Can data fusion help BERT recognize predatory litigation patterns?

Jonathan O. Fernandez[1], **Omar A. C. Cortes**[1,2], **Antonio F. L Jacob**[1], **Fábio M. F. Lobato**[1,3]

[1]Universidade Estadual do Maranhão (UEMA)
São Luís – MA – Brasil

[2]Instituto Federal do Maranhão (IFMA)
São Luís – MA – Brasil

[3]Universidade de São Paulo (USP)
São Carlos – SP – Brasil

omar@ifma.edu.br, fabio.lobato@icmc.usp.br

Abstract. *Predatory litigation involves the repetitive filing of lawsuits and the large-scale use of similar petitions, placing an excessive burden on the judiciary. In this context, this study proposes a data fusion-based approach to support the preliminary identification of this phenomenon in initial petitions. Relational, temporal, and semantic indicators are obtained through feature engineering and combined into an adaptive score that guides a weak labeling process. These indicators are subsequently verbalized into standardized textual narratives, which are used to fine-tune specialized BERT models (BumbaBERT and LegalBERT-pt) on a dataset of 8,729 cases. The fine-tuned LegalBERT-pt achieved a macro F1-score of 0.96 on the proposed weak labeling target, showing that the models effectively recover the constructed indicators from the verbalized narratives and support courts in focusing analytical efforts on potentially predatory cases while ensuring the prompt resolution of legitimate claims.*

Resumo. *A litigância predatória consiste no ajuizamento reiterado de ações e na utilização, em larga escala, de petições semelhantes, o que provoca sobrecarga ao Poder Judiciário. Nesse contexto, este trabalho propõe uma abordagem baseada em fusão de dados para apoiar a identificação preliminar desse fenômeno em petições iniciais, na qual indícios relacionais, temporais e semânticos são obtidos por meio de engenharia de atributos, combinados em um score adaptativo que orienta um processo de rotulagem fraca e, posteriormente, verbalizados em narrativas textuais padronizadas, utilizadas no fine-tuning de modelos BERT especializados (BumbaBERT e LegalBERT-pt) em um dataset de 8.729 processos. O LegalBERT-pt ajustado alcançou F1-score macro de 0,96 sobre o alvo de rotulagem fraca proposto, evidenciando que os indícios construídos são efetivamente recuperados pelos modelos a partir das narrativas verbalizadas, fornecendo subsídios para que os tribunais concentrem o esforço analítico nos casos potencialmente predatórios e resguardem a celeridade no atendimento às demandas legítimas.*

1. Introdução

A transformação digital do Poder Judiciário brasileiro, impulsionada pela implementação do Processo Judicial Eletrônico (PJe) e por iniciativas de governança tecnológica, como o Programa Justiça 4.0, ampliou o acesso e a velocidade de tramitação processual, ao mesmo tempo em que exigiu novos instrumentos de gestão e de governança do PJe [Rampim and Igreja 2022, de Araújo et al. 2022]. O Relatório "Justiça em Números" registra a dimensão desse acervo processual, com o judiciário encerrando 2025 com 75 milhões de processos pendentes, dos quais 59 milhões estavam em tramitação[1].

Neste cenário, emerge o fenômeno da litigância predatória, caracterizada pelo ajuizamento massivo e automatizado de ações judiciais com causas de pedir genéricas, fatos artificialmente replicados e uso abusivo do direito de ação, visando obter vantagens indevidas por meio do congestionamento dos sistemas judiciais [Silva and Neris 2025, de Souza 2024]. O fenômeno manifesta-se por meio de diversos padrões comportamentais, como a distribuição simultânea de ações substancialmente idênticas, o uso de modelos de petição com alterações mínimas e a concentração temporal de protocolos que evidenciam automação [Paradela and Cavaco 2025]. O principal desafio para sua detecção reside na sofisticação das estratégias de camuflagem textual, nas quais variações superficiais são introduzidas com o objetivo de contornar filtros tradicionais baseados em palavras-chave ou em regras sintáticas simples [Neto et al. 2025].

Nesse sentido, os avanços recentes em Processamento de Linguagem Natural (PLN) e em aprendizado de máquina oferecem oportunidades para a análise automatizada de grandes volumes de documentos jurídicos [Melo 2023]. Modelos de linguagem baseados na arquitetura *Transformer*, como o *Bidirectional Encoder Representations from Transformers* (BERT), têm demonstrado desempenho consistente em tarefas de classificação textual, extração de informação e análise de similaridade semântica [Justino et al. 2025].

No contexto brasileiro, a adaptação desses modelos para o processamento de textos jurídicos tem sido objeto de crescente interesse, com desenvolvimento de versões especializadas treinadas em dados jurídicos, como Legal-BERT-pt [Silveira et al. 2023] e BumbaBERT [Carmo et al. 2023]. Diante desse panorama, este trabalho propõe uma abordagem para apoiar a identificação preliminar de litigância predatória em petições iniciais, combinando atributos relacionais, temporais e semânticos. Os indícios extraídos são combinados em um *score* adaptativo, utilizado no processo de rotulagem fraca, e convertidos em narrativas curtas padronizadas para serem usadas como entrada dos modelos. Para avaliar em que medida os modelos reproduzem os indícios verbalizados e como os padrões convergem com as marcações institucionais existentes, utilizaram-se os modelos BumbaBERT e LegalBERT-pt, considerando diferentes estratégias de treinamento e dois alvos distintos em cenários experimentais.

