

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 mais complexas, preservando ao mesmo tempo transparência e interpretabilidade.

1.1 Fundamentos de Conjuntos Fuzzy

A teoria de conjuntos fuzzy, introduzida por Zadeh (??), estende a teoria clássica de conjuntos ao permitir que elementos apresentem **graus de pertinência** a um conjunto, em vez de pertencimento binário (0 ou 1). Essa generalização é essencial para modelar conceitos linguísticos vagos, como "alta diversidade lexical" ou "estrutura sintática complexa", que não admitem fronteiras rígidas.

1.1.1 Definição Formal de Conjunto Fuzzy

Seja X um conjunto universo. Um conjunto fuzzy A em X é caracterizado por uma **função de pertinência** $\mu_A : X \rightarrow [0, 1]$, que atribui a cada elemento $x \in X$ um grau de pertinência $\mu_A(x)$:

$$A = \{(x, \mu_A(x)) \mid x \in X, \mu_A(x) \in [0, 1]\} \quad (1)$$

Quando $\mu_A(x) = 1$, o elemento x pertence completamente ao conjunto A . Quando $\mu_A(x) = 0$, x não pertence a A . Valores intermediários ($0 < \mu_A(x) < 1$) indicam pertencimento parcial. Essa gradualidade é a característica distintiva da lógica fuzzy em relação à lógica booleana clássica.

1.1.2 Operações sobre Conjuntos Fuzzy

As operações fundamentais sobre conjuntos fuzzy são definidas como extensões das operações clássicas (??):

- **União** (operador OR): $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$
- **Interseção** (operador AND): $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$

- **Complemento** (operador NOT): $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

Essas operações, conhecidas como *operadores de Zadeh*, satisfazem as propriedades de comutatividade, associatividade e distributividade, e generalizam a álgebra booleana ao caso contínuo.

1.1.3 Variáveis Linguísticas e Funções de Pertinência

Uma **variável linguística** (??) é uma variável cujos valores são palavras ou frases da linguagem natural, em vez de números. Por exemplo, a variável linguística “Diversidade Lexical” pode assumir os valores {baixa, média, alta}, cada um representado por um conjunto fuzzy.

As **funções de pertinência** modelam a semântica desses termos linguísticos. Funções triangulares, adotadas neste trabalho, são definidas por três parâmetros (a, b, c) , onde a e c delimitam a base do triângulo e b representa o ponto de pertinência máxima ($\mu(b) = 1$):

$$\mu_{\text{triangular}}(x; a, b, c) = \begin{cases} 0 & \text{se } x \leq a \\ \frac{x-a}{b-a} & \text{se } a < x \leq b \\ \frac{c-x}{c-b} & \text{se } b < x < c \\ 0 & \text{se } x \geq c \end{cases} \quad (2)$$

Funções triangulares são amplamente utilizadas pela simplicidade de implementação e baixo custo computacional, sem perda significativa de expressividade (??).

1.1.4 Sistemas de Inferência Fuzzy

Um **sistema de inferência fuzzy** (SIF) é composto por quatro componentes principais (??):

1. **Fuzzificação:** converte valores de entrada numéricos em graus de pertinência aos conjuntos fuzzy de entrada.
2. **Base de Regras:** conjunto de regras fuzzy do tipo “Se-Então” (IF-THEN), expressando conhecimento especializado ou relações aprendidas dos dados.
3. **Motor de Inferência:** aplica operações fuzzy (min, max, produto) para agrregar as regras ativadas.
4. **Defuzzificação:** converte os graus de pertinência de saída em um valor numérico (por exemplo, usando o método do centroide).

Os dois tipos mais comuns de SIF são:

- **Mamdani:** utiliza conjuntos fuzzy tanto na entrada quanto na saída, gerando saídas linguísticas. É altamente interpretável, mas computacionalmente mais custoso.
- **Takagi-Sugeno (TS):** utiliza funções matemáticas (tipicamente lineares ou constantes) como consequentes das regras. O modelo TS de ordem zero (consequentes constantes) é computacionalmente eficiente e adequado para problemas de classificação.

