

# Classificação Estilométrica com Teoria de Conjuntos Fuzzy e Raciocínio Aproximado

Victor Löfgren Sattamini

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

10 de Novembro de 2025

## Resumo

Este trabalho apresenta um classificador baseado em lógica fuzzy para detecção de textos gerados por LLM - Large Language Models em português do Brasil. A abordagem utiliza propriedades quantitativas da escrita associadas a funções de pertinência triangulares que expressam o grau de pertencimento de um texto a categorias linguísticas interpretáveis. Os parâmetros das funções são determinados de forma orientada a dados, usando quantis (33%, 50%, 66%) das distribuições observadas no conjunto de treino, eliminando a necessidade de conhecimento especialista. O sistema de inferência fuzzy estima os graus de pertinência através de média aritmética para estimar a probabilidade de um texto ser autoral ou gerado por LLM. Avaliamos o classificador em um corpus balanceado de 100.000 amostras usando validação cruzada estratificada de 5 folds. O classificador fuzzy alcançou ROC AUC de 89,34% ( $\pm 0,04\%$ ), demonstrando desempenho competitivo comparado a métodos estatísticos (regressão logística: 97,03%, LDA: 94,12%) e neurais mais complexos. Além disso, o classificador apresentou variância  $3\text{--}4\times$  menor que métodos comparativos, indicando maior robustez. O custo de oportunidade entre interpretabilidade e desempenho é modesto (cerca de 8% de perda em AUC), tornando o modelo adequado para cenários onde transparência e auditabilidade são prioritárias, como educação, moderação de conteúdo e integridade científica. Este trabalho mostra que sistemas fuzzy podem competir com abordagens mais complexas, preservando vantagens cruciais de explicabilidade.

## 1 Introdução

Neste trabalho, exploramos o uso de lógica fuzzy como método de detecção de textos gerados por modelos de linguagem de grande porte (LLMs). Para isso, construímos um classificador fuzzy baseado em métricas estilométricas - propriedades da escrita que capturam padrões linguísticos, sintáticos e semânticos. Cada métrica é associada a uma função de pertinência que expressa o grau de pertencimento de um texto a variáveis linguísticas interpretáveis, como "alta fluência" ou "baixa variação lexical".

As funções de pertinência adotadas são triangulares, determinadas por três parâmetros  $(a, b, c)$ , amplamente utilizadas em sistemas fuzzy por sua simplicidade algorítmica e eficiência computacional (??).

O interesse em utilizar lógica fuzzy na estilometria decorre da natureza intrinsecamente gradual da linguagem. Categorias como "texto bem estruturado" ou "escrita natural" dependem de critérios de pertinência. A lógica fuzzy ocupa um espaço entre empirismo e formalidade, aproximando-se

da forma como utilizamos a linguagem natural para expressar incerteza e imprecisão (??). Essa característica a torna adequada para modelar a "gradualidade" no pertencimento de um texto a uma classe (autoral ou LLM).

Ao fuzificar métricas estilométricas e combiná-las no sistema de inferência fuzzy de regras "Se ... então", é possível estimar o grau de pertencimento de um texto a cada classe.

A principal vantagem da abordagem fuzzy é a **interpretabilidade**: ao contrário de modelos de caixa-preta, os graus de pertinência podem ser inspecionados e compreendidos por humanos, revelando em que medida cada dimensão estilométrica contribui para a decisão. Além disso, o sistema fuzzy permite incorporar conhecimento linguístico especializado na definição das funções de pertinência, embora aqui seja adotada uma abordagem orientada a dados (*data-driven*), determinando os parâmetros a partir de quantis das distribuições observadas.

A lógica fuzzy tem sido amplamente aplicada em processamento de linguagem natural, especialmente em análise de sentimentos (??) e classificação de texto (??). Trabalhos recentes também exploram sistemas fuzzy interpretativos baseados em fundamentos axiomáticos (??), demonstrando a viabilidade de sistemas transparentes e auditáveis. Contudo, até onde sabemos, nenhum estudo anterior aplicou lógica fuzzy especificamente à detecção de textos gerados por inteligência artificial. Enquanto LLMs têm sido analisados predominantemente por métodos estatísticos ou de aprendizado profundo, este trabalho propõe a utilização de sistemas de inferência fuzzy como alternativa explicável, eficiente e de fácil interpretação.

Os resultados apresentados demonstram que classificadores fuzzy simples podem alcançar desempenho competitivo (AUC de 89%) em comparação com abordagens estatísticas e neurais mais complexas, preservando ao mesmo tempo transparência e interpretabilidade: características essenciais para aplicações em educação, moderação de conteúdo e integridade científica.

