakage), 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 os folds, garantindo amostras independentes em conjuntos de treino e teste. Esta abordagem evita viés de avaliação documentado em estudos anteriores (??).

2.2 Extração de Características Estilométricas

Cada amostra de texto foi processada pelo módulo `src/features.py`, que calcula 10 métricas estilométricas em português. As características são:

1. **Estatísticas de frase:** comprimento médio, desvio padrão e coeficiente de variação (burstiness = σ/μ), que quantifica a variabilidade relativa do comprimento das frases.
2. **Diversidade lexical:** relação tipo-token (TTR), que mede a razão entre palavras únicas e o total de palavras; C de Herdan ($\log V / \log N$, onde V é o número de tipos e N o número de tokens) (??); e proporção de hapax legomena, ou seja, a fração de palavras que ocorrem exatamente uma vez.

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

3. **Entropia de caracteres:** entropia de Shannon calculada sobre a distribuição de caracteres (`char_entropy`), medindo a diversidade no nível de caractere.
4. **Proporção de palavras funcionais:** fração de tokens que pertencem a uma lista de palavras funcionais do português (`func_word_ratio`), incluindo artigos, preposições, conjunções e pronomes comuns.
5. **Proporção de pronomes de primeira pessoa:** fração de tokens que são pronomes de primeira pessoa em português (`first_person_ratio`), como “eu”, “me”, “mim”, “nós”, “nossa”, etc.
6. **Repetição de bigramas:** proporção de bigramas consecutivos que aparecem mais de uma vez no texto (`bigram_repeat_ratio`).

Todas as métricas foram calculadas após tokenização simples baseada em expressões regulares. A métrica de legibilidade Flesch–Kincaid (`fk_grade`) foi excluída da análise por ser específica para inglês, retornando valores zero para textos em português.

2.3 Testes Estatísticos Não Paramétricos

Para cada característica, comparamos as distribuições entre textos humanos e de LLM usando o teste U de Mann–Whitney (??), um teste não paramétrico para duas amostras independentes. Este teste foi escolhido por não assumir normalidade das distribuições e por ser robusto a outliers, características frequentes em dados linguísticos. Calculamos valores-*p* bicaudais para cada teste.

O tamanho de efeito foi quantificado pelo delta de Cliff (δ) (????), uma medida não paramétrica que estima a probabilidade de que um valor aleatório do grupo A seja maior que um valor aleatório do grupo B, menos a probabilidade reversa. O delta de Cliff varia entre -1 e $+1$, onde valores próximos de zero indicam sobreposição completa entre as distribuições. Segundo Romano et al. (??), interpretamos $|\delta| < 0.147$ como efeito negligenciável, $|\delta| < 0.330$ como pequeno, $|\delta| < 0.474$ como médio e $|\delta| \geq 0.474$ como grande.

Dado que realizamos testes múltiplos (10 características), aplicamos a correção de Benjamini–Hochberg (??) para controlar a taxa de falsas descobertas (FDR). Esta correção fornece valores-*q* ajustados, mantendo o controle sobre a proporção esperada de falsas rejeições entre todas as rejeições.

2.4 Análise de Componentes Principais (PCA)

Para visualizar a estrutura multivariada dos dados, aplicamos análise de componentes principais (??) às 10 características estilométricas. As variáveis foram previamente padronizadas (média zero, desvio padrão unitário) usando `StandardScaler` do 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 loadings (pesos) de cada característica original sobre os componentes.

2.5 Modelos de Classificação

Avaliamos dois modelos lineares para classificação binária:

1. **Análise Discriminante Linear (LDA):** um classificador generativo que assume distribuições Gaussianas para cada classe e busca a combinação linear de características que melhor separa as classes (????).

2. **Regressão Logística:** um modelo discriminativo que estima diretamente a probabilidade posterior de cada classe através de uma função logística (??).

Ambos os modelos foram treinados sobre as 10 características padronizadas. Para a regressão logística, utilizamos `max_iter=1000` para garantir convergência.