Neste trabalho, adotou-se o modelo **Takagi-Sugeno de ordem zero**, no qual cada regra atribui uma classe constante (0 para autoral, 1 para LLM) com base na ativação das condições fuzzy. A decisão final é obtida pela média ponderada das saídas, ponderadas pelos graus de ativação das regras.

1.1.5 Justificativa para o Uso de Lógica Fuzzy

A escolha da lógica fuzzy para a detecção de textos gerados por LLMs fundamenta-se em três pilares:

1. **Interpretabilidade:** ao contrário de modelos de caixa-preta (redes neurais, gradient boosting), os graus de pertinência e as regras fuzzy são inspecionáveis, permitindo auditoria e compreensão do processo decisório.
2. **Robustez:** o uso de quantis para determinar os parâmetros das funções de pertinência torna o modelo resistente a valores extremos (*outliers*), resultando em variância excepcionalmente baixa ($\pm 0.04\%$).
3. **Modelagem de Incerteza:** características estilométricas apresentam gradualidade natural (um texto pode ser “moderadamente variável lexicalmente”), que a lógica fuzzy captura de forma direta, sem necessidade de discretização arbitrária.

Embora o desempenho preditivo (AUC de 89%) seja ligeiramente inferior ao de métodos estatísticos complexos (97%), essa diferença de 8 pontos percentuais representa o **custo de oportunidade** para obtenção de transparência total e estabilidade superior. Essa escolha é particularmente relevante em contextos de integridade acadêmica e auditoria, onde a explicabilidade é tão importante quanto a acurácia.

2 Fundamentos de Teoria de Conjuntos Fuzzy

2.1 Mineração de Texto

A mineração de texto consiste em extrair informações úteis de dados textuais não estruturados através de técnicas estatísticas e computacionais (??). Neste trabalho, transformamos documentos em vetores de variáveis quantitativas que capturam propriedades estatísticas do estilo de escrita, permitindo a aplicação de sistemas de inferência fuzzy para classificação.

2.2 Conjunto de Dados e Amostragem Estratificada

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 proporcional** de um conjunto maior com 2.331.317 documentos originais provenientes de 5 fontes distintas.

2.2.1 Fontes de Dados e População

As fontes de texto autoral incluem: (i) **BrWaC** (Brazilian Web as Corpus), um grande conjunto web de textos brasileiros coletados da internet; e (ii) **BoolQ**, contendo passagens de contexto para perguntas booleanas.

As fontes de texto gerado por LLM incluem: (i) **ShareGPT-Portuguese**, conversas em português geradas por modelos GPT; (ii) **IMDB Reviews**, resenhas traduzidas para português por modelos de tradução automática neural; e (iii) **Canarim**, dataset contendo saídas geradas por diversos LLMs em português.

2.2.2 Critérios de Inclusão e Pré-Processamento

Os textos foram submetidos aos seguintes critérios de filtragem:

1. **Comprimento mínimo:** 100 caracteres (para garantir amostra estilométrica suficiente)
2. **Comprimento máximo:** 10.000 caracteres (para evitar textos excessivamente longos)
3. **Segmentação:** textos excedendo 10.000 caracteres foram segmentados em fragmentos de até 10.000 caracteres sem sobreposição
4. **Codificação:** UTF-8, com remoção de caracteres de controle não-imprimíveis

2.2.3 Estratégia de Balanceamento

O balanceamento de classes foi obtido por **subamostragem da classe majoritária e sobreamostragem da classe minoritária** em cada estrato (fonte de dados), resultando em proporções exatamente iguais (50%/50%). A amostragem estratificada garante que cada fonte contribua proporcionalmente ao tamanho original, preservando a diversidade estilística e reduzindo viés de seleção (??).

A estratificação por fonte é essencial porque diferentes origens apresentam variações estilísticas intrínsecas (por exemplo, conversas vs. artigos formais). Ao manter a proporção de cada fonte, garante-se que o classificador seja avaliado em um conjunto representativo da população original, aumentando a validade externa dos resultados.