O restante do artigo encontra-se organizado como segue. Na Seção 2, discutem-se os trabalhos correlatos. Os materiais e métodos são detalhados na Seção 3. Os resultados são apresentados e discutidos na Seção 4. Por fim, na Seção 5, são apresentadas as considerações finais, as limitações do estudo e as sugestões para trabalhos futuros.

[1]Disponível em: https://justica-em-numeros.cnj.jus.br/painel-estatisticas/

2. Trabalhos relacionados

A classificação automática de textos tem sido amplamente investigada por meio de diferentes estratégias de representação textual e de modelos de aprendizado supervisionado. [Borges et al. 2025] apresentam uma análise comparativa entre modelos baseados em *Transformers* (BERT e BERTimbau) e algoritmos tradicionais (*Decision Tree*, XGBoost, *Naive Bayes*, *Support Vector Machine* e *Multilayer Perceptron*), em cenários binários e multiclasse, com representações baseadas em *embeddings*. Os modelos baseados em *Transformers* apresentaram melhor desempenho, sobretudo quando a compreensão contextual do texto é mais relevante, embora os métodos tradicionais permaneçam competitivos quando associados a representações adequadas. O estudo também aponta dificuldades em tarefas multiclasse, atribuídas ao desbalanceamento das classes e à sobreposição semântica entre categorias, aspectos relevantes para problemas em que os limites entre as classes não são rigidamente definidos.

De forma complementar, [Santos et al. 2025] comparam modelos clássicos de aprendizado de máquina, modelos *Transformers* ajustados por *fine-tuning* e *Large Language Models* (LLMs) baseados em *prompt engineering* em uma tarefa de classificação multiclasse de petições da Defensoria Pública do Estado de Goiás, distribuídas em 24 categorias. Os resultados demonstraram melhor desempenho dos modelos *Transformers* ajustados, com *F1-Score* de 94,70% para o BERTimbau Large e de 90,91% para o BERTimbau Base. Embora as abordagens com LLMs tenham apresentado resultados competitivos, o estudo reforça que o ajuste fino de modelos pré-treinados ainda constitui uma estratégia eficaz para a classificação jurídica em português. Esse resultado é relevante para o presente trabalho, por indicar a adequação de modelos baseados em BERT a tarefas multiclasse com textos jurídicos, um contexto próximo à identificação de indícios de litigância predatória em petições iniciais.

Outra linha relacionada envolve a adaptação de dados estruturados para formatos compatíveis com modelos de linguagem. [Gheibi and Hougen 2025] propõem uma abordagem para a classificação do risco de câncer cervical a partir da conversão de dados tabulares em descrições textuais, permitindo a aplicação desses modelos a uma tarefa originalmente estruturada. Para tal, os autores avaliaram modelos generativos, como GPT-2, GPT-3, GPT-4o, LLaMA, Mistral e Phi, em estratégias *zero-shot* e *few-shot*, além de modelos especializados no domínio biomédico, como BioBERT, PubMedBERT, ALBERT e DistilBERT, ajustados por meio de *fine-tuning*. Os resultados indicaram desempenho limitado dos LLMs de propósito geral para uma classificação confiável, enquanto modelos BERT especializados e ajustados apresentaram desempenho superior, com menor demanda computacional. Essa abordagem contribui para o presente estudo ao demonstrar que a transformação de atributos estruturados em descrições textuais pode viabilizar o uso desses modelos em tarefas classificatórias que combinam diferentes tipos de informação.

Em contraste com os trabalhos apresentados, este estudo se diferencia em dois aspectos. Enquanto [Borges et al. 2025] e [Santos et al. 2025] realizam a classificação utilizando rótulos já existentes, neste caso, não havia uma variável alvo de litigância predatória na base, exigindo que a mesma fosse criada a partir de uma rotulagem fraca orientada por indícios estruturados que são construídos por meio da comparação entre processos, como a similaridade de petições e repetição de atores, padrões que não são notáveis em uma petição isolada.

3. Materiais e Métodos

Este estudo foi guiado pela metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [Wirth and Hipp 2000], amplamente utilizada na mineração de textos, inclusive no domínio jurídico, como demonstrado em [Justino et al. 2025]. O CRISP-DM é composto por 6 etapas, a saber: Entendimento do negócio, Entendimento dos dados, Preparação dos dados, Modelagem, Avaliação e Implantação, detalhadas a seguir.

