Avaliação de Ferramentas de Extração de Texto em Documentos Jurídicos: Uma Análise de Soluções OCR e Baseadas em Texto.

Aline M. B. Silva[1], **Ewaldo E. C. Santana**[1],
Fábio M. F. Lobato[1,2] **Antonio F. L. Jacob Junior**[1]

[1] Centro de Ciências e Tecnologias, Departamento de Engenharia de Computação
Universidade Estadual do Maranhão (UEMA)

[2]Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará (UFOPA)

alinesilva17@aluno.uema.br, ewaldosantana@professor.uema.br

fabio.lobato@ufopa.edu.br, antoniojunior@professor.uema.br

Resumo. *Este estudo realiza uma análise comparativa de oito ferramentas de extração de texto aplicadas a documentos jurídicos em formato PDF, divididas entre técnicas de OCR e extração baseada em texto, utilizando a tarefa de classificação de texto jurídico como critério de avaliação da eficácia. As ferramentas avaliadas incluem PDFMiner, PDFX, PyPDF e PyMuPDF para documentos nato-digitais, e Pytesseract, docTR, EasyOCR e PaddleOCR para documentos digitalizados. Os testes foram conduzidos com documentos do sistema do Conselho Nacional de Justiça (CNJ), e os resultados evidenciaram que ferramentas de extração baseada em texto, como PDFX, demonstraram melhor desempenho em documentos nato-digitais, enquanto docTR se destacou entre as soluções de OCR. O estudo oferece insights valiosos para a escolha de ferramentas mais adequadas em cenários jurídicos, considerando a natureza dos documentos a serem processados.*

1. Introdução

A extração de informações em grandes volumes de dados (Big Data) tornou-se essencial na sociedade contemporânea, facilitando a geração de conhecimento e a tomada de decisões mais embasadas. No entanto, esse processo enfrenta desafios, especialmente na coleta, estruturação e análise de dados não estruturados, como ocorre no setor jurídico [González 2023]. No Brasil, o crescente volume de processos judiciais torna indispensável o uso de soluções tecnológicas baseadas em Big Data para otimizar a gestão dessas informações[CNJ 2024].

A extração de texto de documentos PDF, particularmente em processos judiciais, depende da natureza do arquivo. Existem dois tipos principais de PDFs: nato-digitais, onde o texto pode ser selecionado diretamente, e digitalizados, nos quais o conteúdo textual está embutido em imagens. O Decreto 8.539 de 2015 [Brasil 2015] formaliza a distinção entre esses documentos, destacando a complexidade de manipulação dos documentos escaneados.

Diante desse cenário, duas abordagens são comumente utilizadas para a extração de texto: a extração baseada em texto (EBT) e o Reconhecimento Óptico de Caracteres

(OCR). A EBT é ideal para documentos nato-digitais, enquanto o OCR é necessário para os digitalizados, convertendo imagens em texto editável. Cada técnica apresenta diferentes níveis de confiabilidade e eficiência, o que torna a escolha da abordagem crucial, especialmente no contexto jurídico [Neudecker et al. 2021].

Este trabalho visa realizar um estudo comparativo entre ferramentas de extração de texto, explorando tanto técnicas de EBT quanto OCR em documentos jurídicos. A pesquisa avalia oito ferramentas e suas respectivas capacidades, com o objetivo de identificar as mais eficientes em termos de tempo de processamento e precisão usando a tarefa de classificação de texto. Além disso, um protótipo foi desenvolvido em colaboração com o Tribunal de Justiça do Maranhão (TJMA), integrando as ferramentas mais eficazes em uma arquitetura de microserviços para otimizar o processamento de documentos no âmbito judicial.

2. Metodologia

Este estudo utilizou uma abordagem experimental baseada no trabalho de [Sancar et al. 2023], composta por cinco etapas principais, para a avaliação de ferramentas de extração de texto aplicadas a documentos jurídicos. Inicialmente, foi realizada uma revisão de literatura com o propósito de identificar as ferramentas mais robustas e amplamente utilizadas na extração de texto de arquivos PDF. Foram selecionadas oito ferramentas, divididas em duas categorias: extração baseada em texto (PDFMiner, PDFX, PyPDF, PyMuPDF) e extração baseada em OCR (Pytesseract, docTR, EasyOCR, PaddleOCR).Na sequência, foi definido o conjunto de dados, composto por 7.440 documentos jurídicos em formato PDF provenientes do sistema do Conselho Nacional de Justiça (CNJ). A amostra incluía tanto documentos nato-digitais, quanto documentos digitalizados. Foram estabelecidos procedimentos rigorosos de coleta de dados para assegurar a integridade e a representatividade dos documentos selecionados, garantindo que fossem contempladas as diferentes naturezas dos arquivos no processo de análise.

Posteriormente, as ferramentas foram aplicadas a cada um dos documentos, seguindo um protocolo padronizado de pré-processamento de texto. Este processo incluiu técnicas de Processamento de Linguagem Natural (PLN), como stemização e remoção de stop words, com o objetivo de uniformizar o texto extraído e prepará-lo para as etapas subsequentes de vetorização e classificação.Além disso, o tempo de processamento necessário para que cada ferramenta extraísse o texto dos documentos foi registrado e considerado na avaliação de desempenho das ferramentas, uma vez que a eficiência computacional é um fator crucial na escolha de soluções de extração de texto em grandes volumes de documentos jurídicos.Para avaliar a precisão das extrações, os conjunto de textos extraídos pelas ferramentas foram convertidos em vetores numéricos utilizando um vetorizador TF-IDF (Term Frequency-Inverse Document Frequency), que foram usados em uma tarefa de classificação de texto jurídico – especificamente, na identificação de petição inicial.