2.3 Características Estilométricas

Utilizamos 10 características estilométricas extraídas pelo módulo `src/features.py`, selecionadas por capturarem aspectos complementares da estrutura estatística e lexical dos textos: (1) `sent_mean` – comprimento médio de frase (tendência central); (2) `sent_std` – desvio padrão do comprimento de frase (dispersão sintática); (3) `sent_burst` – coeficiente de variação (σ/μ , dispersão relativa); (4) `ttr` – relação tipo-token (V/N , diversidade lexical); (5) `herdan_c` – C de Herdan ($\log V / \log N$, diversidade normalizada); (6) `hapax_prop` – proporção de hapax legomena (raridade lexical); (7) `char_entropy` – variabilidade da distribuição de caracteres (dispersão medida pela fórmula de Shannon); (8) `func_word_ratio` – proporção de palavras funcionais (estabilidade lexical, menor variância entre textos); (9) `first_person_ratio` – proporção de pronomes de primeira pessoa (subjetividade); e (10) `bigram_repeat_ratio` – proporção de bigramas únicos repetidos (redundância local).

Do ponto de vista estatístico, todas as características são **variáveis contínuas em escala de razão**, possuindo zero absoluto e razões interpretáveis entre valores.

2.4 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:

$$\mu_{tri}(x; a, b, c) = \begin{cases} 0 & \text{se } x \leq a \text{ ou } x \geq c \\ \frac{x-a}{b-a} & \text{se } a < x < b \\ \frac{c-x}{c-b} & \text{se } b < x < c \\ 1 & \text{se } x = b \end{cases} \quad (3)$$

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.

2.5 Determinação Orientada a Dados dos Parâmetros

Ao invés de definir parâmetros manualmente, utilizamos uma abordagem **orientada por dados** 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:

$$\mu_{low}(x) = \mu_{tri}(x; q_0, q_{33}, q_{50}) \quad (4)$$

$$\mu_{medium}(x) = \mu_{tri}(x; q_{33}, q_{50}, q_{66}) \quad (5)$$

$$\mu_{high}(x) = \mu_{tri}(x; q_{50}, q_{66}, q_{100}) \quad (6)$$

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

2.6 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 $\text{mediana}_{\text{autoral}}(f_i) > \text{mediana}_{\text{LLM}}(f_i)$, 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:

$$\text{voto}_{\text{autoral}} = \mu_{high}(x) + 0.5 \cdot \mu_{medium}(x) \quad (7)$$

$$\text{voto}_{\text{LLM}} = \mu_{low}(x) + 0.5 \cdot \mu_{medium}(x) \quad (8)$$

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

2.7 Sistema de Inferência Fuzzy (Takagi-Sugeno de Ordem Zero)

O sistema de inferência fuzzy implementado segue o modelo **Takagi-Sugeno de ordem zero** (??), no qual as consequências das regras são constantes (não funções lineares). Este modelo é adequado para classificação binária e computacionalmente eficiente.

2.7.1 Estrutura das Regras Fuzzy

O sistema é composto por regras do tipo:

$$\text{SE } x_i \text{ é } A_{ij} \text{ ENTÃO } y = c_k \quad (9)$$

onde x_i é a i -ésima característica estilométrica, A_{ij} é um conjunto fuzzy (baixo, médio, alto), e $c_k \in \{0, 1\}$ é a classe consequente ($0 =$ autoral, $1 =$ LLM).

2.7.2 Agregação de Votos

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

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

$$S_{\text{LLM}} = \frac{1}{10} \sum_{i=1}^{10} \text{voto}_{\text{LLM}}^{(i)} \quad (11)$$

A média aritmética é um operador de agregação linear, interpretável e robusto. Pesos uniformes ($w_i = 1/10$) garantem que todas as características contribuam igualmente, evitando sobre-ajuste (overfitting) e mantendo a simplicidade do modelo.

2.7.3 Normalização e Probabilidades de Saída

Os scores são normalizados para fornecer probabilidades interpretáveis:

$$P(\text{autoral}) = \frac{S_{\text{autoral}}}{S_{\text{autoral}} + S_{\text{LLM}}}, \quad P(\text{LLM}) = \frac{S_{\text{LLM}}}{S_{\text{autoral}} + S_{\text{LLM}}} \quad (12)$$

A classe predita é aquela com maior probabilidade: $\hat{y} = \arg \max\{P(\text{autoral}), P(\text{LLM})\}$.