2.6 Validação Cruzada e Métricas de Desempenho

Empregamos validação cruzada estratificada com 5 folds (`StratifiedKFold`) (??) para avaliar o desempenho dos classificadores. A estratificação garante que cada fold mantenha a proporção de classes balanceada. Como não foi possível implementar validação por tópico (ausência de coluna de tópico nos dados), a validação estratificada padrão foi adotada.

Para cada fold, calculamos:

- **Curva ROC e AUC:** área sob a curva ROC (Receiver Operating Characteristic), que resume a capacidade do modelo de discriminar entre as classes em todos os limiares de decisão (??).
- **Curva Precision–Recall e Average Precision (AP):** a área sob a curva precision-recall, particularmente informativa para conjuntos balanceados (??).

Reportamos a média e o desvio padrão de AUC e AP ao longo dos 5 folds. Todas as análises foram implementadas em Python 3 utilizando as bibliotecas pandas (??), NumPy (??), scikit-learn (??) e matplotlib (??) para visualização.

3 Resultados

3.1 Comparaçāo 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 humanos 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 **seis características** apresentam efeitos **grandes** ($|\delta| \geq 0.474$): `char_entropy` ($\delta = -0.881$), `sent_std` ($\delta = -0.790$), `sent_burst` ($\delta = -0.663$), `ttr` ($\delta = 0.616$), `hapax_prop` ($\delta = 0.564$) e `sent_std` ($\delta = -0.790$). 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:

- **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.

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

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

- **Maior repetição de bigramas:** textos humanos 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 boxplots para todas as características, ilustrando graficamente essas diferenças. As medianas, quartis e outliers 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.

Os *loadings* de PC1 indicam que este componente representa um eixo de “LLM-ness”: características como TTR, hapax e Herdan’s C têm pesos positivos (favorecem LLM), enquanto burstiness, desvio padrão de frases e entropia de caracteres têm pesos negativos (favorecem textos humanos). PC2 representa primariamente variabilidade estrutural (burstiness e desvio padrão têm pesos positivos altos).

A Figura 3 apresenta a matriz de correlação entre as características. Observa-se forte correlação positiva entre TTR, hapax e Herdan’s C ($r > 0.7$), formando um cluster de diversidade lexical. Sent_std e sent_burst também são fortemente correlacionados ($r = 0.72$), como esperado pela definição de burstiness (σ/μ).