O **Entendimento do negócio** teve por finalidade compreender o domínio jurídico e o problema investigado, com base na literatura científica e em documentos institucionais e normativos do Poder Judiciário. Para tal, foi realizado um levantamento bibliográfico de caráter *Ad hoc* [Ralph and Baltes 2022], voltado à identificação dos principais conceitos, critérios normativos e indícios associados à litigância predatória. Nesse contexto, a Recomendação n^o 159/2024[2] do CNJ forneceu parâmetros institucionais para a identificação e o tratamento do problema, enquanto o Diagnóstico Nacional sobre o Enfrentamento da Litigância Abusiva e Predatória no Poder Judiciário[3] contribuiu para contextualizar o enfrentamento do fenômeno no âmbito dos tribunais brasileiros. A partir desse levantamento, o trabalho focou em indícios observáveis na base de dados disponível, uma vez que o fenômeno pode manifestar-se por meio de diversos padrões comportamentais [Paradela and Cavaco 2025, Magalhães et al. 2024]. O embasamento teórico resultante dessa etapa está disponível como material complementar em https://github.com/jonoliveiraf/litigancia_predatoria_eniac_2026.

No **Entendimento dos dados**, o conjunto de dados é composto por petições iniciais provenientes de processos judiciais, disponibilizadas pelo Tribunal de Justiça do Maranhão (TJMA) no formato *JavaScript Object Notation* (JSON), contendo 11.621 processos judiciais e abrangendo o período de 2024 a 2026. A Tabela 1 apresenta os campos do *dataset* e suas respectivas descrições.

Tabela 1. Sumarização dos dados disponibilizados.

Campo	Tipo	Descrição
id_processo	*Int*	Identificador do processo
id_orgao_julgador	*Int*	Identificador do órgão julgador do processo
ds_orgao_julgador	*String*	Nome do órgão julgador do processo
cd_classe_judicial	*String*	Código da classe judicial
ds_classe_judicial	*String*	Nome da classe judicial
str_partes_polos_ativos	*String*	Identificadores dos polos ativos (separados por ponto e vírgula)
str_partes_polos_passivos	*String*	Identificadores dos polos passivos (separados por ponto e vírgula)
nr_processo	*String*	Número do processo
dt_distribuicao	*Datetime*	Data e horário de distribuição
id_peticao	*Int*	Identificador do documento identificado como petição inicial do processo
texto_peticao	*String*	Texto do conteúdo extraído do documento da petição inicial
str_nm_etiqueta	*String*	Etiquetas aplicadas pelo TJMA
ids_advogados	*String*	Identificadores dos advogados do processo (separados por ponto e vírgula)
advogados	*String*	Lista de informação de cada advogado (id, nome, OAB)

Além dos campos estruturados descritos na Tabela 1, o campo *str_nm_etiqueta* contém marcações previamente existentes nos processos, agrupadas em três categorias: *Sem Etiqueta*, *Possível Prevenção* e *Possível Demanda Predatória*.

Na etapa de **Preparação dos dados**, inicialmente, foram removidos registros

[2]Disponível em: https://atos.cnj.jus.br/atos/detalhar/5822

[3]Disponível em: https://tinyurl.com/diagnostico-cnj-litigancia

inválidos, correspondentes a petições sem identificadores de processo e conteúdos textuais. Em seguida, foram tratados os casos duplicados, definidos pela repetição do campo *id_peticao*, preservando o registro mais antigo, considerando a data e o horário de distribuição, o que resultou em um conjunto final de 8.729 processos válidos. Os textos foram submetidos a uma limpeza mínima, voltada à remoção de caracteres especiais, à quebra de linha e à normalização de espaços. Procedimentos clássicos como remoção de *stopwords*, *stemming* e lematização não foram aplicados, de modo a preservar o texto completo para a análise com modelos *Transformers*, que dependem do contexto léxico e sintático original do documento [Devlin et al. 2019].

A fase de **Modelagem** compreende a construção da base experimental e a preparação dos dados para o treinamento e a avaliação dos modelos supervisionados. Essa etapa integra atributos relacionais, temporais e textuais extraídos dos processos, com o objetivo de representar indícios computacionais associados à litigância predatória. A Figura 1 sintetiza o fluxo adotado, que parte dos processos judiciais e resulta em modelos treinados e avaliados a partir de uma base rotulada com rotulagem fraca.

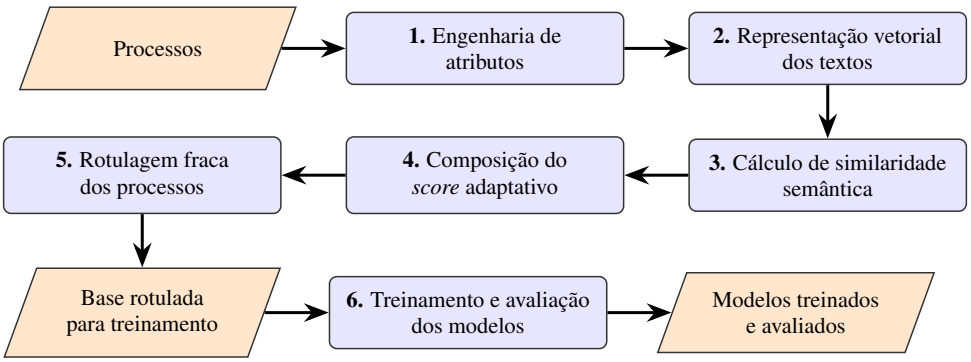


Figura 1. Fluxo geral da construção da base experimental e da etapa de treinamento dos modelos.