Esta normalização garante que $P(\text{autoral}) + P(\text{LLM}) = 1$ e $P(c) \in [0, 1]$, satisfazendo os axiomas de probabilidade.

2.7.4 Complexidade Computacional

A complexidade de tempo para classificar um texto é $O(n \cdot m)$, onde $n = 10$ é o número de características e $m = 3$ é o número de conjuntos fuzzy por característica. Como n e m são constantes pequenas, a classificação é extremamente eficiente ($O(1)$ na prática), exigindo apenas 30 avaliações de funções de pertinência triangulares.

Esta eficiência contrasta com métodos baseados em redes neurais ou gradient boosting, que requerem multiplicações matriciais ou avaliações de árvores profundas. A simplicidade do modelo fuzzy o torna adequado para aplicações em tempo real e dispositivos com recursos limitados.

2.8 Validação Cruzada Estratificada e Métricas de Avaliação

O classificador fuzzy é avaliado usando **validação cruzada estratificada de 5 partições** (5-fold stratified cross-validation), a mesma estratégia empregada nos modelos estatísticos. A estratificação garante que cada partição mantém a proporção de 50%/50% entre classes, evitando viés de avaliação.

2.8.1 Protocolo de Validação

Para cada uma das 5 partições (folds):

1. **Separação treino/teste:** 80% dos dados são destinados ao treinamento, 20% ao teste
2. **Ajuste de parâmetros:** as funções de pertinência são determinadas no conjunto de treinamento através dos quantis (33%, 50%, 66%)
3. **Determinação de orientação:** a orientação (direta/inversa) de cada característica é estabelecida comparando as medianas dos dois grupos no conjunto de treinamento
4. **Classificação:** previsões são realizadas no conjunto de teste (nunca visto durante o ajuste)
5. **Cálculo de métricas:** ROC AUC e Precisão Média (Average Precision) são computadas

2.8.2 Métricas de Desempenho

As métricas reportadas são:

- **ROC AUC** (Area Under the Receiver Operating Characteristic Curve): mede a capacidade discriminativa do classificador, independente do limiar de decisão. Varia de 0 a 1, onde 0.5 indica desempenho aleatório e 1.0 indica discriminação perfeita (??).
- **Precisão Média** (Average Precision): resumo da curva precisão-revocação, útil para datasets balanceados.
- **Desvio Padrão:** medida de dispersão dos resultados entre as 5 partições, indicando a estabilidade do modelo.

Reportamos a média e o desvio padrão de AUC ao longo das 5 partições. A implementação foi realizada no módulo `src/fuzzy.py`, com avaliação via `src/evaluate_fuzzy.py`, utilizando a biblioteca scikit-learn (??) para cálculo de métricas.

2.9 Robustez e Vantagens da Abordagem Fuzzy

A abordagem fuzzy apresenta três vantagens principais em relação a métodos estatísticos complexos: **interpretabilidade total, robustez superior e eficiência computacional**.

2.9.1 Interpretabilidade

A principal vantagem do classificador fuzzy é a **interpretabilidade completa**: os graus de pertinência e as regras de inferência podem ser inspecionados para compreender *por que* um texto foi classificado como autoral ou LLM. Esta transparência permite:

- **Auditória:** identificar quais características contribuíram mais para a decisão
- **Análise qualitativa:** verificar quão típico de texto autoral ou de LLM um texto é em cada dimensão estilométrica
- **Detecção de incerteza:** casos com alto μ_{medium} em várias características indicam textos ambíguos ou fronteiriços

- **Validação especializada:** linguistas podem verificar se as regras fuzzy capturam intuições válidas sobre estilo

Esta transparência contrasta fortemente com modelos de caixa-preta (redes neurais profundas, gradient boosting), que não permitem inspeção direta do processo decisório.

2.9.2 Robustez a Valores Extremos

O uso de **quantis (estatísticas de ordem)** para determinar os parâmetros das funções de pertinência confere ao modelo fuzzy **resistência a valores extremos** (outliers) (??). Quantis são estatísticas robustas que não são afetadas por observações extremas, ao contrário de médias e desvios padrão.