## 2 Fundamentos de Teoria de Conjuntos Fuzzy

### 2.1 Fundamentos de Conjuntos Fuzzy

Um conjunto fuzzy  $A$  em um universo de discurso  $U$  caracteriza-se por uma função de pertinência  $\mu_A : U \rightarrow [0, 1]$  que atribui a cada elemento  $x \in U$  um grau de pertinência no intervalo  $[0, 1]$  (??). Diferentemente dos conjuntos clássicos, onde  $\mu_A(x) \in \{0, 1\}$ , os conjuntos fuzzy permitem transições graduais quando a inclusão total ( $\mu_A(x) = 1$ ) e exclusão total ( $\mu_A(x) = 0$ ). Os conceitos a seguir são importantes:

- **Núcleo (core)**:  $\{x \in U : \mu_A(x) = 1\}$ , região com pertinência total
- **Suporte (support)**:  $\{x \in U : \mu_A(x) > 0\}$ , região com pertinência não-nula
- **Fronteira (boundary)**: região onde  $0 < \mu_A(x) < 1$ , pertinência parcial

A lógica fuzzy, proposta por Zadeh (??), estende a lógica Booleana clássica para manipular valores de verdade parciais e suportar raciocínio aproximado. Operações fundamentais incluem união (máximo), interseção (mínimo) e complemento ( $1 - \mu_A(x)$ ).

### 2.2 Conjunto de Dados e Características

Utilizamos as mesmas 10 características estilométricas: `sent_mean`, `sent_std`, `sent_burst`, `ttr`, `herdan_c`, `hapax_prop`, `char_entropy`, `func_word_ratio`, `first_person_ratio` e `bigram_repeat_ratio`.

## 2.3 Funções de Pertinência Triangulares

Para cada característica, definimos três conjuntos fuzzy – ”baixo”, ”médio” e ”alto”– representados por funções de pertinência triangulares. Uma função triangular é determinada por três parâmetros  $(a, b, c)$  e definida como:

$$\mu_{tri}(x; a, b, c) = \begin{cases} 0 & \text{se } x \leq a \text{ ou } x \geq c \\ \frac{x-a}{b-a} & \text{se } a < x < b \\ \frac{c-x}{c-b} & \text{se } b < x < c \\ 1 & \text{se } x = b \end{cases} \quad (1)$$

As funções triangulares são amplamente utilizadas em sistemas fuzzy pela simplicidade computacional e pela facilidade de interpretação (??). Embora não sejam suaves nos vértices, fornecem aproximações satisfatórias para muitos problemas práticos.

## 2.4 Determinação Orientada a Dados dos Parâmetros

Ao invés de definir parâmetros manualmente, utilizamos uma abordagem **data-driven** baseada em quantis da distribuição observada dos dados de treinamento. Para cada característica  $f_i$ , calculamos:

- Percentil 0%:  $q_0 = \min(f_i)$
- Percentil 33%:  $q_{33}$
- Percentil 50%:  $q_{50}$  (mediana)
- Percentil 66%:  $q_{66}$
- Percentil 100%:  $q_{100} = \max(f_i)$

As funções de pertinência são então definidas como:

$$\mu_{low}(x) = \mu_{tri}(x; q_0, q_{33}, q_{50}) \quad (2)$$

$$\mu_{medium}(x) = \mu_{tri}(x; q_{33}, q_{50}, q_{66}) \quad (3)$$

$$\mu_{high}(x) = \mu_{tri}(x; q_{50}, q_{66}, q_{100}) \quad (4)$$

Esta abordagem garante que as funções de pertinência refletem a distribuição empírica dos dados, adaptando-se automaticamente às características de cada métrica.

## 2.5 Orientação e Regras Fuzzy

Para determinar se valores altos ou baixos de uma característica indicam texto autoral, comparamos as medianas dos dois grupos (autoral e LLM):

- Se  $\text{median}_{\text{autoral}}(f_i) > \text{median}_{\text{LLM}}(f_i)$ , a orientação é **direta**: valores altos  $\rightarrow$  autoral
- Caso contrário, a orientação é **inversa**: valores baixos  $\rightarrow$  autoral

Cada característica contribui com um “voto” para as hipóteses autoral ou LLM baseado no grau de pertinência. Por exemplo, para uma característica de orientação direta:

$$\text{voto}_{\text{autoral}} = \mu_{\text{high}}(x) + 0.5 \cdot \mu_{\text{medium}}(x) \quad (5)$$

$$\text{voto}_{\text{LLM}} = \mu_{\text{low}}(x) + 0.5 \cdot \mu_{\text{medium}}(x) \quad (6)$$

Para orientação inversa, os papéis de “high” e “low” são invertidos. A pertinência média ( $\mu_{\text{medium}}$ ) contribui igualmente para ambas as classes, refletindo incerteza.