Seguindo o fluxo apresentado na Figura 1, a construção da base teve início com a engenharia de atributos (Etapa 1), responsável por transformar os dados processuais em variáveis capazes de representar padrões potencialmente relevantes para a identificação de litigância predatória. Inicialmente, foram construídos atributos relacionais com base nos vínculos entre advogados, entre partes ativas (autor) e entre partes passivas (réu). Para cada processo, foram geradas combinações entre esses três tipos de atores, em que um processo com dois advogados, uma parte ativa e dois réus distintos gera quatro trios para análise, correspondentes às combinações possíveis entre esses atores.

A partir dos trios formados, foram obtidos atributos de volume e de repetição, considerando a quantidade de processos em cada trio. Quanto maior a recorrência do trio, mais forte é o indício de atuação reiterada, sobretudo quando esse padrão se associa a outros indícios, como proximidade temporal e semelhança textual entre as petições. A análise estrutural é complementada pelos atributos temporais relacionados ao trio, calculados a partir dos intervalos, em dias, entre as datas de distribuição, permitindo medir se as ações foram propostas em intervalos próximos ou distribuídas ao longo do tempo, pois a concentração de muitos processos em um curto período reforça a repetição.

Por fim, foram considerados atributos contextuais para minimizar falsos positi-

vos e evitar que padrões legítimos de litigância fossem interpretados como indícios de litigância predatória. Para isso, foi analisada a combinação entre advogado e autor, verificando quantas partes passivas distintas estavam associadas a essa mesma dupla, pois alguns tipos de demanda podem envolver muitos processos semelhantes por razões legítimas, como ações de cobrança contra diferentes devedores.

Após a engenharia de atributos, os textos das petições foram convertidos em representações *embeddings* (Etapa 2), de modo que petições semelhantes possam ser comparadas em um espaço vetorial. Para tal, foi utilizado o modelo BumbaLM[4], treinado a partir de dados jurídicos em português brasileiro, adequado ao uso no conjunto de dados. Contudo, devido ao limite de 1.024 *tokens* processáveis por inferência, utilizou-se a técnica de *chunking*, com janelas de 900 *tokens* e sobreposição de 128 *tokens* para evitar a perda de contexto semântico. Posteriormente, os vetores foram agregados em uma única representação do processo, normalizados pela norma $L2$, resultando em um único vetor representativo por processo.

Com as representações vetoriais geradas e normalizadas, foi calculado o valor de similaridade semântica (Etapa 3) entre todos os processos por meio do cosseno, de modo que valores mais próximos de 1 indicassem maior proximidade semântica entre os textos. Após o cálculo, foram mantidos apenas os pares de processos que atingiram o limiar $s \geq 0,95$, e os pares acima desse limiar foram associados aos trios. Para cada par de processos, foi verificado se ambos compartilhavam o mesmo trio e, quando havia uma interseção entre eles, a similaridade era associada ao respectivo trio, sendo possível calcular a quantidade de pares similares dentro do mesmo trio e os valores médio e máximo de similaridade textual.

Na composição do *score* adaptativo (Etapa 4), assim denominado por ser definido a partir das características do conjunto de dados analisado, os diferentes indícios foram reunidos em uma medida para ordenar os processos conforme a intensidade dos padrões observados, considerando que o potencial predatório pode estar relacionado a valores altos (*e.g.* repetição do trio) ou baixos (*e.g.* concentração temporal). Nesse sentido, os valores brutos foram convertidos em percentis e reescalados para o intervalo $[0, 1]$ por meio da transformação Min-Max, invertendo a escala quando necessário. Os pesos dos indícios foram definidos automaticamente com base na matriz de correlação de Spearman, visando evitar a dupla contagem de indícios redundantes, como a similaridade máxima e média do processo, que receberam pesos menores na composição final.

Com o *score* calculado, foi realizada a etapa de rotulagem fraca (Etapa 5) com base na posição relativa de cada processo na distribuição do *score*. Para isso, os processos foram organizados em quantis e separados em três níveis de prioridade, a saber: a classe de *Alta prioridade* é atribuída aos processos com os 5% maiores *scores* da distribuição (acima do percentil 95); processos entre os percentis 80 e 95 foram rotulados como *Suspeito*; e processos abaixo do percentil 80 foram rotulados como *Baixa prioridade*. Desse modo, a saída dessa etapa é uma base rotulada, composta pelos atributos calculados e por uma variável alvo, utilizada na etapa seguinte para o treinamento dos modelos.

