ARTIGO

# A ERA DOS MODELOS DE LINGUAGEM ESPECIALIZADOS NA JUSTIÇA 4.0

**POR**

Fabrício Almeida do Carmo, Ewaldo Eder Carvalho Santana, Omar Andres Carmona Cortes e José Jorge Figueiredo dos Anjos Junior

fabricio.almeida@discente.ufma.br, ewaldoeder@gmail.com, omar@ifma.edu.br e jjjunior@tjma.jus.br

O Brasil possui um dos sistemas judiciários mais volumosos do mundo. Segundo o relatório "Justiça em Números 2025", o Poder Judiciário finalizou o ano de 2024 com um acervo de 80,6 milhões de processos em tramitação[1]. Diante desse cenário, onde a taxa de congestionamento se mantém elevada, a busca por celeridade processual deixa de ser apenas uma meta administrativa para se tornar um imperativo de cidadania. Nesse contexto, o Conselho Nacional de Justiça (CNJ) impulsionou o programa "Justiça 4.0", fomentando a transformação digital dos tribunais e a adoção de ferramentas e técnicas de Inteligência Artificial (IA). A plataforma Codex, é um exemplo notório desse programa, focando no processamento e estruturação de bases de dados processuais, visando alimentar as chamadas aplicações LegalTech [2]..

No entanto, a digitalização dos processos é apenas o primeiro passo. O verdadeiro salto de produtividade reside na capacidade das máquinas de "lerem" e "compreenderem" a grande variedade de textos jurídicos produzidos diariamente. A IA, especificamente através do PLN, surge como a ferramenta-chave para essa tarefa. Contudo, o domínio jurídico apresenta um obstáculo peculiar: a lin-

guagem. O "juridiquês", com seus jargões, arcaísmos e estruturas sintáticas complexas, desafia modelos de IA treinados em textos genéricos da Internet.

## A Evolução da Tecnologia: De Vetores a Modelos Generativos

Para que um computador processe texto, ele precisa convertê-lo em representações matemáticas. Inicialmente, técnicas de *word embeddings* (como Word2Vec) cumpriam esse papel, transformando palavras em vetores onde termos semanticamente similares ficavam próximos no espaço matemático. Um modelo genérico pode entender que "banco" é uma instituição financeira ou um assento de praça. Porém, no contexto jurídico, a distinção precisa entre termos como "deferimento", "tutela" ou "agravo" exige treinamento contextualizado.

A evolução para modelos baseados na arquitetura *Transformer*, como o BERT, trouxe a capacidade de entender o contexto da sentença inteira, e não apenas de palavras isoladas. Mais recentemente, a área avançou para os LLMs generativos. No entanto, pesquisas indicam que há uma lacuna de recursos para línguas com menos representatividade digital que o inglês, criando um desequilíbrio no desenvolvimento de modelos de IA. Iniciativas recentes, como o modelo Tucano [3], treinado em um corpus massivo de português (GigaVerbo), o LegalBert-pt[4], e Jurema [5], entre outros, buscam mitigar essa escassez para a língua geral, mas o setor jurídico exige um refinamento ainda maior.

## A Necessidade de Especialização no Domínio Jurídico

A base de uma IA jurídica eficiente é o dado. Modelos generalistas, treinados na Wikipédia ou em notícias, frequentemente falham em captar as nuances do Direito. Em esforços recentes de pesquisa no Brasil, já se nota a existência de compilados de dados (*corpora*) contendo diferentes documentos jurídicos, incluindo acórdãos, petições iniciais e movimentações processuais, provenientes de tribunais de diversas esferas, como o STF, o TST e os tribunais estaduais. A estruturação desses dados permite o treinamento e adaptação de modelos especializados.

Exemplos práticos, como o LegalBert--pt e o BumbaBert, ilustram essa tendência. O BumbaBERT, por exemplo, é uma adaptação da arquitetura BERT, treinada especificamente com dados do Judiciário brasileiro. Experimentos comparativos demonstram que esses modelos especializados superam modelos genéricos (como o BERTimbau) em tarefas críticas como identificação de Precedentes Judiciais em Petições iniciais [2].

Uma aplicação comum é a classificação automática de peças processuais. Ao tentar identificar se um processo se enquadra em um Incidente de Resolução de Demandas Repetitivas (IRDR), mecanismo vital para garantir isonomia em casos idênticos, modelos treinados com textos jurídicos apresentam métricas de acurácia e F1-score superiores. Isso ocorre porque o modelo especializado "entende" que, no contexto de uma petição, certas palavras carregam um peso decisório que

um modelo comum ignoraria. Além disso, técnicas avançadas de ajuste fino, como o uso de *Low-Rank Adaptation* (LoRA) [6], têm permitido adaptar esses modelos a tarefas específicas com menor custo computacional, viabilizando a implementação e implantação de aplicações nos tribunais com recursos de hardware limitados.

**Aplicações Práticas e a Visão de Futuro: Justiça 5.0**

A aplicação prática desses modelos já é uma realidade no Tribunal de Justiça do Maranhão (TJMA), onde a parceria entre o laboratório de inovação (ToadaLab) e a academia evoluiu de modelos classificadores para soluções de IA Generativa integradas ao fluxo de trabalho.