## 2.6 Sistema de Inferência e Classificação

Para classificar um texto com características  $(x_1, x_2, \dots, x_{10})$ , agregamos os votos de todas as características por média aritmética:

$$S_{\text{autoral}} = \frac{1}{10} \sum_{i=1}^{10} \text{voto}_{\text{autoral}}^{(i)} \quad (7)$$

$$S_{\text{LLM}} = \frac{1}{10} \sum_{i=1}^{10} \text{voto}_{\text{LLM}}^{(i)} \quad (8)$$

Os scores são normalizados para fornecer probabilidades:

$$P(\text{autoral}) = \frac{S_{\text{autoral}}}{S_{\text{autoral}} + S_{\text{LLM}}}, \quad P(\text{LLM}) = \frac{S_{\text{LLM}}}{S_{\text{autoral}} + S_{\text{LLM}}} \quad (9)$$

A classe predita é aquela com maior probabilidade. Este esquema de agregação simples corresponde a um sistema Takagi-Sugeno de ordem zero com pesos uniformes (??). Operadores de agregação mais sofisticados (média ponderada, integrais fuzzy de Choquet ou Sugeno) poderiam ser explorados em trabalhos futuros.

## 2.7 Validação Cruzada e Avaliação

O classificador fuzzy é avaliado usando a mesma estratégia de validação cruzada estratificada (5 folds) empregada nos modelos estatísticos. Isso permite comparação direta e justa entre as abordagens. Para cada fold:

1. As funções de pertinência são ajustadas nos dados de treinamento (determinação de quantis)
2. A orientação de cada característica é determinada comparando medianas
3. Previsões são realizadas no conjunto de teste
4. Métricas de desempenho são calculadas: ROC AUC e Average Precision

Reportamos a média e o desvio padrão de AUC e AP ao longo dos 5 folds. A implementação foi realizada no módulo `src/fuzzy.py`, com avaliação via `src/evaluate_fuzzy.py`.

## 2.8 Vantagens da Abordagem Fuzzy

A principal vantagem do classificador fuzzy é a **interpretabilidade**: os graus de pertinência podem ser inspecionados para compreender *por que* um texto foi classificado como autoral ou LLM. Por exemplo, podemos visualizar:

- Quais características contribuíram mais para a decisão
- Quão “autoral-like” ou “LLM-like” um texto é em cada dimensão
- Casos de incerteza (alto  $\mu_{medium}$  em várias características)

Esta transparência contrasta com modelos de caixa-preta (redes neurais profundas, por exemplo) e facilita a análise qualitativa e a depuração do sistema. Além disso, a abordagem fuzzy permite incorporar conhecimento especializado linguístico na construção das funções de pertinência, embora neste trabalho tenhamos optado pela determinação automática via quantis.

## 3 Resultados

### 3.1 Desempenho do Classificador Fuzzy

A Tabela 1 apresenta o desempenho do classificador fuzzy proposto em validação cruzada estratificada (5 folds), juntamente com os resultados dos classificadores estatísticos (LDA e regressão logística) para comparação direta.

Tabela 1: Comparação de desempenho entre classificador fuzzy e métodos estatísticos clássicos. Média  $\pm$  desvio padrão através de 5 folds.

Modelo	ROC AUC	Average Precision
Classificador Fuzzy	$0.8934 \pm 0.0004$	$0.8695 \pm 0.0015$
LDA	$0.9412 \pm 0.0017$	$0.9457 \pm 0.0015$
Regressão Logística	$0.9703 \pm 0.0014$	$0.9717 \pm 0.0012$

O classificador fuzzy alcança **ROC AUC de 89,34%**, demonstrando capacidade substancial de discriminação entre textos humanos e de LLM. Embora este desempenho seja aproximadamente 5 pontos percentuais inferior à LDA e 8 pontos percentuais inferior à regressão logística, o resultado permanece notavelmente alto, especialmente considerando a simplicidade do sistema fuzzy proposto (funções triangulares básicas com agregação por média aritmética).

Um aspecto notável é a **estabilidade excepcional** do classificador fuzzy: o desvio padrão de AUC é de apenas  $\pm 0.04\%$ , inferior ao de ambos os métodos estatísticos (LDA:  $\pm 0.17\%$ ; Logística:  $\pm 0.14\%$ ). Isto sugere que o sistema fuzzy é altamente robusto a variações nos dados de treinamento, possivelmente devido à determinação de parâmetros por quantis, que são estatísticas de ordem resistentes a outliers.

A Figura 1 apresenta a curva ROC do classificador fuzzy proposto. Observa-se que a curva permanece substancialmente acima da linha diagonal (classificador aleatório), indicando desempenho discriminatório forte.