Prosseguindo para a etapa de treinamento dos modelos (Etapa 6), optou-se por não utilizar o texto das petições iniciais como entrada para os classificadores. Como os

[4]Disponível em: https://huggingface.co/fabricioalmeida/BumbaLM-Embedding-4B-v0.1

documentos possuem um número de *tokens* superior ao permitido pelos modelos BERT, o truncamento para 512 *tokens* descartaria informações relevantes. Por outro lado, a estratégia de *chunking*, embora preserve o conteúdo textual, aumentou o custo computacional e o consumo de memória, tornando-a inviável para a fase de treinamento. Diante desses obstáculos, foi adotada uma estratégia baseada em [Lin et al. 2024], que propõe a conversão de dados tabulares em texto, transformando os atributos estruturados em descrições textuais para serem processados por modelos de linguagem. No presente estudo, os atributos foram convertidos em narrativas curtas e padronizadas para serem utilizadas como entrada dos modelos, reduzindo o custo de processamento. Um exemplo pode ser observado na Tabela 2.

Tabela 2. Exemplo da narrativa textual

Rótulo associado: Alta prioridade
similaridade_maxima: 1.00; faixa_similaridade: muito_alta; similaridade_media: 0.96; pares_similares: 199; texto_duplicado_exato: sim; processos_mesmo_texto: 2; processos_mesmo_trio: 2; pares_similares_trio: 1; similaridade_media_trio: 1.00; similaridade_maxima_trio: 1.00; intervalo_minimo_dias: 20; urgencia_temporal: moderada; baixa_diversidade_passiva: sim; partes_passivas_distintas: 3; processos_dupla_advogado: 6; razao_diversidade_passiva: 0.50.

A partir disso, o aprendizado foi conduzido sobre dois modelos BERT especializados no domínio jurídico, o BumbaBERT *Small* [Carmo et al. 2023][5] e o LegalBERT-pt FP [Silveira et al. 2023][6], em que cada modelo foi treinado a partir de duas estratégias de treinamento. Na primeira, o codificador foi mantido congelado, com treinamento apenas da camada de classificação, enquanto, na segunda, foi realizado o *fine-tuning* completo, permitindo que todos os parâmetros do modelo fossem treinados. Os experimentos foram executados em dois cenários, um utilizando como variável alvo o rótulo fraco (proposta do presente estudo) e outro empregando as etiquetas da coluna *str_nm_etiqueta* como referência comparativa, de modo a observar em que medida os modelos capturam padrões convergentes às marcações institucionais disponíveis.

Foi adotada a estratégia *StratifiedGroupKFold* com $k = 5$ para o treinamento dos modelos, usando o trio como unidade de agrupamento para preservar a distribuição das classes em cada divisão e impedir que processos pertencentes ao mesmo grupo fossem alocados simultaneamente em treino e teste, prevenindo o vazamento de dados. Em cada *fold*, a divisão dos dados seguiu a proporção 70, 15 e 15 para os conjuntos de treino, teste e validação, respectivamente, e os mesmos *folds* foram reutilizados em cada cenário experimental para evitar que diferenças de desempenho fossem atribuídas a variações nas partições da base. Em relação ao desbalanceamento de classes, a estratégia adotada foi a de *WeightedRandomSampler*, em que cada amostra recebe um peso inversamente proporcional à frequência de sua classe, de modo que instâncias de classes minoritárias tenham maior probabilidade de serem selecionadas nos lotes de treinamento, produzindo um efeito de *oversampling*, sem duplicar os registros da base.

Os hiperparâmetros foram definidos com base na estratégia padrão de *fine-tuning* de modelos BERT proposta por [Devlin et al. 2019], com 5 épocas, tamanho de lote de

[5]Disponível em: https://huggingface.co/fabricioalmeida/bumbabert-small-v3
[6]Disponível em: https://huggingface.co/raquelsilveira/legalbertpt_fp

16, otimizador AdamW, taxa de aprendizado de $2e{-}5$, regularização por *weight decay* de $0,01$ e parada antecipada com paciência de 2. A seleção do melhor modelo em cada *fold* foi realizada com base no *F1-score* macro obtido no conjunto de validação, de modo que, na ausência de melhora nessa métrica por duas épocas consecutivas, o treinamento era interrompido e o estado mais bem-sucedido do modelo era restaurado antes da avaliação no conjunto de teste. O ambiente de execução possuía 8 GB de memória RAM e uma GPU NVIDIA L40S com 48 GB de VRAM.

Para a **Avaliação**, as métricas calculadas incluíram acurácia, *F1-score* macro, *F1-score* ponderado, precisão macro, *recall* macro, acurácia balanceada, coeficiente de correlação de Matthews (MCC) e matriz de confusão. A métrica principal para comparação foi o *F1-score* macro, por atribuir o mesmo peso a cada classe, independentemente da frequência de cada uma, o que é relevante em um cenário de classes desbalanceadas [Farhadpour et al. 2024]. Na etapa de **Implementação**, os resultados obtidos fornecem subsídios técnicos para a futura incorporação de modelos de apoio à triagem de processos judiciais, auxiliando equipes técnicas, magistrados e servidores na seleção de processos que demandam análise mais detalhada. A Seção seguinte apresenta estes resultados, considerando as métricas de desempenho, as matrizes de confusão e a comparação entre as configurações experimentais adotadas.