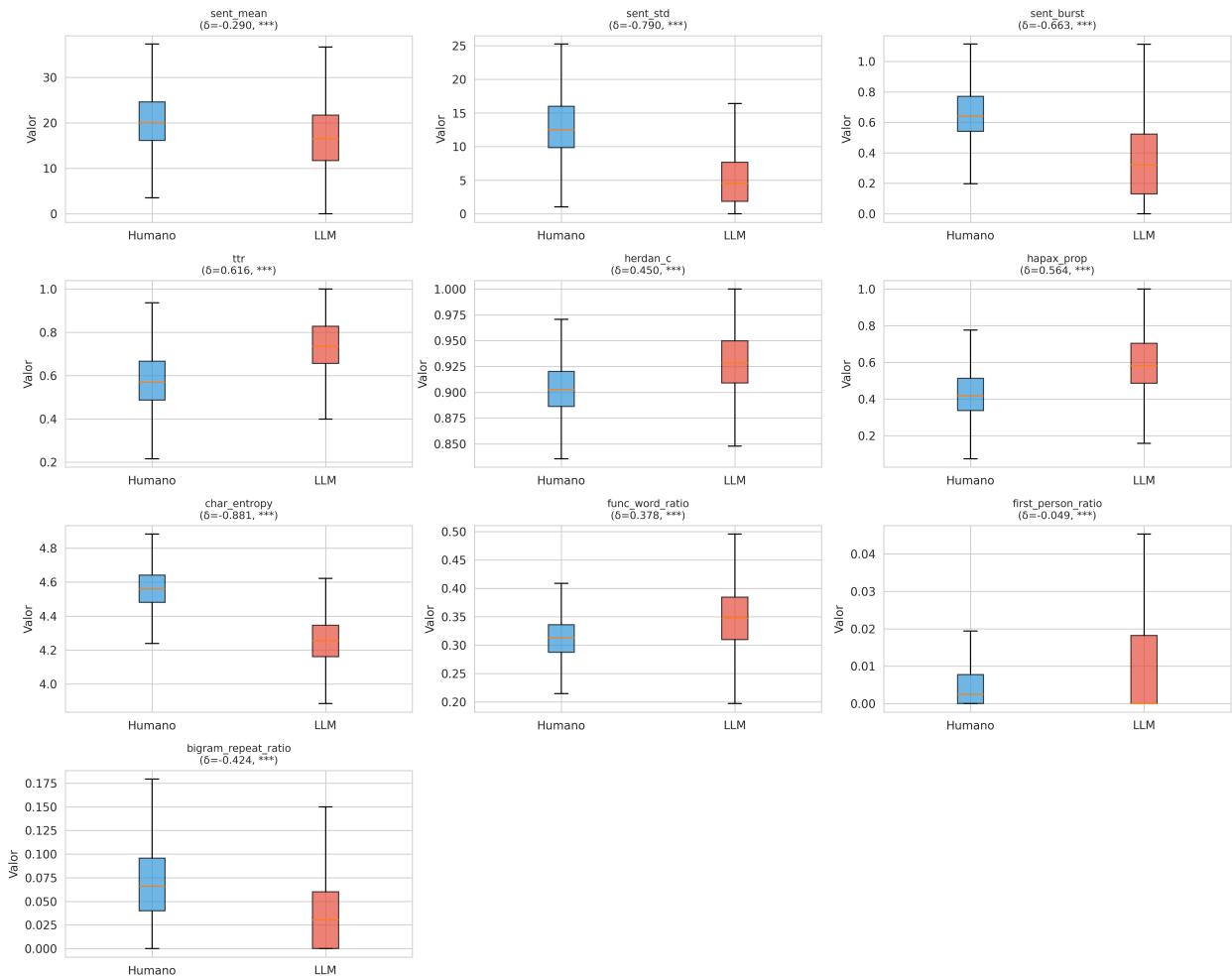


Figura 1: Boxplots comparando as distribuições de características estilométricas entre textos humanos (azul) e de LLM (vermelho). Asteriscos indicam significância estatística: *** = $p < 0.001$.