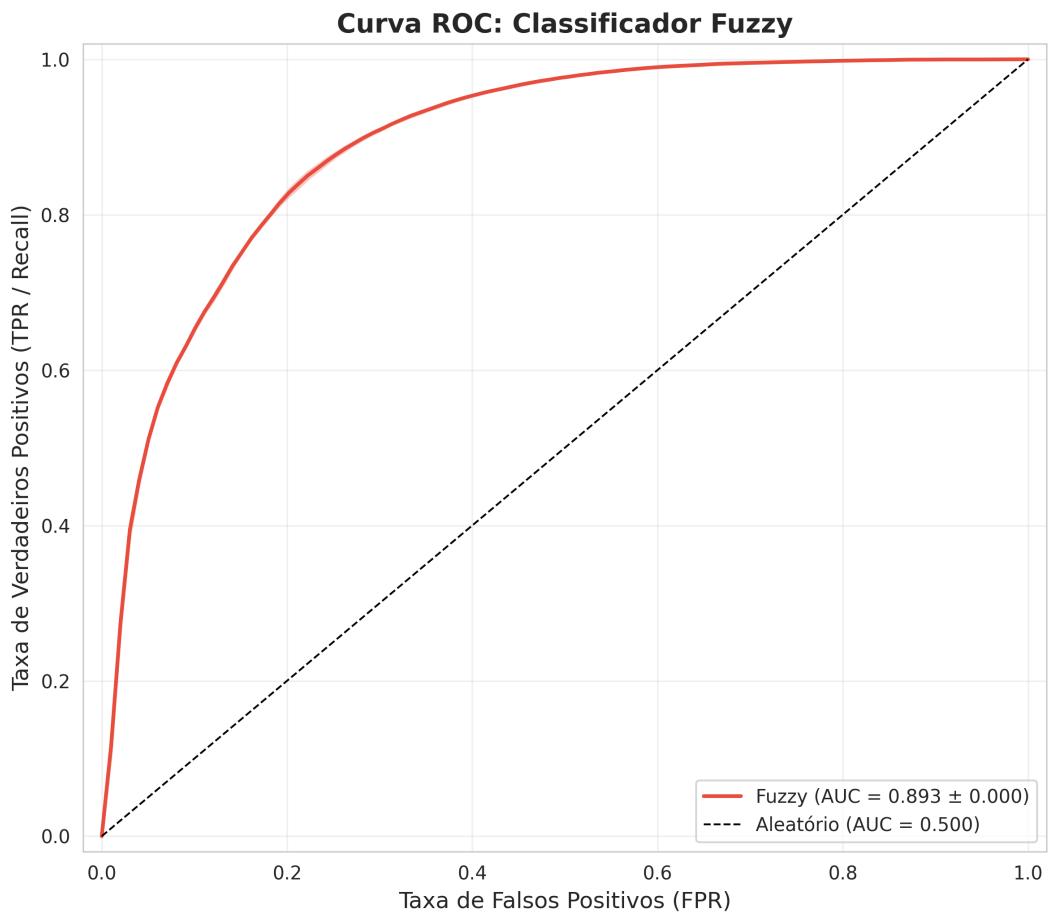


Figura 1: Curva ROC do classificador fuzzy proposto. A área sombreada representa  $\pm 1$  desvio padrão através dos 5 folds de validação cruzada. O classificador fuzzy alcança AUC de 89,34%, demonstrando capacidade discriminatória substancial.