Para isso, foram realizados 100 experimentos de Monte Carlo utilizando cinco classificadores de aprendizado de máquina: Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Árvores de Decisão (CART), Random Forest Classifier (RFC), e Regressão Logística (LR).Os resultados obtidos foram organizados em tabelas de dados e submetidos a uma análise estatística rigorosa. Para cada métrica de desempenho

(acurácia, precisão, recall e F1-score), foram calculados parâmetros descritivos, como média, mediana, desvio padrão, entre outros. A fim de comparar as médias amostrais de cada métrica para os diferentes classificadores, foram aplicados dois testes preliminares: o teste de Shapiro-Wilk, para avaliar a normalidade das distribuições, e o teste de Levene, para verificar a homogeneidade das variâncias (homoscedasticidade). Como as hipóteses nulas foram rejeitadas para as amostras analisadas, optou-se pelo uso do teste não-paramétrico de Kruskal-Wallis, com o objetivo de verificar a igualdade das medianas entre os grupos. Por fim, foi conduzida uma análise post-hoc utilizando o teste de Dunn com correção de Bonferroni, a fim de identificar quais classificadores apresentaram diferenças estatisticamente significativas nas médias das métricas. Essa análise permitiu determinar, com maior precisão, quais ferramentas e classificadores se destacaram em termos de desempenho na tarefa de extração e classificação de textos em documentos jurídicos.

3. Resultados e Discussão

Com base nos resultados obtidos na tabela 1 e na tabela 2 a partir dos testes de Kruskal-Wallis, seguidos por comparações post-hoc (Teste de Dunn), confirmaram que o modelo docTR-SVM, para o reconhecimento óptico de caracteres (OCR) e o modelo PDFX-RFC para a extração de informações em arquivos PDF demonstraram-se como as abordagens mais eficazes e adequadas para o presente contexto.

Tabela 1. Os melhores resultados na tarefa de classificação para os OCR

Métrica	Ferramenta	Classificador	Valor
ACURÁCIA	docTR	SVM	0,9916
F1-BINARY	docTR	SVM	0,9677
F1-MACRO	docTR	SVM	0,9814
F1-MICRO	docTR	SVM	0,9916
PRECISÃO	PaddleOCR	RFC	0,9525
RECALL	docTR	SVM	0,9880

No que se refere ao tempo de processamento em documentos nato-digitais, as ferramentas PDFMiner, PDFX, PyPDF e PyMuPDF foram testadas em documentos nato-digitais. Dentre essas, o PyMuPDF destacou-se por apresentar o menor tempo de processamento, completando a extração de todo o conjunto de dados em 69 segundos. Em contraste, PDFMiner, PDFX e PyPDF demonstraram tempos significativamente maiores, variando de 697 a 1671 segundos.

Tabela 2. Os melhores resultados na tarefa de classificação para os EBT

Métrica	Ferramenta	Classificador	Valor
ACURÁCIA	PDFX	RFC	0,9862
F1-BINARY	PDFX	RFC	0,9697
F1-MACRO	PDFX	RFC	0,9803
F1-MICRO	PDFX	RFC	0,9862
PRECISÃO	PDFX	RFC	0,9517
RECALL	PDFX	KNN	0,9936

Por outro lado , entre as soluções de OCR, as ferramentas Pytesseract, docTR, EasyOCR e PaddleOCR foram testadas. PaddleOCR apresentou o melhor desempenho, com um tempo de processamento de apenas 10.063 segundos em todo conjunto de dados , demonstrando ser altamente eficiente para a extração de texto de imagens. docTR, embora mais lento, com 108.925 segundos, mostrou-se confiável e pode ser uma alternativa viável quando a velocidade não é o fator mais crítico. As ferramentas Pytesseract e EasyOCR com tempos de 45.944 e 71.043 segundos, respectivamente, revelaram ser opções menos eficientes em termos de tempo e eficiência.

4. Considerações Finais

Este estudo comparativo demonstrou a eficácia de diferentes ferramentas de extração de texto em documentos jurídicos, destacando as particularidades de abordagens baseadas em texto e OCR em uma tarefa de classificação de texto. A partir dos resultados, foi possível desenvolver um protótipo eficiente, alinhado com a arquitetura adotada pelo TJMA, que otimiza o processamento de documentos no contexto judicial. A integração das ferramentas mais eficazes em uma arquitetura de microserviços não apenas valida sua aplicabilidade, mas também aponta para a viabilidade de implementação em larga escala. Assim, este trabalho contribui significativamente para a modernização dos processos judiciais, promovendo maior eficiência na gestão de dados jurídicos.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq Nº 045/2021; e pelo Acordo de Cooperação Tecnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA).

Referências

Brasil (2015). Dispõe sobre o uso do meio eletrônico para a realização do processo administrativo no âmbito dos órgãos e das entidades da administração pública federal direta, autárquica e fundacional. diário oficial [da] república federativa do brasil. *BRASIL. Decreto nº*, 8539.

CNJ (2024). Justiça em números 2023. relatório CNJ, 2023.disponível em https://justica-em-numeros.cnj.jus.br/. acesso 22 de ago de 2024. In *Conselho Nacional de Justiça (CNJ)*.

González, J. A. G. (2023). La inteligencia artificial en el campo jurídico. *Revista Académica CUNZAC*, 6(2):96–103.

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., and Pletschacher, S. (2021). A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, pages 13–18.

Sancar, Y., Karabey Aksakallı, I., and Karacalı, T. (2023). Text classification-based petition recognition and routing system: a turkish case study. *International Journal of Information Technology*, 15(4):2139–2146.