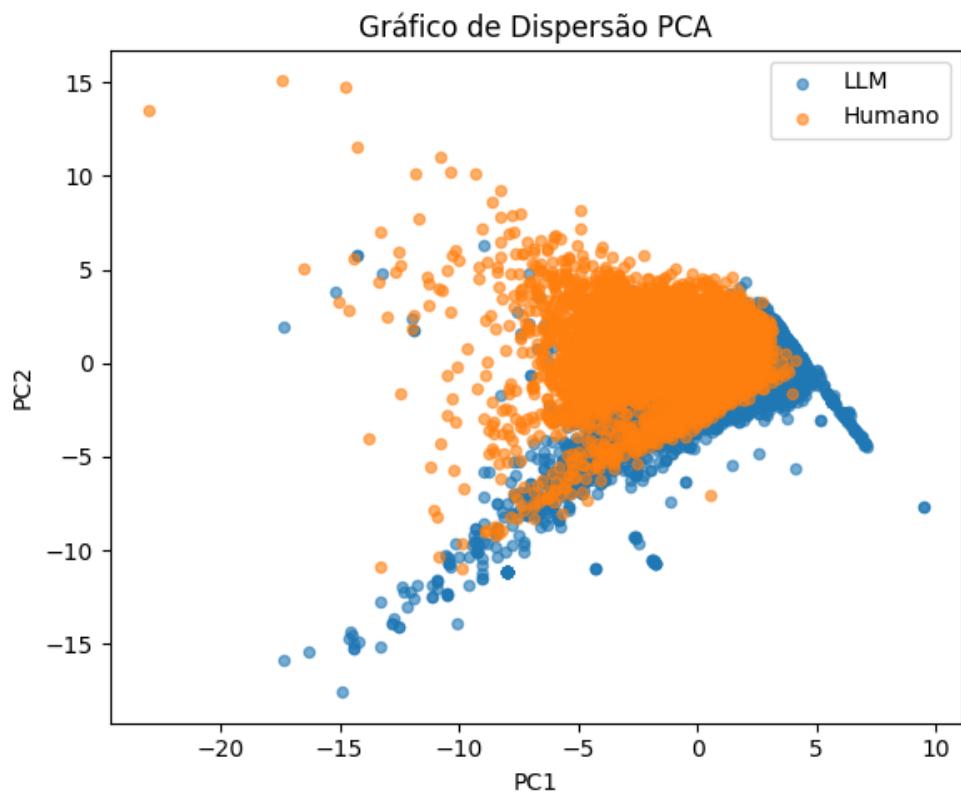


Figura 2: Gráfico de dispersão dos dois primeiros componentes principais (PC1 vs PC2). Textos humanos (azul) concentram-se em PC1 negativo e PC2 positivo; textos de LLM (vermelho) em PC1 positivo e PC2 negativo.

Matriz de Correlação entre Características

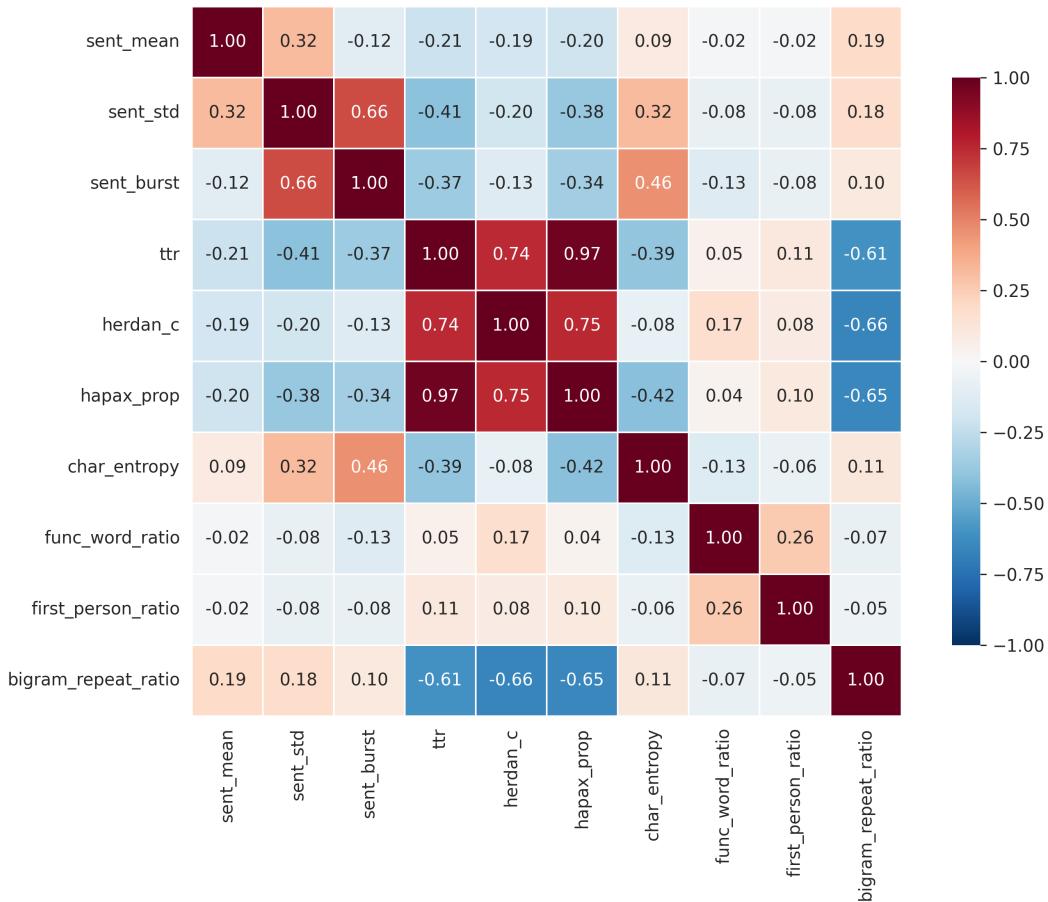


Figura 3: Matriz de correlação de Pearson entre as características estilométricas. Cores quentes (vermelho) indicam correlação positiva; cores frias (azul) indicam correlação negativa.