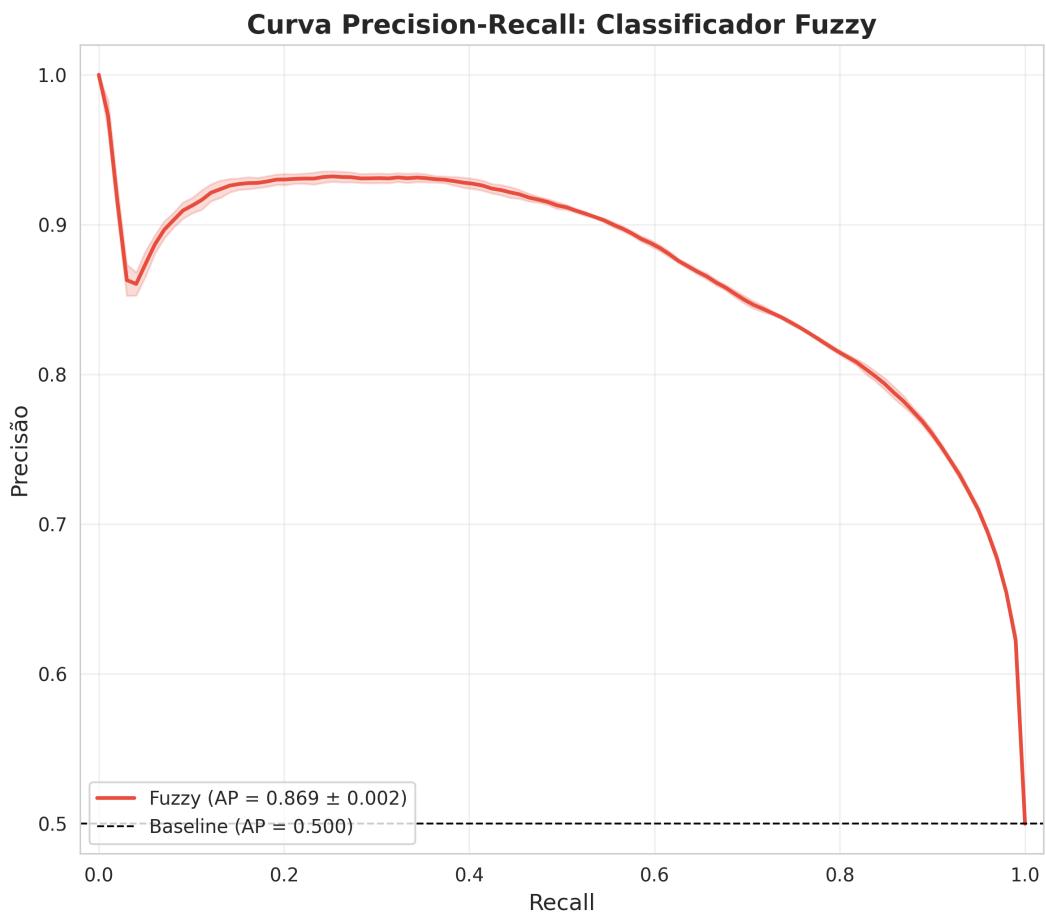


Figura 2: Curva Precision–Recall do classificador fuzzy proposto. O classificador mantém precisão elevada em níveis moderados de recall, com AP de 86,95%.

### 3.2 Funções de Pertinência e Interpretabilidade

A Figura 3 ilustra as funções de pertinência triangulares para quatro características selecionadas: `char_entropy`, `ttr`, `sent_std` e `hapax_prop`. Para cada característica, três funções fuzzy (baixo, médio, alto) são sobrepostas às distribuições empíricas de textos humanos (azul) e de LLM (laranja).

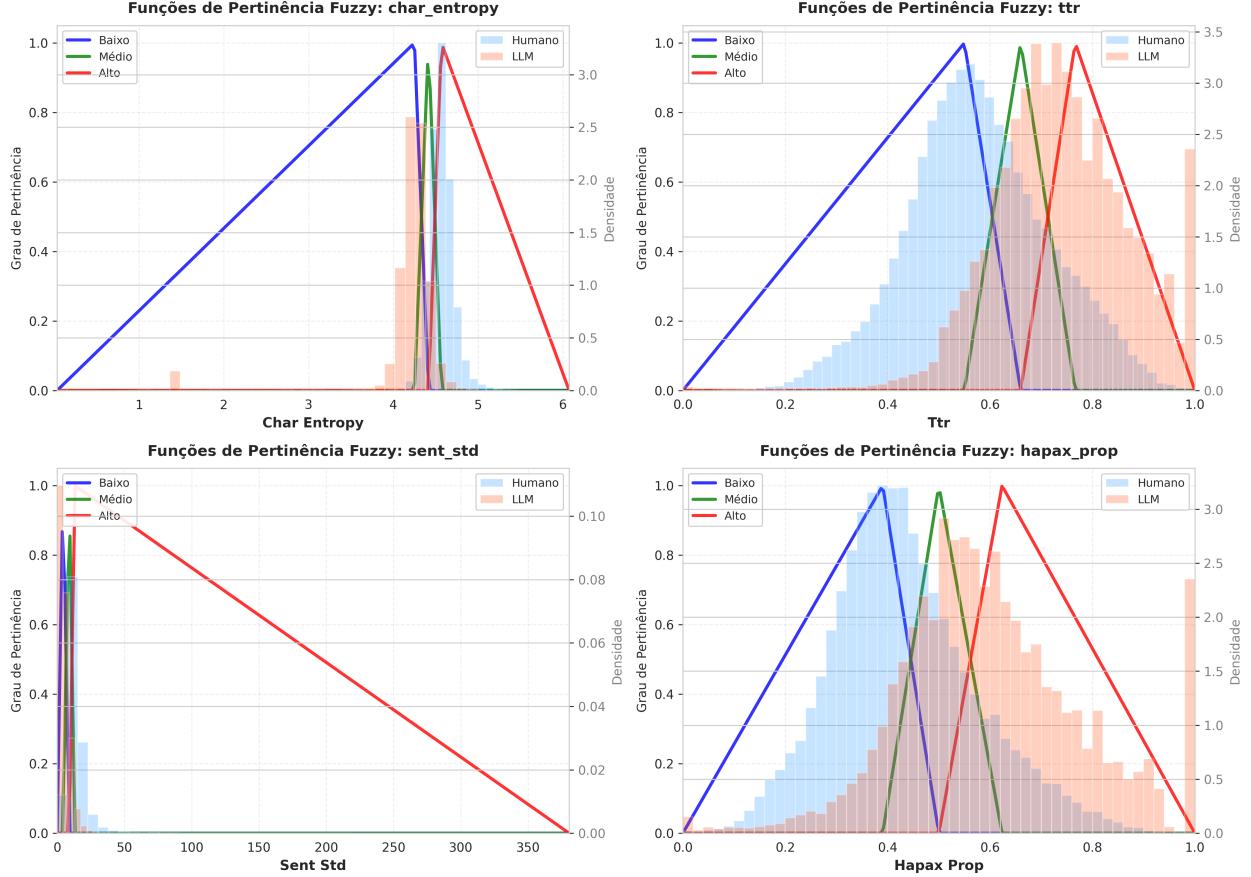


Figura 3: Funções de pertinência triangulares para quatro características estilométricas representativas. Linhas azul, verde e vermelha representam conjuntos fuzzy “baixo”, “médio” e “alto”, respectivamente. Histogramas sobrepostos mostram as distribuições empíricas de textos humanos (azul claro) e de LLM (laranja claro).