4. Resultados

A distribuição dos rótulos fracos e das etiquetas institucionais apresenta comportamento semelhante, com maior concentração em categorias intermediárias ou não conclusivas e menor presença nas categorias associadas a indícios mais fortes de litigância predatória. Na rotulagem fraca, a maior parte dos processos foi classificada como *Suspeito* (4.139), seguida de *Baixa prioridade* (3.522) e *Alta prioridade* (1.068). Nas etiquetas institucionais, observa-se a predominância de *Possível Prevenção* (4.018), seguida por *Sem Etiqueta* (3.788) e *Possível Demanda Predatória* (923). Em relação ao desempenho dos modelos, a Figura 2 apresenta o *F1-score* macro obtido nos oito cenários experimentais, sendo a principal métrica utilizada, pois permite comparar os modelos em um cenário multiclasse desbalanceado.

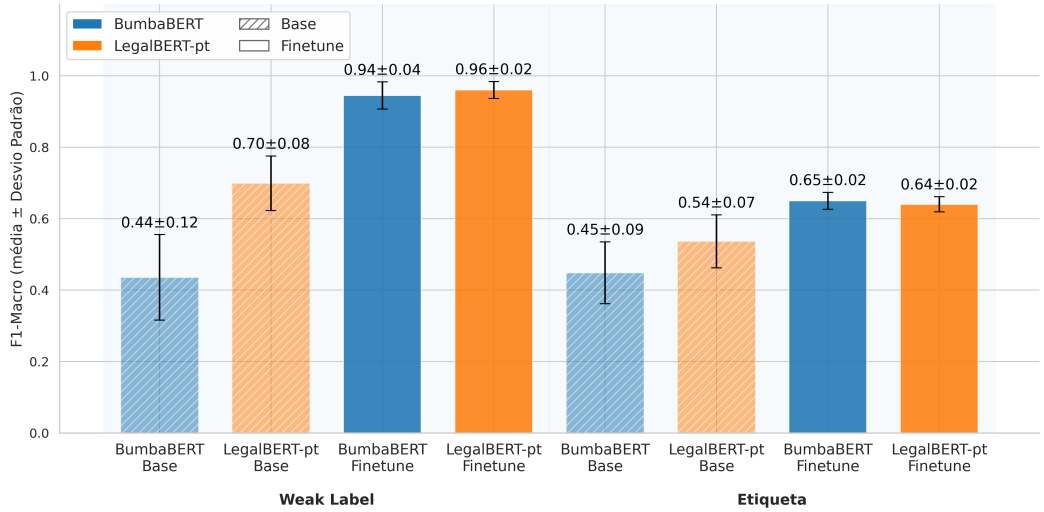


Figura 2. Desempenho geral dos modelos segundo o *F1-score* macro.

Na Figura 2, observa-se que a escolha do modelo não foi determinante para o desempenho. Apesar de o LegalBERT-pt obter vantagem no modelo base, com 0,70 contra 0,44 no rótulo fraco, os dois modelos se equiparam após o *fine-tuning*, configuração que apresentou os melhores resultados. O LegalBERT-pt alcançou 0,96 enquanto o BumbaBERT obteve desempenho próximo com o valor de 0,94. O fator determinante para este resultado foi o modo de treinamento, com ganhos de até 0,50 em relação aos modelos base, dado que as narrativas verbalizadas seguem um formato diferente do padrão jurídico aprendido no pré-treinamento, ao qual não estão plenamente adaptados. Esse comportamento está em consonância com o estudo de [Santos et al. 2025], no qual o ajuste fino de modelos BERT também se mostrou eficaz em tarefas de classificação multiclasse. Para complementar a análise baseada no *F1-score* macro, a Tabela 3 apresenta outras métricas obtidas nos experimentos.

Tabela 3. Comparação das métricas dos modelos com *fine-tuning* (média e desvio padrão)

Métrica	BumbaBERT Rótulo fraco	LegalBERT-pt Rótulo fraco	BumbaBERT Etiqueta	LegalBERT-pt Etiqueta
F1-score macro	0,94 ± 0,04	**0,96 ± 0,02**	0,65 ± 0,02	0,64 ± 0,02
Precisão macro	0,94 ± 0,05	**0,95 ± 0,03**	0,68 ± 0,01	0,67 ± 0,02
Recall macro	0,96 ± 0,02	**0,97 ± 0,01**	0,70 ± 0,01	0,67 ± 0,02
Acurácia balanceada	0,96 ± 0,02	**0,97 ± 0,01**	0,70 ± 0,01	0,67 ± 0,02
MCC	0,94 ± 0,03	**0,96 ± 0,02**	0,59 ± 0,02	0,59 ± 0,09

Como observado na Tabela 3, os modelos treinados com o alvo de rótulo fraco mantiveram desempenho elevado em todas as métricas, com *recall* macro e acurácia balanceada acima de 0,95 e MCC de até 0,96 no LegalBERT-pt, o que indica bom desempenho entre as classes e baixa concentração de acertos na classe majoritária. Nos experimentos com o alvo de Etiqueta, os resultados foram mais moderados, com *F1-score* macro entre 0,64 e 0,65, e MCC de 0,59 para ambos os modelos. Ainda assim, os valores de *recall* macro e de acurácia balanceada indicam que os classificadores conseguiram capturar parte da separação entre as classes, embora com menor estabilidade em comparação ao alvo de rótulo fraco.