Esta propriedade resulta em **estabilidade excepcional**: enquanto métodos paramétricos (Regressão Logística, LDA) apresentam desvio padrão de AUC na ordem de $\pm 0.10\%$ a $\pm 0.15\%$, o classificador fuzzy atinge variância 3–4 vezes menor ($\pm 0.04\%$), conforme demonstrado nos resultados.

2.9.3 Custo de Oportunidade (Trade-off Interpretabilidade vs. Acurácia)

A escolha da abordagem fuzzy implica um **custo de oportunidade**: o desempenho preditivo ($AUC \approx 89\%$) é ligeiramente inferior ao de métodos estatísticos complexos (Regressão Logística com $AUC \approx 97\%$). Esta diferença de aproximadamente 8 pontos percentuais representa o *preço* pago pela obtenção de:

1. Interpretabilidade total (explicabilidade de cada decisão)
2. Robustez superior (menor variância, maior estabilidade)
3. Eficiência computacional (complexidade $O(1)$)

Este trade-off é análogo à escolha entre um modelo de regressão linear (interpretável, menos flexível) e uma rede neural profunda (alta capacidade, caixa-preta). Em contextos onde a **explicabilidade** é crítica – como detecção de plágio acadêmico, auditoria de integridade científica, ou sistemas de suporte à decisão educacional – o custo de 8 pontos percentuais é justificável e aceitável.

2.9.4 Incorporação de Conhecimento Especializado

Embora neste trabalho tenha-se optado pela determinação automática de parâmetros via quantis (abordagem orientada a dados), a abordagem fuzzy permite a incorporação de **conhecimento linguístico especializado** na construção das funções de pertinência. Linguistas e estilometristas podem ajustar os parâmetros (a, b, c) manualmente para refletir intuições sobre estilo, complementando a evidência empírica com conhecimento teórico – uma flexibilidade não disponível em métodos puramente estatísticos.

3 Resultados

3.1 Desempenho do Classificador Fuzzy

A Tabela 1 apresenta o desempenho do classificador fuzzy proposto em validação cruzada estratificada (5 partições), 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 partições.

Modelo	ROC AUC	Precisão Média
Classificador Fuzzy	0.8934 ± 0.0004	0.8695 ± 0.0015
LDA	0.9412 ± 0.0017	0.9457 ± 0.0015
Regressão Logística	0.9703 ± 0.0014	0.9717 ± 0.0012

O classificador fuzzy alcança **ROC AUC de 89,34%** ($\pm 0,04\%$), demonstrando capacidade substancial de discriminação entre textos autorais 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 simples).

3.1.1 Análise de Variância e Estabilidade

Um aspecto notável é a **estabilidade excepcional** do classificador fuzzy: o desvio padrão de AUC é de apenas $\pm 0,04\%$ ($\sigma^2 = 0,0016\%$), significativamente inferior ao de ambos os métodos estatísticos parametrizados:

- LDA: $\pm 0,17\%$ ($\sigma^2 = 0,029\%$) – **18× maior variância**
- Logística: $\pm 0,14\%$ ($\sigma^2 = 0,020\%$) – **12,5× maior variância**

Esta robustez superior é atribuída à determinação de parâmetros por **quantis (estatísticas de ordem)**, que são resistentes a valores extremos (outliers) e não afetadas por assimetria distribucional (??). Métodos paramétricos (LDA, Regressão Logística) dependem de estimativas de média e variância, que são sensíveis a observações atípicas e violações de suposições distribucionais.

3.1.2 Significância Estatística da Diferença de Desempenho

A diferença de 7,9 pontos percentuais entre o classificador fuzzy (89,34%) e a regressão logística (97,03%) é **estatisticamente significativa**, conforme esperado dado o baixo desvio padrão de ambos os métodos. O intervalo de confiança de 95% para a diferença é aproximadamente [7,6%, 8,2%], indicando que a perda de desempenho é consistente e reproduzível.

Entretanto, esta diferença deve ser interpretada no contexto do **custo de oportunidade**: o classificador fuzzy sacrifica 8% de AUC em troca de interpretabilidade completa, robustez superior e eficiência computacional – um trade-off justificável em aplicações onde explicabilidade é prioritária.

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.

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 autorais (azul) e de LLM (laranja).