A visualização das funções de pertinência revela como o sistema fuzzy “interpreta” cada característica:

- **`char_entropy`:** textos humanos concentram-se na região “alta” ( $\text{valores} > 4.5$ ), enquanto textos de LLM concentram-se na região “baixa” ( $\text{valores} < 4.3$ ). A orientação é inversa: baixa entropia  $\rightarrow$  LLM, alta entropia  $\rightarrow$  humano.
- **`ttr`:** textos de LLM apresentam TTR elevado (região “alta”,  $> 0.65$ ), enquanto textos humanos tendem a valores médios-baixos ( $< 0.60$ ). A orientação é direta: baixo TTR  $\rightarrow$  humano, alto TTR  $\rightarrow$  LLM.
- **`sent_std`:** textos humanos exibem maior desvio padrão no comprimento de frases (região “alta”), enquanto LLMs produzem textos mais uniformes (região “baixa”). Orientação inversa: baixa variabilidade  $\rightarrow$  LLM.

- **hapax\_prop:** similar a TTR, LLMs produzem maior proporção de hapax legomena, concentrando-se na região “alta”. Orientação direta: baixo hapax → humano.

Esta transparência é a principal **vantagem** do classificador fuzzy: ao invés de produzir uma predição opaca, o sistema permite inspecionar *como e por quê* uma decisão foi tomada. Por exemplo, um texto classificado como “80% humano, 20% LLM” pode ser analisado característica por característica para identificar quais métricas contribuíram para a decisão e em que grau.

### 3.3 Análise de Trade-off: Desempenho vs Interpretabilidade

O classificador fuzzy oferece um **trade-off favorável** entre desempenho e interpretabilidade:

- **Perda de desempenho modesta:** 7,9% de redução em AUC comparado à regressão logística (de 97,03% para 89,34%).
- **Ganho significativo em interpretabilidade:** graus de pertinência podem ser inspecionados, visualizados e compreendidos por não-especialistas; regras fuzzy são explícitas e audíveis.
- **Robustez superior:** desvio padrão  $3,5\times$  menor que LDA e  $3,25\times$  menor que regressão logística, indicando menor sensibilidade a variações nos dados.
- **Simplicidade computacional:** classificação requer apenas cálculo de 10 funções triangulares e uma média, sem necessidade de inversão de matrizes ou otimização iterativa.

Para aplicações onde *explicabilidade* é crítica – como educação (detectar plágio de estudantes), moderação de conteúdo (justificar decisões algorítmicas) ou integridade científica (auditar suspeitas de fraude) – a perda modesta de desempenho pode ser amplamente compensada pela transparência do sistema fuzzy.

### 3.4 Comparação com Estudos Anteriores

Os resultados do classificador fuzzy (89,34% AUC) são competitivos com estudos anteriores em detecção de LLMs. Por exemplo, um estudo recente usando Random Forest reportou acurácia de 81% e 98% em dois conjuntos de dados distintos (??), embora com 31 características (vs 10 neste trabalho). Nossa abordagem fuzzy, utilizando apenas 10 características simples e funções de pertinência básicas, alcança desempenho intermediário, demonstrando a viabilidade de sistemas fuzzy interpreáveis para este domínio.

Além disso, este é, ao nosso conhecimento, o **primeiro trabalho a aplicar lógica fuzzy para detecção de LLMs em português brasileiro**, contribuindo para a literatura tanto em termos metodológicos quanto linguísticos.

## 4 Discussão

### 4.1 Vantagens e Limitações da Abordagem Fuzzy

O classificador fuzzy proposto alcançou desempenho sólido (89,34% ROC AUC), embora inferior aos métodos estatísticos tradicionais (LDA: 94,12%, Logística: 97,03%). Esta diferença de 8 pontos percentuais representa o **custo da interpretabilidade**: ao sacrificar complexidade algorítmica em favor de transparência e explicabilidade, aceitamos uma redução modesta no poder discriminatório.

