

Relato das Etapas e Estratégia Adotada

Objetivo: Construir um pipeline completo para análise de sentenças judiciais da classe processual Usucapião, com foco em duas dimensões: classificação do resultado da sentença (taxonomia ternária: procedente, improcedente, neutro) e auditoria de fairness considerando o atributo sensível gênero do julgador (Feminino/Masculino), para avaliar possíveis disparidades e mitigar viés.

1. Preparação e Rotulagem Heurística

- Base original continha metadados das sentenças e a coluna decisão com o texto do dispositivo.
- Aplicamos regras heurísticas para mapear expressões típicas do dispositivo em três classes: procedente, improcedente, neutro.
- Resultado: geração de dois arquivos: resultado_tjma_rotulado.csv e dataset_minimal_tjma.csv.

2. Treinamento do Modelo Base

- Modelo escolhido: BERTimbau (neuralmind/bert-base-portuguese-cased), otimizado para português.
- Pipeline: tokenização, fine-tuning para classificação ternária, métrica principal macro-F1.
- Avaliação: acurácia geral, relatório por classe, matriz de confusão.

3. Auditoria de Fairness

- Atributo sensível: magistrado_genero (Feminino/Masculino).
- Métricas aplicadas: Demographic Parity, Equal Opportunity, Equalized Odds.
- Ferramentas: cálculo manual e integração opcional com Fairlearn.
- Visualizações: TPR/FPR por gênero e gráfico de disparidades Base vs. Adversarial.

4. Mitigação via Adversarial Debiasing

- Protótipo adversarial com encoder BERTimbau e duas cabeças (classificação e adversária).
- Loss total: $L_{cls} - \lambda * L_{adv}$ ($\lambda = 0.3$), usando gradient reversal.
- Objetivo: reduzir sinal do atributo sensível sem comprometer acurácia.
- Comparação: métricas de disparidade antes e depois da mitigação.

5. Saídas e Relatórios

- Arquivos gerados: resultado_tjma_rotulado.csv, dataset_minimal_tjma.csv, modelo treinado, fairness_report.json.
- Gráficos: matriz de confusão, TPR/FPR por gênero, disparidades Base vs. Adversarial.
- Observações: fairness é sociotécnico; diferenças podem refletir mix de casos por julgador.

Resumo da Abordagem

- Estratégia combinou NLP jurídico (BERTimbau) para classificação com auditoria de fairness baseada em métricas reconhecidas.
- Mitigação explorada via adversarial debiasing, técnica moderna para reduzir sinal do atributo sensível.
- Processo automatizado em notebooks, garantindo reproduzibilidade e transparência.