As matrizes de confusão foram construídas a partir da agregação das predições obtidas nos conjuntos de teste de cada um dos cinco *folds*. Em cada matriz, as linhas indicam as classes verdadeiras e as colunas, as classes preditas pelos modelos. Para os rótulos fracos, foram utilizadas as abreviações AP para *Alta prioridade predatória*, BP para *Baixa prioridade* e SUS para *Suspeito*. Para as etiquetas, foram utilizadas as abreviações PDP para *Possível demanda predatória*, PP para *Possível prevenção* e SE para *Sem etiqueta*. As Tabelas 4 e 5 apresentam as matrizes.

Tabela 4. Matrizes de confusão para o alvo de rótulo fraco

	BumbaBERT / Base			LegalBERT-pt / Base			BumbaBERT / Finetuned			LegalBERT-pt / Finetuned		
Real \ Pred.	AP	BP	SUS	AP	BP	SUS	AP	BP	SUS	AP	BP	SUS
AP	354	409	305	209	20	839	1.003	0	65	1.025	0	43
BP	0	3.491	31	0	3.522	0	0	3.522	0	0	3.522	0
Sus	1.154	1.932	1.053	194	140	3.805	200	33	3.906	141	0	3.998

Tabela 5. Matrizes de confusão para o alvo de etiqueta

	BumbaBERT / Base			**LegalBERT-pt / Base**			**BumbaBERT / Finetuned**			**LegalBERT-pt / Finetuned**		
Real \ Pred.	PDP	PP	SE	PDP	PP	SE	PDP	PP	SE	PDP	PP	SE
PDP	158	371	394	511	383	29	614	294	15	406	502	15
PP	752	1.582	1.684	2.141	1.737	140	1.730	2.187	101	1.208	2.709	101
SE	26	138	3.624	153	157	3.478	151	162	3.475	80	236	3.472

A Tabela 4 confirma o desempenho superior dos modelos com *fine-tuning* no alvo de rótulo fraco. Nos dois modelos ajustados, a classe BP foi praticamente recuperada, enquanto a classe Sus também apresentou um alto número de acertos, especialmente no LegalBERT-pt, com 3.998 predições corretas. A principal diferença entre os modelos ocorre na classe AP, na qual o LegalBERT-pt apresentou o maior número de acertos e a menor confusão com Sus. Nos cenários com modelo base, há maior dispersão dos erros no BumbaBERT, o que confundiu parte dos casos de AP e Sus com outras classes.

Na Tabela 5, os erros aparecem de forma mais distribuída entre as classes, o que é consistente com os resultados observados nas métricas gerais. A classe SE foi a mais bem reconhecida em todos os cenários, com mais de 3.400 acertos nos modelos base e de *fine-tuning*. Já as classes PDP e PP apresentaram maior sobreposição, principalmente porque muitos casos de PP foram classificados como PDP ou SE, dependendo do modelo. Entre os cenários ajustados, o BumbaBERT apresentou melhor recuperação da classe PDP, enquanto o LegalBERT-pt concentrou mais acertos na classe PP. Esses resultados indicam que, no alvo institucional, os modelos distinguem bem os processos sem etiqueta, mas encontram maior dificuldade na separação entre as categorias institucionais associadas à possível prevenção e à possível demanda predatória.

Os resultados indicam que as narrativas verbalizadas sustentam a priorização proposta e convergem parcialmente com a triagem institucional, com maior segurança na identificação dos processos sem etiqueta do que na separação entre as duas categorias de alerta.

5. Considerações Finais

Este trabalho apresentou uma abordagem para apoiar a identificação preliminar de litigância predatória em processos judiciais, combinando atributos relacionais, temporais e semânticos com uma estratégia de rotulagem fraca e de treinamento supervisionado. Para tal, os processos foram estruturados com base em indícios, como a repetição de atores, a similaridade textual entre petições e a concentração temporal de protocolos, o que viabilizou a construção de uma base de dados para os experimentos de classificação. Os resultados demonstraram que os modelos apresentaram melhor desempenho nos cenários de *fine-tuning* quando o alvo era o rótulo fraco. O LegalBERT-pt obteve as melhores métricas, com um resultado próximo ao do BumbaBERT, indicando que as narrativas verbalizadas preservam os indícios utilizados na rotulagem fraca e permitem distinguir os níveis de prioridade definidos. Nos experimentos com as etiquetas, o desempenho foi mais moderado, sobretudo na separação entre *Possível Prevenção* e *Possível Demanda Predatória*, embora os modelos tenham capturado parte da estrutura dessas marcações.

