Identificação Automática da Fala do Magistrado em Audiências Trabalhistas Utilizando Machine Learning

Antonio J. S. Araújo[1], **Adam Santos**[1]

[1]Universidade Federal do Pará (UFPA)
Programa de Pós-graduação em Computação Aplicada (PPCA)
Núcleo de Desenvolvimento Amazônico em Engenharia (NDAE) - Tucuruí/PA - Brazil

j.garibald@gmail.com, adamdreyton@unifesspa.edu.br

Resumo. *A adoção da inteligência artificial no Judiciário brasileiro tem crescido para aumentar a eficiência dos processos se tornando uma prática comum dentre os tribunais brasileiros. Nesse cenário, o Tribunal Regional do Trabalho da 8ª Região desenvolveu uma ferramenta de transcrição automática de audiências, e este trabalho propõe um modelo complementar capaz de identificar quando o(a) magistrado(a) está falando. Devido às limitações de infraestrutura, foram utilizados algoritmos tradicionais de aprendizado de máquina, com destaque para o Perceptron Multicamadas (MLP), que atingiu 99% de acurácia. O estudo contribui para o avanço de sistemas judiciais inteligentes, aprimorando a precisão das transcrições e promovendo maior automação nos processos.*

1. Introdução

Um dos grandes problemas da justiça brasileira é a morosidade, despertando na sociedade a necessidade de se cobrar por uma justiça célere, capaz de entregar serviços efetivos a custos satisfatórios [Staats et al. 2016]. Segundo o relatório anual da Justiça em Números 2024[1], ano base 2023, disponibilizado pelo Conselho Nacional de Justiça (CNJ), foram registrados 35,3 milhões de novos processos. Ao final de 2023, havia 83,8 milhões de processos em curso; desse total, 18,5 milhões estavam suspensos.

Segundo o mesmo relatório, na Justiça do Trabalho, ramo com maior aumento percentual comparado ao ano de 2022, foram registrados aproximadamente 4,2 milhões de novos processos, o que representa um crescimento aproximado de 28,8% sobre 2022. Além disso, vale ressaltar que o tempo médio de tramitação dos processos pendentes está acima de 3 anos, e de processos baixados, resolvidos, está próximo a 2 anos.

A crescente demanda de ações trabalhistas exige a otimização na tramitação processual. De modo a mitigar essa problemática, grande parte dos tribunais brasileiros já está fazendo uso de soluções internas baseadas em inteligência artificial (IA), como é o caso do robô VICTOR, do Supremo Tribunal Federal (STF), que auxilia na resolução de recursos extraordinários do órgão, reduzindo consideravelmente o tempo de execução dessas tarefas [Inazawa et al. 2019].

No Tribunal Regional do Trabalho da 8ª Região (TRT8), foi desenvolvida uma ferramenta de degravação de audiência, Degrava[2], com o objetivo de acelerar e facilitar

[1]https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

[2]https://intranet.trt8.jus.br/noticia/trt-8-implementa-sistema-degrava-para-otimizar-analise-processual

a análise processual do Segundo Grau. Com a implantação da ferramenta, não há mais a necessidade dos analistas assistirem a toda a gravação da audiência para elaboração dos votos ou da minuta, tornando a análise mais eficiente, já que é possível realizar buscas por palavras-chave e navegar para os pontos mais relevantes da audiência. Dentre outras funcionalidades, a ferramenta realiza a separação de locutores, entretanto, não realiza a identificação dos mesmos.

Partindo deste contexto, neste trabalho foi desenvolvido um complemento a esta ferramenta: um modelo capaz de identificar o momento de fala do magistrado. Vale ressaltar que para o desenvolvimento do modelo levou-se em consideração a infraestrutura atual do TRT8, que atualmente não dispõe de servidores com GPUs para implantação de modelos mais robustos como os da arquitetura *transformers*.

2. Metodologia

Esta seção descreve a coleta e preparação de dados, bem como o treinamento do modelo conforme demonstrado na Figura 1.