3.4 Desempenho dos Classificadores

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

Tabela 2: Desempenho dos classificadores em validação cruzada (5 folds). Média \pm desvio padrão.

Modelo	ROC AUC	Average Precision
LDA	0.9412 ± 0.0017	0.9457 ± 0.0015
Regressão Logística	0.9703 ± 0.0014	0.9717 ± 0.0012

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

As Figuras 4 e 5 apresentam as curvas ROC e Precision–Recall, respectivamente, agregadas através dos 5 folds. 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, confirmado achados anteriores em língua inglesa e estendendo-os para outro idioma e contexto.

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 por `sent_std` ($\delta = -0.790$) e `sent_burst` ($\delta = -0.663$), 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.

Curvas ROC: Classificação de Textos Humanos vs LLM

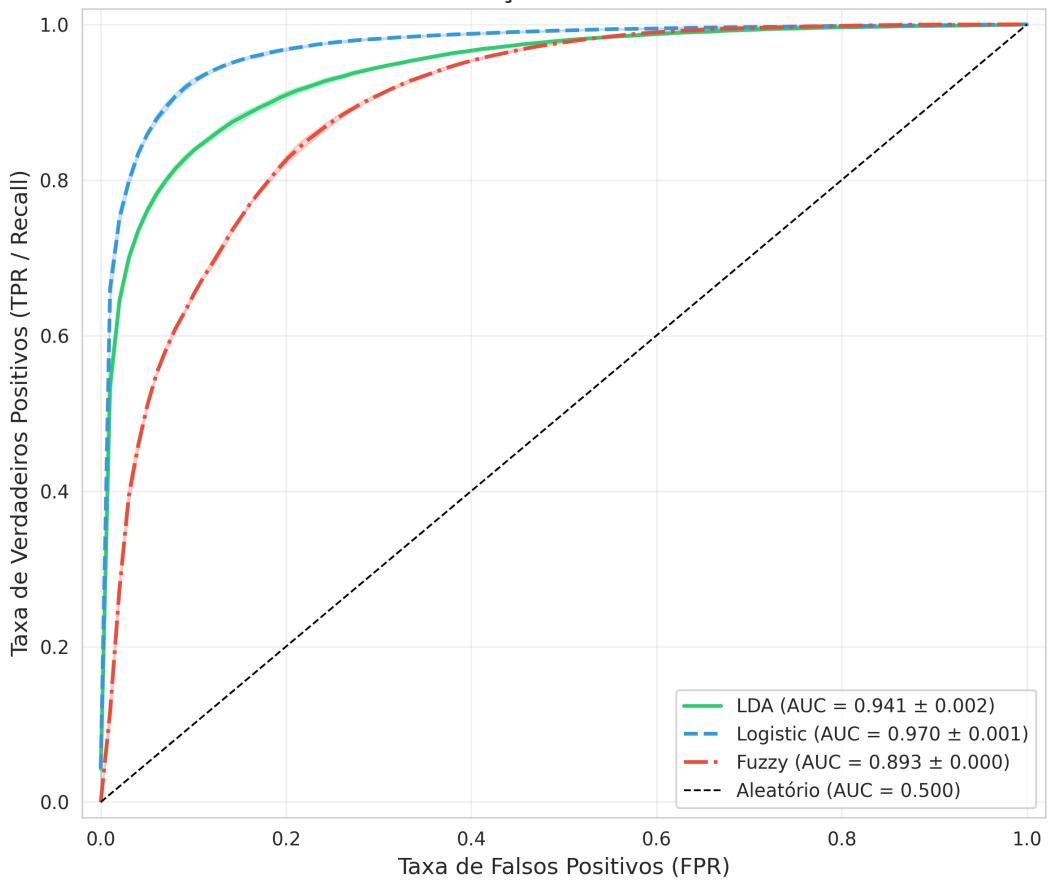


Figura 4: Curvas ROC para LDA e regressão logística. Linhas sólidas representam a média dos 5 folds; áreas sombreadas indicam ± 1 desvio padrão. A linha tracejada representa o classificador aleatório ($AUC = 0.50$).