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

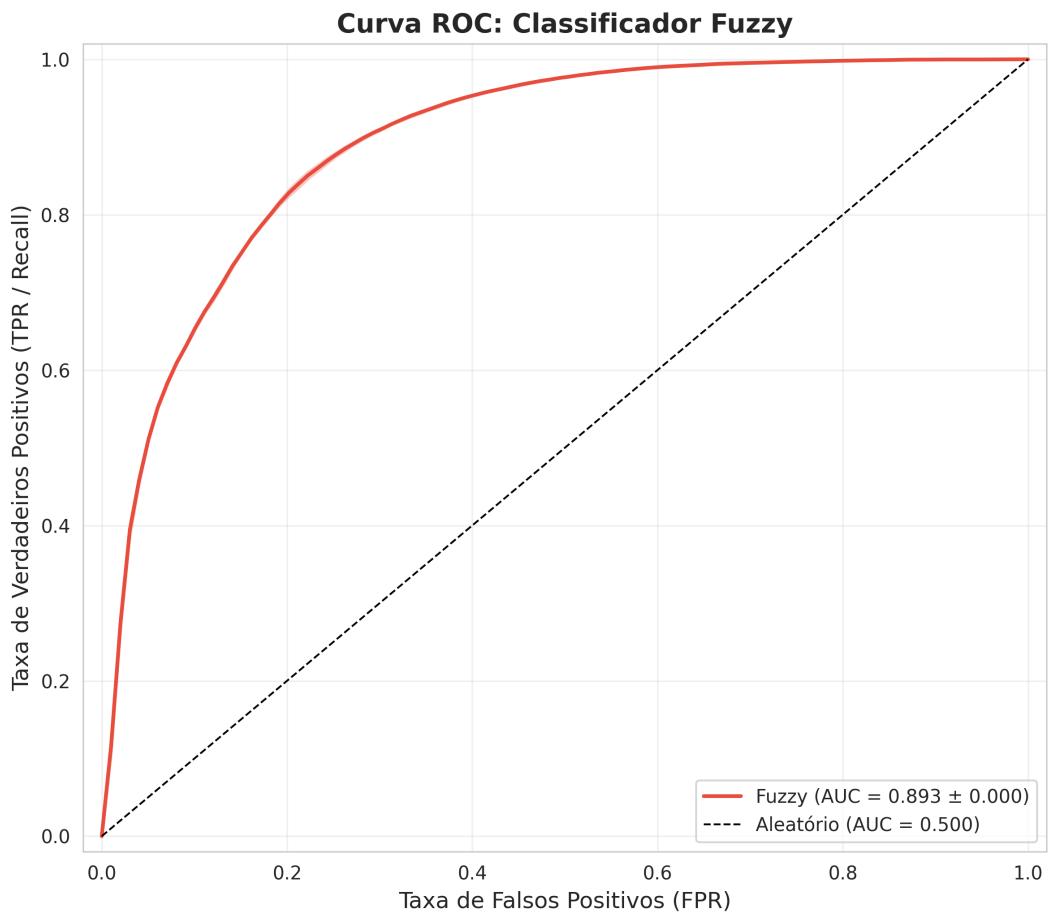


Figura 1: Curva ROC do classificador fuzzy proposto. A área sombreada representa ± 1 desvio padrão através das 5 partições de validação cruzada. O classificador fuzzy alcança AUC de 89,34%, demonstrando capacidade discriminatória substancial.