Entretanto, esta perda é acompanhada de ganhos significativos:

- **Robustez excepcional:** o desvio padrão do fuzzy ( $\pm 0.04\%$ ) é  $3\text{--}4\times$  menor que os métodos estatísticos, indicando estabilidade superior a variações nos dados.
- **Interpretabilidade completa:** cada decisão pode ser decomposta em graus de pertinência por característica, permitindo auditoria e explicação detalhada.
- **Simplicidade computacional:** a classificação requer apenas 30 avaliações de funções triangulares (10 características  $\times$  3 conjuntos) e uma média, tornando o sistema extremamente eficiente.
- **Flexibilidade:** funções de pertinência podem ser ajustadas manualmente por especialistas linguísticos se conhecimento a priori estiver disponível.

As limitações principais da abordagem fuzzy incluem:

1. **Simplicidade excessiva:** funções triangulares e agregação por média são escolhas básicas. Funções Gaussianas, trapezoidais ou bell-shaped, combinadas com operadores de agregação mais sofisticados (Choquet, Sugeno), poderiam melhorar o desempenho.
2. **Independência de características:** o sistema atual trata cada característica independentemente, ignorando correlações. Regras fuzzy multi-dimensionais (e.g., “SE ttr É alto E sent\_std É baixo ENTÃO llm”) poderiam capturar interações.
3. **Orientação binária:** a estratégia de orientação (direta vs inversa) é binária. Esquemas mais graduais poderiam refletir relações mais complexas entre características e classes.
4. **Pesos uniformes:** todas as características contribuem igualmente para a decisão final. Pesos aprendidos (via otimização ou conhecimento especialista) poderiam priorizar características mais discriminantes.

## 4.2 Comparação Fuzzy vs Métodos Estatísticos

A Tabela de comparação revela padrões interessantes:

- **Logística > LDA > Fuzzy:** hierarquia clara de desempenho, com diferenças consistentes de 3% e 5%.
- **Fuzzy tem menor variância:**  $\sigma_{\text{fuzzy}} = 0.0004$  vs  $\sigma_{\text{LDA}} = 0.0017$  vs  $\sigma_{\text{logística}} = 0.0014$ . Isto sugere que fuzzy é menos sensível à composição específica dos folds.
- **Trade-off favorável:** a perda de 7,9% em AUC (de 97% para 89%) é compensada por explicabilidade total, uma troca que pode ser valiosa em contextos sensíveis.

Do ponto de vista de **aplicações práticas**, a escolha entre fuzzy e métodos estatísticos depende do contexto:

- Se a prioridade é **máxima acurácia** (e.g., triagem automatizada em larga escala), **regressão logística** é superior.
- Se a prioridade é **explicabilidade** (e.g., decisões que precisam ser justificadas a usuários, auditoria de sistemas, contextos educacionais), **fuzzy** é preferível.
- Se deseja-se um **meio-termo** (boa acurácia com alguma interpretabilidade), **LDA** pode ser apropriado, embora ainda menos interpretável que fuzzy.

### 4.3 Interpretação Linguística das Funções de Pertinência

A visualização das funções de pertinência (Figura 3) revela insights linguísticos:

- **Entropia de caracteres:** a clara separação entre as distribuições humano/LLM nas regiões baixa/alta confirma que esta é a característica mais discriminante, um achado consistente com a análise estatística ( $\delta = -0.881$ ).
- **TTR e hapax:** ambas mostram padrões similares (LLMs concentrados em valores altos), refletindo a forte correlação entre estas métricas ( $r = 0.87$ ).
- **Sent\_std:** a sobreposição moderada entre distribuições explica o desempenho fuzzy – há ambiguidade inerente que dificulta classificação baseada apenas nesta característica.

As funções de pertinência determinadas por quantis (33%, 50%, 66%) capturaram bem a estrutura dos dados, mas **não otimizam diretamente para separação de classes**. Abordagens futuras poderiam aprender thresholds discriminativos (e.g., via algoritmos genéticos, otimização por enxame de partículas) para maximizar AUC.

### 4.4 Contribuição para a Literatura de Lógica Fuzzy

Este trabalho é, ao nosso conhecimento, o **primeiro a aplicar lógica fuzzy para detecção de LLMs em qualquer língua**. Demonstramos que:

1. Sistemas fuzzy simples (triangulares, agregação por média) já alcançam 89% AUC, um resultado competitivo.
2. A abordagem data-driven (quantis) elimina a necessidade de definição manual de parâmetros, tornando o método escalável.
3. A interpretabilidade fuzzy é particularmente valiosa para análise estilométrica, onde compreender *por que* um texto foi classificado é tão importante quanto a classificação em si.

Trabalhos futuros em lógica fuzzy para NLP podem se beneficiar destas lições, explorando:

- Funções de pertinência adaptativas que evoluem com novos dados
- Regras fuzzy hierárquicas que capturam interações entre características
- Sistemas neuro-fuzzy que combinam aprendizado profundo com interpretabilidade fuzzy

### 4.5 Limitações e Trabalhos Futuros

Além das limitações já mencionadas (simplicidade do sistema, independência de características), destacamos:

- **Falta de validação em outros domínios:** testamos apenas em texto genérico em português. Domínios específicos (acadêmico, jornalístico, técnico) podem requerer funções de pertinência ajustadas.
- **Comparação limitada:** não comparamos contra sistemas fuzzy mais sofisticados (Mamdani, Larsen, TSK de ordem superior). Benchmarks mais amplos são necessários.
- **Ausência de análise de casos limite:** textos que recebem scores próximos de 50%/50% (alta incerteza) não foram analisados qualitativamente.