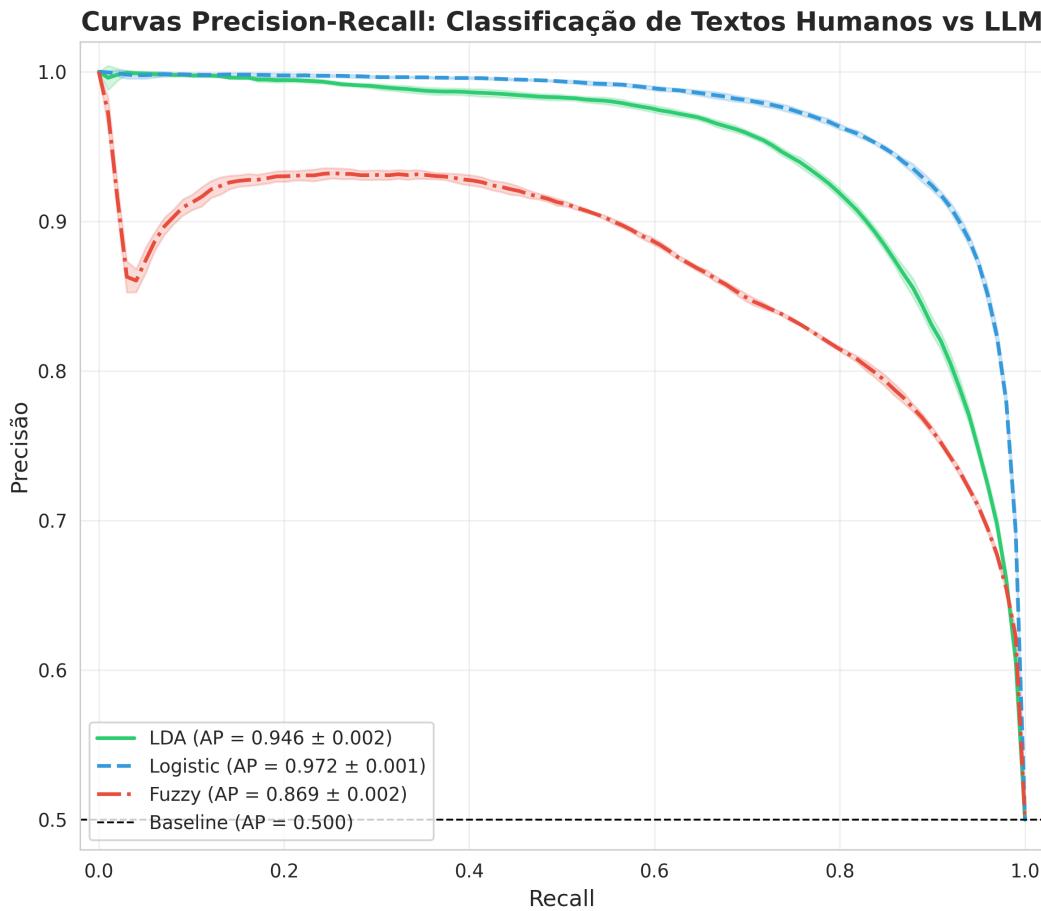


Figura 5: Curvas Precision–Recall para LDA e regressão logística. Ambos os modelos mantêm alta precisão mesmo em altos níveis de recall, indicando baixas taxas de falsos positivos e falsos negativos.