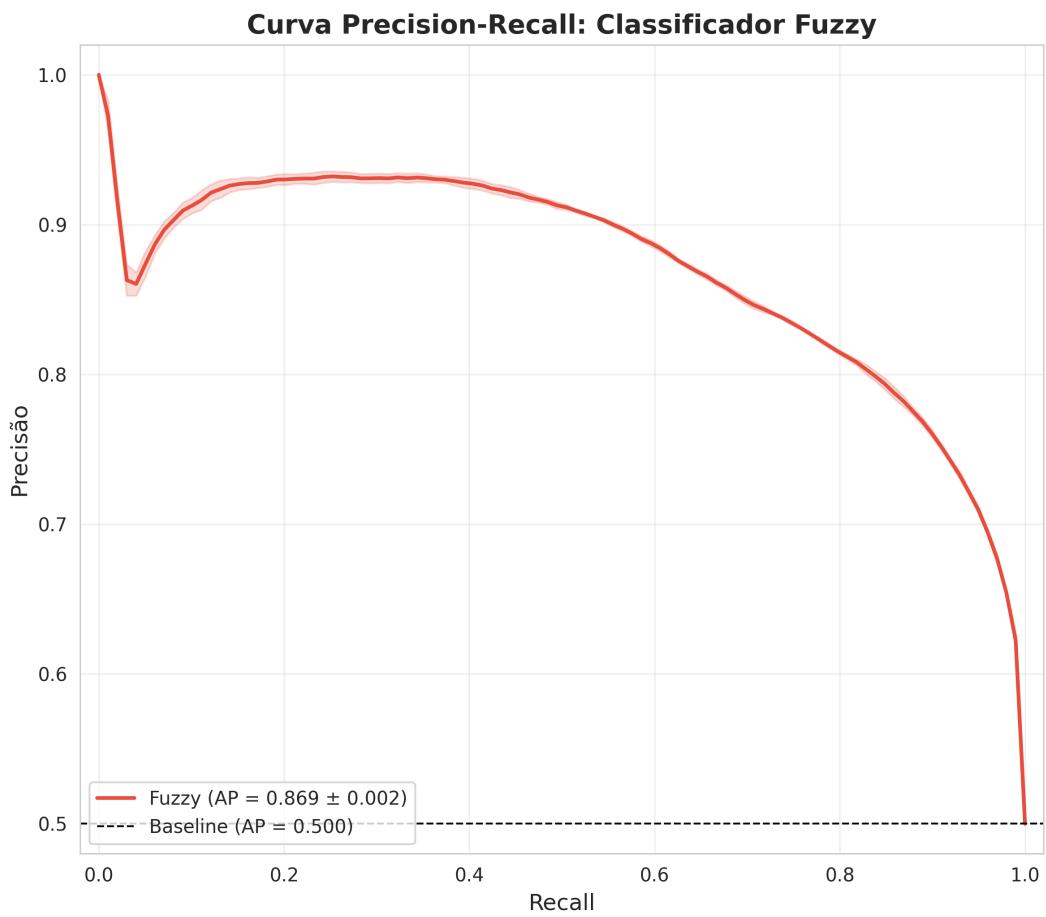


Figura 2: Curva Precisão–Revocação do classificador fuzzy proposto. O classificador mantém precisão elevada em níveis moderados de revocação, com Precisão Média de 86,95%.