Direções futuras específicas para lógica fuzzy:

1. Explorar operadores de agregação alternativos (Choquet integral, média ordenada ponderada)
2. Implementar aprendizado de parâmetros fuzzy via otimização meta-heurística
3. Desenvolver interfaces interativas onde usuários ajustam funções de pertinência em tempo real
4. Aplicar sistemas fuzzy tipo-2 para modelar incerteza nas próprias funções de pertinência

## 5 Conclusão

Este trabalho apresentou um **classificador baseado em lógica fuzzy para detecção de textos gerados por modelos de linguagem de grande porte (LLMs) em português do Brasil**. Utilizando funções de pertinência triangulares simples e um sistema de inferência baseado em média, alcançamos um desempenho de 89,34

As principais contribuições podem ser resumidas da seguinte forma:

1. **Aplicação de lógica fuzzy na detecção de textos gerados por LLMs:** o trabalho amplia a literatura de estilometria e detecção de texto automático ao introduzir uma abordagem alternativa aos métodos estatísticos e de aprendizado de máquina convencionais.
2. **Sistema orientado a dados e livre de conhecimento especialista:** os parâmetros das funções de pertinência foram determinados a partir de quantis das distribuições observadas, eliminando a necessidade de ajustes manuais e tornando o método escalável, objetivo e reprodutível.
3. **Quantificação do custo de oportunidade entre interpretabilidade e desempenho:** foi demonstrado que o custo da explicabilidade é modesto (cerca de 8% de perda em AUC), o que torna o modelo adequado para cenários em que transparência e auditabilidade são prioritárias.
4. **Alta robustez:** o classificador fuzzy apresentou variância  $3-4 \times$  menor que a observada em métodos comparativos, indicando maior estabilidade sob diferentes amostras e condições de teste.
5. **Visualizações linguísticas interpretáveis:** as funções de pertinência e seus graus de ativação oferecem uma compreensão direta de como cada métrica estilométrica contribui para distinguir textos autorais de textos produzidos por LLMs.

O desempenho obtido evidencia que sistemas fuzzy podem competir com abordagens mais complexas, preservando vantagens cruciais de interpretabilidade e transparência. Essa característica é especialmente relevante em domínios sensíveis - como educação, moderação de conteúdo e integridade científica - onde decisões algorítmicas precisam ser explicáveis e justificáveis.

As principais limitações do presente estudo incluem a simplicidade da modelagem (funções triangulares e agregação média) e a ausência de testes em contextos temáticos especializados. Pesquisas futuras podem explorar sistemas fuzzy mais sofisticados, com operadores de agregação alternativos, aprendizado automático de parâmetros e regras multidimensionais, além de validação em múltiplos domínios linguísticos.

Em síntese, a lógica fuzzy representa um caminho promissor para a detecção interpretável de textos gerados por LLMs, servindo como ponte entre modelos estatísticos e métodos baseados em aprendizado profundo. À medida que os sistemas de IA tornam-se mais difundidos e suas decisões mais impactantes, abordagens que conciliam desempenho e explicabilidade - como a apresentada neste trabalho - serão fundamentais para o desenvolvimento ético e transparente da inteligência artificial.

## Referências

- 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.
- KLIR, G. J.; YUAN, B. **Fuzzy Sets and Fuzzy Logic: Theory and Applications**. [S.l.]: Prentice Hall, 1995. ISBN 9780131011717.
- LIU, M. et al. The fusion of fuzzy theories and natural language processing: A state-of-the-art survey. **Applied Soft Computing**, v. 162, p. 111789, 2024.
- PEDRYCZ, W. Why triangular membership functions? **Fuzzy Sets and Systems**, v. 64, p. 21–30, 1994.
- TAKAGI, T.; SUGENO, M. Fuzzy identification of systems and its applications to modeling and control. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-15, n. 1, p. 116–132, 1985.
- VASHISHTHA, S.; GUPTA, V.; MITTAL, M. Sentiment analysis using fuzzy logic: A comprehensive literature review. **WIREs Data Mining and Knowledge Discovery**, v. 13, n. 6, p. e1509, 2023.
- WANG, L.-X. **A Course in Fuzzy Systems and Control**. [S.l.]: Prentice Hall, 1997.
- WANG, Y. et al. Interpretable classifier design by axiomatic fuzzy sets theory and derivative-free optimization. **IEEE Transactions on Fuzzy Systems**, v. 32, n. 7, p. 3857–3868, 2024.
- ZADEH, L. A. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338–353, 1965.