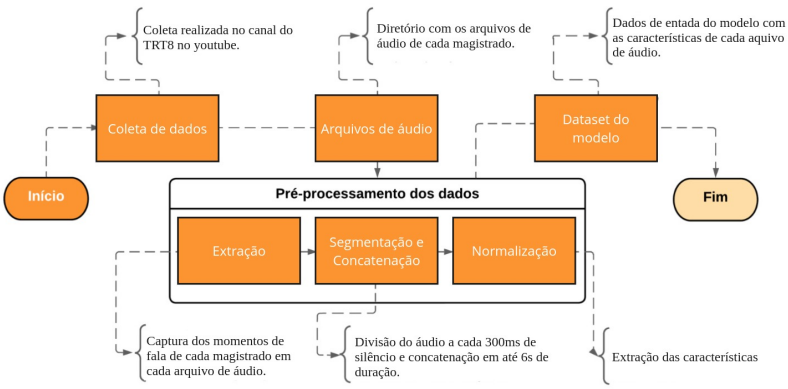


Figura 1. Processo de preparação do dataset.

2.1. *Dataset*

Os dados para o treinamento do modelo foram obtidos no canal do TRT8 no YouTube, na playlist “Sessões da 1ª Turma”, composta por vídeos de domínio público conforme a Lei 12.527/2011. Foram selecionados vídeos de diferentes ambientes para capturar áudios de cinco magistrados, garantindo diversidade de qualidade sonora e aprimorando a acurácia do modelo.

Em seguida, implementou-se um pipeline em Python com a biblioteca pydub para extrair e segmentar as falas individuais, gerando arquivos CSV com metadados de início, fim e identificação dos áudios. A segmentação considerou pausas de silêncio iguais ou superiores a 300 ms e os trechos foram ajustados para durar entre 4 e 6 segundos, conforme estudo de Chu et al. [Chu et al. 2009], resultando em um conjunto balanceado de amostras para cada classe de magistrado.

Por fim na etapa de ”Normalização”, apresentada na Figura 1, foi realizado o processamento de cada arquivo de áudio extraindo as características MFCCs, assim como [de Oliveira Guedes 2019], com número de coeficiêntes igual a 40. Para extração das características foi utilizada a biblioteca *librosa* (versão 0.12.0), usando o interpolador *kai-*

ser_fast. Essas características serão utilizadas como atributos ("*features*") para o modelo de classificação.

2.2. Modelo de classificação

Com a base de dados final definida, foi desenvolvido e validado um modelo supervisionado de classificação de áudio capaz de identificar qual magistrado está falando. O treinamento utilizou algoritmos tradicionais de aprendizado de máquina — SVC, Logistic Regression, Random Forest, AdaBoost e MLP — com divisão dos dados em 80% para treino e 20% para teste conforme trabalho de [Bauder and Khoshgoftaar 2017].

De acordo com a Tabela 1, todos os modelos apresentaram alto desempenho (acima de 94%), com destaque para o MLP e o SVC, que atingiram cerca de 99% em precisão, revocação e f1-score. O AdaBoost teve desempenho inferior, principalmente em revocação.

Tabela 1. Métricas dos modelos

Métrica	SVC	MLP	LR	RF	AdaBoost
Precision (macro avg)	**0.99**	0.98	0.97	0.98	0.91
Recall (macro avg)	0.98	**0.99**	0.97	0.97	0.83
F1-score (macro avg)	0.98	**0.99**	0.97	0.97	0.86
Acurácia	0.98	**0.99**	0.97	0.97	0.86