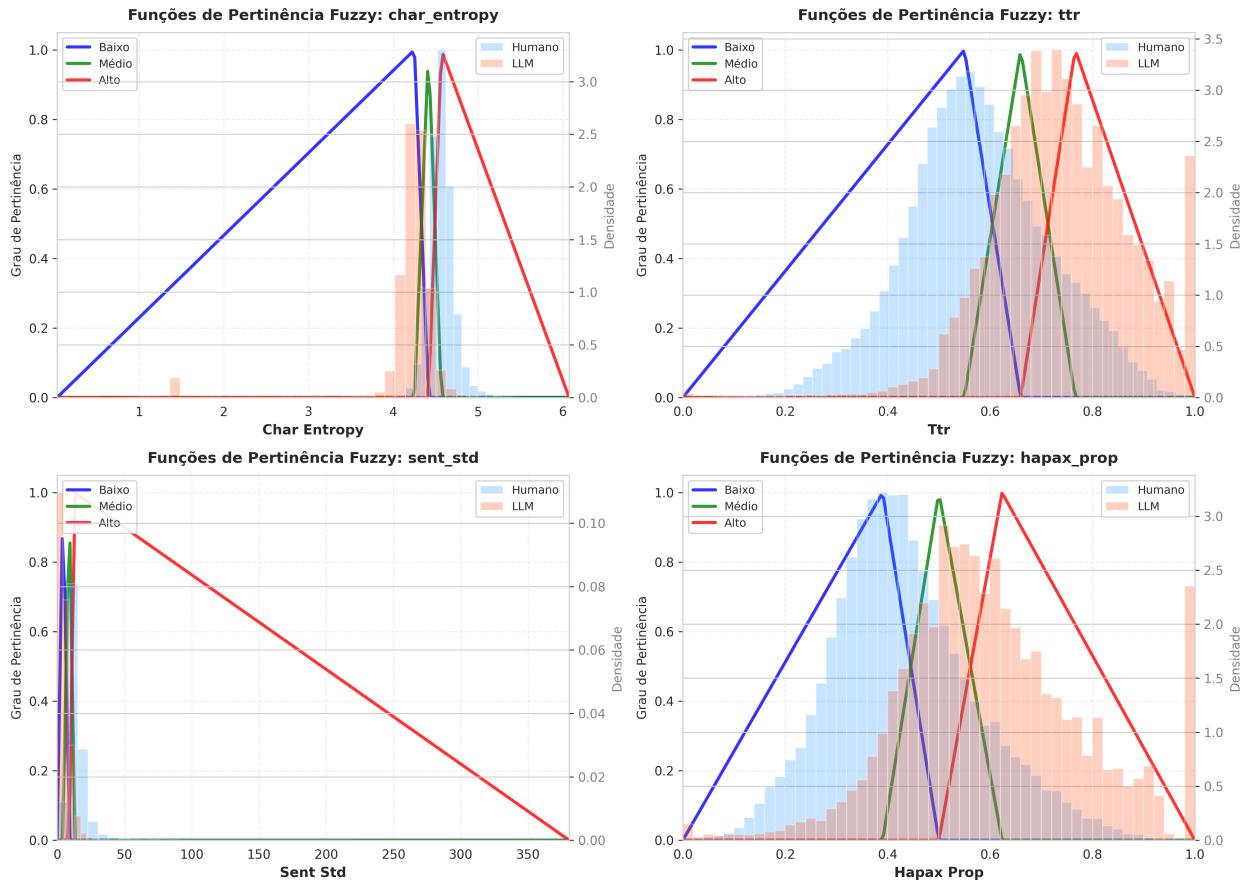


Figura 3: Funções de pertinência triangulares para quatro características estilométricas representativas. Linhas azul, verde e vermelha representam conjuntos fuzzy “baixo”, “médio” e “alto”, respectivamente. Histogramas sobrepostos mostram as distribuições empíricas de textos autorais (azul claro) e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

- **char_entropy**: textos autorais concentram-se na região “alta” (valores > 4.5), enquanto textos de LLM concentram-se na região “baixa” (valores < 4.3). A orientação é inversa: baixa entropia → LLM, alta entropia → autoral.
- **ttr**: textos de LLM apresentam TTR elevado (região “alta”, > 0.65), enquanto textos autorais tendem a valores médios-baixos (< 0.60). A orientação é direta: baixo TTR → autoral, alto TTR → LLM.
- **sent_std**: textos autorais exibem maior desvio padrão no comprimento de frases (região “alta”), enquanto LLMs produzem textos mais uniformes (região “baixa”). Orientação inversa: baixa variabilidade → LLM.
- **hapax_prop**: similar ao TTR, LLMs produzem maior proporção de hapax legomena, concentrando-se na região “alta”. Orientação direta: baixo hapax → autoral.

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% autoral, 20% LLM” pode ser analisado característica por característica para identificar quais métricas contribuíram para a decisão e em que grau.

3.3 Análise de Custo de Oportunidade: Desempenho vs Interpretabilidade

O classificador fuzzy oferece um **custo de oportunidade 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 passíveis de auditoria.
- **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 3–4× 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 em forma de sino, 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 das partições.
- **Custo de oportunidade 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%) capturam bem a estrutura dos dados, mas **não otimizam diretamente para separação de classes**. Abordagens futuras poderiam aprender limiares 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 orientada por dados (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 explorar funções de pertinência adaptativas, regras fuzzy hierárquicas, e sistemas neuro-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
3. Aplicar sistemas fuzzy tipo-2 para modelar incerteza nas próprias funções de pertinência
4. Validar em domínios textuais específicos e com múltiplos modelos de LLM

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× 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.

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

Em síntese, a lógica fuzzy representa um caminho promissor para a detecção interpretável de textos gerados por LLMs, oferecendo transparência e explicabilidade em contextos onde essas características são valorizadas.

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.