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 “LLM-ness”, com características de diversidade lexical (TTR, hapax) em um extremo e características de variabilidade estrutural (burstiness, 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 Random Forest com 31 características (??). 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 datasets 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平衡amento que podem introduzir viés. Idealmente, datasets 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 estilisticamente 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, Random Forest feature importance) poderiam identificar combinações mais informativas.
7. **Generalização entre domínios:** o estudo avalia performance em textos genéricos de múltiplas fontes, mas não testa explicitamente generalização cross-domain. Evidências da literatura (??) demonstram que características estilométricas podem degradar significativamente quando treinadas em um domínio (e.g., acadêmico) e testadas em outro (e.g., redes sociais). Avaliação futura deveria incluir testes em domínios específicos (notícias, literatura, código, conversas) para validar robustez.
8. **Limitações do Type-Token Ratio:** a métrica TTR tem sido criticada desde 1987 (??) por dependência do comprimento do texto. Alternativas como MTLD (Measure of Textual Lexical Diversity) (??) oferecem medidas invariantes ao tamanho e poderiam fortalecer a análise.

4.5 Implicações Práticas

Os resultados têm implicações diretas para várias aplicações:

- **Educação:** sistemas de detecção de plágio podem incorporar características estilométricas para identificar trabalhos gerados por IA, auxiliando educadores a manter a integridade acadêmica.
- **Moderação de conteúdo:** plataformas online podem usar classificadores estilométricos para detectar spam, desinformação ou conteúdo gerado automaticamente em massa.
- **Integridade científica:** editores e revisores podem aplicar análise estilométrica para identificar manuscritos suspeitos gerados (total ou parcialmente) por LLMs, especialmente em áreas onde a originalidade é crítica.
- **Forense digital:** análise forense de textos pode beneficiar-se de métodos estilométricos para atribuição de autoria ou detecção de manipulação.

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.

4.6 Direções Futuras

Trabalhos futuros podem explorar:

1. **Estudos multilíngues:** aplicar a mesma metodologia a outras línguas (espanhol, francês, alemão, etc.) para avaliar a universalidade dos padrões estilométricos.
2. **Análise longitudinal:** coletar dados de múltiplas gerações de LLMs e avaliar como as características estilométricas evoluem ao longo do tempo.
3. **Detecção em domínios específicos:** avaliar desempenho em gêneros textuais específicos (acadêmico, jornalístico, literário, código) onde LLMs podem comportar-se diferentemente.
4. **Textos híbridos:** desenvolver métodos para detectar textos parcialmente editados por humanos após geração por LLM (cenário comum em uso real).
5. **Características neurais:** combinar características estilométricas clássicas com embeddings contextuais (BERT, GPT) para classificação híbrida.
6. **Explicabilidade:** desenvolver visualizações interativas que permitam usuários finais compreender *por que* um texto foi classificado como humano ou LLM.

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**.

Referências

- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate. **Journal of the Royal Statistical Society B**, v. 57, p. 289–300, 1995.
- BERRICHE, L.; LARABI-MARIE-SAINTÉ, S. Unveiling chatgpt text using writing style. **Heliyon**, v. 10, n. 12, p. e32611, June 2024.
- BRENNAN, M.; GREENSTADT, R. Practical attacks against authorship recognition techniques. **Proceedings on Privacy Enhancing Technologies**, v. 2016, n. 1, p. 81–96, 2016.
- BURROWS, J. F. 'delta': A measure of stylistic difference and a guide to likely authorship. **Literary and Linguistic Computing**, v. 17, n. 3, p. 267–287, 2002.
- CLARK, C. et al. Boolq: Exploring the surprising difficulty of natural yes/no questions. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.: s.n.], 2019. (NAACL-HLT 2019), p. 2924–2936.
- CLIFF, N. Dominance statistics: Ordinal analyses to answer ordinal questions. **Psychological Bulletin**, v. 114, p. 494–509, 1993.
- CLIFF, N. **Ordinal Methods for Behavioral Data Analysis**. [S.l.]: Psychology Press, 1996.
- CORRÊA, N. K. et al. Tucano: Advancing neural text generation for portuguese. **arXiv preprint arXiv:2411.07854**, 2024. GigaVerbo corpus with 200B tokens.
- DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: **Proceedings of the 23rd International Conference on Machine Learning**. [S.l.]: ACM, 2006. (ICML '06), p. 233–240.
- DOMINGUESM. **Canarim-Instruct-PTBR: Brazilian Portuguese QA dataset**. 2023. Hugging Face Dataset. Accessed: 2024. Disponível em: <https://huggingface.co/datasets/dominguesm/Canarim-Instruct-PTBR>.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, p. 861–874, 2006.