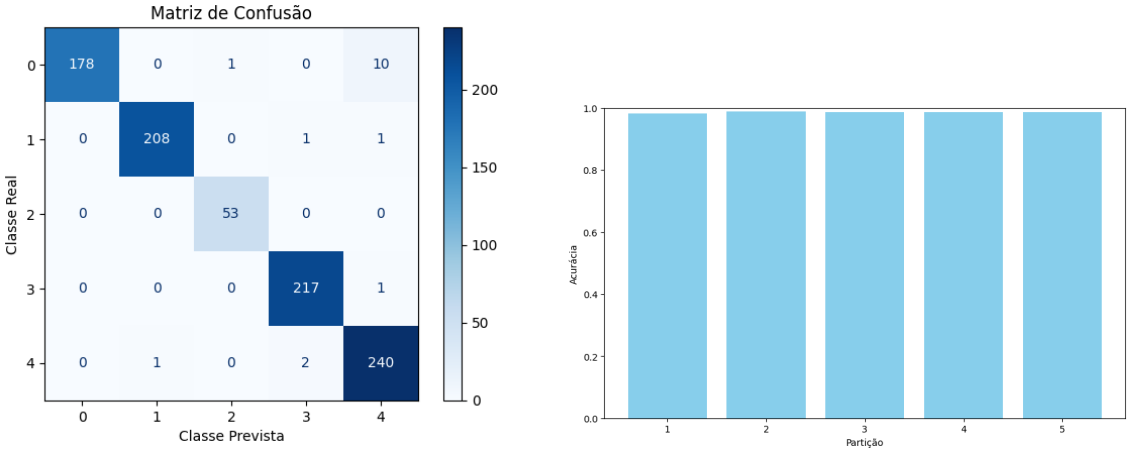
3. Resultados

A análise dos resultados demonstrou que o modelo MLP obteve desempenho superior na classificação dos áudios, apresentando alta precisão e poucas ocorrências de erro entre as classes. A Figura 2 demonstra que o modelo alcançou aproximadamente 100% de acurácia em torno de 20 épocas de treinamento e manteve baixa perda, evidenciando aprendizado estável, eficiente e sem sinais de overfitting. A validação cruzada confirmou sua robustez, com acurácia consistente em todos os cinco folds e variação mínima entre as execuções. A média de acurácia manteve-se entre 0.985 e 0.988, com baixa dispersão e apenas um pequeno outlier, comprovando a confiabilidade e a capacidade de generalização do modelo em diferentes divisões dos dados.

Na comparação entre os demais algoritmos, o Random Forest e o SVC também apresentaram resultados satisfatórios, com alta precisão e poucas confusões, especialmente nas classes mais frequentes. Já o AdaBoost e a Logistic Regression demonstraram maior dificuldade em distinguir determinadas classes, com desempenho inferior nas Classes 0 e 4. Esses resultados reforçam a superioridade das abordagens não lineares, como o MLP e o SVC, em tarefas de classificação de áudio mais complexas.

4. Conclusão

Este trabalho apresentou o desenvolvimento e a validação de um modelo de classificação de áudio voltado à identificação automática dos momentos de fala do magistrado em audiências trabalhistas, como complemento à ferramenta de degravação do TRT8. A metodologia envolveu a coleta de vídeos públicos, a segmentação dos áudios, a extração



(a) Matriz de confusão **(b) Gráfico de acurácia por *fold***

Figura 2. Matriz de confusão e acurácia por *fold* do modelo MLP

de características MFCCs e a comparação entre diferentes algoritmos de aprendizado de máquina tradicionais.

Os resultados mostraram alto desempenho em todos os modelos testados, com destaque para o MLP, que obteve 99% de acurácia e excelente equilíbrio entre precision, recall e f1-score, inclusive nas classes menos representadas. A validação cruzada confirmou a consistência e a estabilidade do modelo, reforçando sua capacidade de generalização.

Além de robusto e eficiente, o modelo proposto é compatível com a infraestrutura atual do TRT8, dispensando hardware especializado e podendo ser facilmente integrado à ferramenta já existente. Como trabalhos futuros, sugere-se investigar arquiteturas mais avançadas, como redes neurais profundas e modelos transformers, bem como expandir o conjunto de dados para abranger diferentes turmas e contextos, ampliando o alcance e a precisão da solução.

Referências

Bauder, R. A. and Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods. In *Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 858–865. IEEE.

Chu, S., Narayanan, S., and Kuo, C.-C. J. (2009). Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158.

de Oliveira Guedes, V. (2019). Deep learning aplicado a classificação de patologias da voz. Master's thesis, Instituto Politecnico de Braganca (Portugal).

Inazawa, P., Hartmann, F., de Campos, T., Silva, N., and Braz, F. (2019). Projeto victor. *Computação Brasil*, (39):19–24.

Staats, J. L., Bowler, S., Hiskey, J. T., Staats, J. L., and Thiskey, J. (2016). Measuring judicial performance in latin america. *Latin American Research Review*, 47(4):77–106. Stable URL.