Um exemplo concreto é o Projeto ANA (Automação de Notas e Análises). Diferente de ferramentas genéricas, o ANA utiliza uma arquitetura que integra três elementos: um modelo de linguagem (LLM), um repositório curado de precedentes e um conector de processos (via protocolo *Model Context Protocol* - MCP) entre documentos do sistema PJe e o fluxo de trabalho. O sistema consulta a jurisprudência do tribunal, cruza com os dados do processo eletrônico e gera minutas padronizadas.

Essa evolução tecnológica não apenas consolida o Programa Justiça 4.0, mas prepara o terreno para o que se vislumbra como Justiça 5.0, um estágio onde a transformação digital evolui para um ambiente de inteligência colaborativa. Neste cenário, a IA atua como apoio à decisão humana, permitindo que magistrados e servidores foquem na atividade intelectual e na humanização do atendimento, enquanto a máquina lida com a complexidade dos dados.

Como destaca o Desembargador Nilo Lacerda [7], é essencial que o uso dessas tecnologias venha acompanhado de letramento digital e validação humana rigorosa para mitigar vieses. Modelos de linguagem especializados, sejam discriminativos (como o BumbaBert) ou generativos (como o ANA), não são apenas uma inovação acadêmica, são ferramentas essenciais para uma justiça mais rápida, acessível e eficiente no Brasil.

**Referências**

1. CNJ. Justiça em Números 2025. Brasília: Conselho Nacional de Justiça, 2025. Disponível em: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/.
2. CARMO, F. A. do. Representações Embeddings Orientadas à Linguagem Jurídica Brasileira. Dissertação (Mestrado em Engenharia da Computação e Sistemas) – Universidade Estadual do Maranhão, São Luís, 2024.
3. CORRÊA, N. K. et al. Tucano: Advancing neural text generation for Portuguese. Patterns, v. 6, 101325, 2025.
4. SILVEIRA, R. et al. LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In: Intelligent Systems, Springer, 2023.
5. JUREMA-BR. Jurema-7B.[S.l.]: HuggingFace, 2024. Disponível em: https://huggingface.co/Jurema-br/Jurema-7B. Acesso em: 20 jan. 2026.
6. HU, Edward J. et al. Lora: Low-rank adaptation of large language models. ICLR, v. 1, n. 2, p. 3, 2022.
7. LACERDA, N. L. Elementos de Redação da Decisão Judicial em Segundo Grau na Era da Inteligência Artificial Generativa. Curitiba e Editora CRV, p.16, 2025.

**FABRÍCIO ALMEIDA DO CARMO** é Mestre em Engenharia da Computação e Sistemas pela Universidade Estadual do Maranhão (UEMA), doutorando em Engenharia Elétrica pela Universidade Federal do Maranhão (UFMA). Pesquisador com ênfase em Processamento de Linguagem Natural aplicado ao domínio jurídico. Atua no desenvolvimento de soluções de Inteligência Artificial em parceria com o Tribunal de Justiça do Maranhão (TJMA), focando na modernização e celeridade da prestação jurisdicional.

**EWALDO EDER CARVALHO SANTANA** é Doutor em Engenharia Elétrica pela Universidade Federal de Campina Grande (2009) e Mestre pela Universidade Federal do Maranhão (2006). Realizou estágio pós-doutoral no GIPSA-Lab (França). Atualmente, é professor na Universidade Estadual do Maranhão (UEMA), onde coordena o Laboratório para Aquisição e Processamento de Sinais (LAPS). Suas áreas de pesquisa incluem processamento de sinais, sensores e aprendizado de máquinas.

**OMAR ANDRES CARMONA CORTES** é Professor Associado do Instituto Federal do Maranhão (IFMA), com mestrado (1999) e doutorado (2004) em Ciências da Computação e Matemática Computacional pela USP e pós-doutorado na Dalhousie University (Canadá). Sua pesquisa tem se focado em Inteligência Computacional, Aprendizado de Máquina e Computação Paralela, com ênfase em algoritmos evolutivos, lógica nebulosa e Deep Learning.

**JOSÉ JORGE FIGUEIREDO DOS ANJOS JUNIOR** é Juiz Auxiliar da Presidência do Tribunal de Justiça do Maranhão (TJMA). Especialista em Direito Notarial e Registral (Fortium) e Criminologia e Processo Penal (PUCRS). Coordena o ToadaLab, laboratório de inovação do Poder Judiciário do TJMA, onde desenvolve projetos de aplicação de Inteligência Artificial ao fluxo de trabalho judicial. É autor do Projeto ANA (Automação de Notas e Análises), que integra tecnologias de IA generativa, RAG (Retrieval-Augmented Generation) e MCP (Model Context Protocol) para automação de minutas judiciais. Tem expertise em automação judicial, integração de sistemas jurídicos e capacitação de magistrados e servidores para o uso ético e eficiente de IA no Judiciário.