Do ponto de vista prático, a abordagem oferece um mecanismo de triagem inicial que pode apoiar o TJMA na priorização de processos com indícios de litigância

predatória, reduzindo o esforço manual de análise. Como limitações, destaca-se que a avaliação do modelo foi conduzida exclusivamente com base em métricas quantitativas, sem validação qualitativa por especialistas do domínio jurídico, o que poderia fornecer subsídios adicionais quanto à relevância prática e à aceitabilidade institucional das predições. Além disso, os dados utilizados são restritos ao TJMA, o que pode limitar a generalização dos resultados a outros contextos institucionais. Como trabalhos futuros, estão planejadas a incorporação de avaliações por especialistas, visando aumentar a robustez e a transparência da solução proposta, a comparação com classificadores tradicionais aplicados aos atributos estruturados, para mensurar o ganho de desempenho a partir do conteúdo das petições, e a investigação do uso de LLMs para auxiliar na análise e na caracterização desses padrões.

Agradecimentos e Uso de IA generativa

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio das bolsas DT-303031/2023-9, DT 302850/2023-6 e DT-315052/2026-0; pelo Tribunal de Justiça do Maranhão, por meio do Acordo de Cooperação Técnica N^o 02/2021(processo N^o 38328/2020-TJ/MA); e pela Financiadora de Estudos e Projetos (FINEP) – Pró-Amazônia – 2.373/24 – CTCCA-II.

Declara-se que os modelos de IA generativa *GPT 5.5*, *NotebookLM* e *Grammarly Pro* foram utilizados como ferramentas de apoio para a revisão gramatical e auxílio na seleção de artigos relevantes para compor as referências da pesquisa, bem como o *Claude Sonnet 4.6* para melhoria de códigos, aprimoramento do *pipeline* e desenho de pesquisa. A autoria e a responsabilidade integral pelo conteúdo final, incluindo a verificação de plágio e as correções, são da primeira pessoa autora.

Referências

Borges, B., Faria, E., and Gabriel, P. (2025). Comparative analysis of text classification algorithms. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 189–200, Porto Alegre, RS, Brasil. SBC.

Carmo, F., Serejo, F., Junior, A. J., Santana, E., and Lobato, F. (2023). Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 188–199, Porto Alegre, RS, Brasil. SBC.

de Araújo, V. S., de Paiva Gabriel, A., and Porto, F. R. (2022). Justiça 4.0: a transformação tecnológica do poder judiciário deflagrada pelo cnj no biênio 2020-2022. *Revista Eletrônica Direito Exponencial-DIEX*, 1(1):1–18.

de Souza, G. (2024). Litigância predatória, tutela coletiva e o porvir do acesso à justiça. *Revista de Processo— vol*, 353.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Farhadpour, S., Warner, T. A., and Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, 16(3):533.

Gheibi, R. and Hougen, D. (2025). Improved cervical cancer classification using generative ai and language models. In *2025 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–8. IEEE.

Justino, A., Junior, A. J., Marcacini, R., and Lobato, F. (2025). Evaluating bert models for semantic retrieval in long portuguese legal documents. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 1045–1056, Porto Alegre, RS, Brasil. SBC.

Lin, T., Yan, J., Jurgens, D., and Tomkins, S. J. (2024). Tab2Text - a framework for deep learning with tabular data. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12925–12935, Miami, Florida, USA. Association for Computational Linguistics.

Magalhães, J. L., de Sousa, I. D. V., et al. (2024). Litigância judicial abusiva e instrumentos de gestão processual conferidos ao juiz no código de processo civil: a necessidade de preservação do direito fundamental de acesso à justiça. *Revista Foco*, 17(4).

Melo, R. M. (2023). Discussão sobre a viabilidade técnica e jurídica para a aplicação de processamento de linguagem natural em decisões vinculantes em processos judiciais. *Anais do 31º Encontro Ibero Americano de Governo Eletrônico e Inclusão Digital*.

Neto, J. C. d. S. et al. (2025). Similaridade semântica entre processos judiciais: Uma abordagem para identificação de litigância predatória.

Paradela, A. S. and Cavaco, B. d. S. B. (2025). Litigância abusiva e a necessária formação de grupo de trabalho multissetorial: o papel do ministério público na defesa da sociedade e na preservação do acesso à justiça. *Diké-Revista Jurídica*, 24(29):128–156.

Ralph, P. and Baltes, S. (2022). Paving the way for mature secondary research: the seven types of literature review. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

Rampim, T. and Igreja, R. L. (2022). Acesso à justiça e transformação digital: um estudo sobre o programa justiça 4.0 e seu impacto na prestação jurisdicional. *Direito Público*.

Santos, W., Filho, A. G., Oliveira, S., and Presa, J. (2025). Comparing prompt-based llms, fine-tuning, and classical models for legal text classification in portuguese. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 1138–1149, Porto Alegre, RS, Brasil. SBC.

Silva, M. E. L. and Neris, J. S. (2025). A litigância predatória nas demandas massificadas consumeristas: Um estudo sob a ótica do poder geral de cautela do juízo e o entendimento do stj no tema de julgamento repetitivo 1.198. *Graduação em Movimento-Ciências Jurídicas*, 4(3):352–376.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Anais da XII Brazilian Conference on Intelligent Systems*, Porto Alegre, RS, Brasil. SBC.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.