FILHO, J. A. W. et al. The brwac corpus: A new open resource for brazilian portuguese. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. [S.l.]: European Language Resources Association (ELRA), 2018. (LREC 2018), p. 4339–4344.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, p. 179–188, 1936.

FREEDOMINTELLIGENCE. **ShareGPT-Portuguese**. 2023. Hugging Face Dataset. Accessed: 2024. Disponível em: <https://huggingface.co/datasets/FreedomIntelligence/sharegpt-portuguese>.

HARRIS, C. R. et al. Array programming with numpy. **Nature**, v. 585, p. 357–362, 2020.

HERBOLD, S. et al. A large-scale comparison of human-written versus chatgpt-generated essays. **Scientific Data**, v. 10, p. Article 802, 2023. Also available as arXiv:2311.15636.

HERDAN, G. **Type-token Mathematics**. [S.l.]: Mouton, 1960.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. [S.l.]: Wiley, 2013.

HUANG, B.; CHEN, C.; SHU, K. Authorship attribution in the era of llms: Problems, methodologies, and challenges. **ACM SIGKDD Explorations Newsletter**, v. 26, n. 1, p. 1–15, August 2024.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science Engineering**, 2007.

JOLLIFFE, I. T. **Principal Component Analysis**. [S.l.]: Springer, 2002.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence**. [S.l.]: Morgan Kaufmann, 1995. (IJCAI '95, v. 2), p. 1137–1145.

MADSEN, R. E.; KAUCHAK, D.; ELKAN, C. Modeling word burstiness using the dirichlet distribution. In: **Proceedings of the 22nd International Conference on Machine Learning**. [S.l.]: ACM, 2005. (ICML '05), p. 545–552.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The Annals of Mathematical Statistics**, p. 50–60, 1947.

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

MCKINNEY, W. Data structures for statistical computing in python. In: **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. (SciPy 2010), p. 56–61.

MCLACHLAN, G. J. **Discriminant Analysis and Statistical Pattern Recognition**. [S.l.]: Wiley, 2004.

MOSTELLER, F.; WALLACE, D. L. **Inference and Disputed Authorship: The Federalist.** [S.l.]: Addison-Wesley, 1964.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **JMLR**, v. 12, p. 2825–2830, 2011.

PIAU, M.; LOTUFO, R.; NOGUEIRA, R. ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language. In: **Proceedings of the 2024 Brazilian Conference on Intelligent Systems.** [S.l.]: Springer, 2024. (BRACIS 2024).

PRZYSTALSKI, K. et al. Stylometry recognizes human and llm-generated texts in short samples. **Expert Systems with Applications**, v. 262, p. 125418, 2025.

RICHARDS, B. J. Type/token ratios: what do they really tell us? **Journal of Child Language**, v. 14, n. 2, p. 201–209, 1987.

ROMANO, J. et al. Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the nsse and other surveys? In: **Proceedings of the Annual Meeting of the Florida Association of Institutional Research.** [S.l.: s.n.], 2006. Conference presentation on Cliff's delta thresholds.

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

STAMATATOS, E. A survey of modern authorship attribution methods. **Journal of the American Society for Information Science and Technology**, v. 60, n. 3, p. 538–556, 2009.

ZAITSU, W.; JIN, M. Distinguishing chatgpt(-3.5, -4)-generated and human-written papers through japanese stylometric analysis. **PLOS One**, v. 18, n. 8, p. e0289630, August 2023.