# Análise comparativa de métodos baseados em modelos de linguagem para documentos jurídicos longos

**Gabriele S. Araújo**[1], **Fabrício A. do Carmo**[2], **Ewaldo E. C. Santana**[1,2], **Antonio F. L. Jacob Junior**[1], **Fábio M. F. Lobato**[1,3]

[1] Programa de Pós-Graduação em Engenharia da Computação e Sistemas (PECS)
Universidade Estadual do Maranhão (UEMA) – São Luís – MA – Brazil

[2]Programa de Pós-graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Maranhão (UFMA) – São Luís – MA – Brazil

[3]Instituto de Engenharia e Geociências (IEG)
Universidade Federal do Oeste do Pará (UFOPA) – Santarém – PA – Brazil

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

***Resumo.*** *Com mais de 80 milhões de processos em trâmite, o judiciário brasileiro enfrenta uma crise de eficiência, comprometendo a celeridade e qualidade da prestação jurisdicional. Modelos de linguagem baseados em Transformers, como o BumbaBERT, têm sido usados para otimizar o processamento de dados jurídicos, mas enfrentam limitações com documentos longos devido ao custo computacional e à restrição de tokens. Com isso, este estudo tem o objetivo de comparar e avaliar métodos existentes na literatura para superar essas limitações. Os resultados indicam que métodos de seleção de sentenças melhoram o desempenho na classificação de documentos. Com isso, este trabalho tem em vista aprimorar a eficiência do sistema judiciário, alinhando-se aos objetivos do programa Justiça 4.0.*

## 1. Introdução

O sistema judiciário brasileiro enfrenta um desafio crescente com o aumento exponencial de processos judiciais, ultrapassando os 80 milhões de processo ativos em julho de 2024[1]. Esta sobrecarga impacta significativamente a eficiência e celeridade da prestação jurisdicional, resultando em morosidade e descrédito do sistema [de Almeida and de Almeida Pinto 2022]. Em resposta a este cenário, o Conselho Nacional de Justiça (CNJ) estabeleceu diretrizes para a implementação de Inteligência Artificial (IA) no Poder Judiciário através da Resolução nº 332/2020 [CNJ 2020], além do programa Justiça 4.0 em 2021, uma iniciativa conjunta do CNJ e do Programa das Nações Unidas para o Desenvolvimento focada em modernizar o acesso à justiça por meio de novas tecnologias [CNJ and PNUD 2021].

Neste contexto, destaca-se o Acordo de Cooperação Técnica entre o Tribunal de Justiça do Maranhão (TJMA) e a Universidade Estadual do Maranhão (UEMA) que resultou no desenvolvimento do modelo de linguagem BumbaBERT baseado na arquitetura *Bidirectional Encoder Representations for Transformers* (BERT) e pré-treinado para o domínio jurídico brasileiro [do Carmo 2024]. Contudo, assim como outros modelos baseados em *Transformers*, o BumbaBERT enfrenta limitações no processamento

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

de documentos longos, um desafio particularmente relevante no contexto jurídico, onde peças processuais frequentemente excedem o limite padrão de 512 *tokens* dos modelos [Kalamkar et al. 2022, Devlin 2018].

Frente a este cenário, o presente trabalho propõe investigar e comparar diferentes métodos baseados em modelos de linguagem para o processamento eficiente de documentos jurídicos. Espera-se que os resultados deste estudo contribuam para a otimização do fluxo processual, alinhando-se aos objetivos do Justiça 4.0 e promovendo uma transformação digital no judiciário brasileiro. O resumo expandido encontra-se organizado como segue: os materiais e métodos e resultados obtidos são descritos nas Seções 2 e 3, respectivamente. Por fim, as considerações finais são apresentadas na Seção 4.

## 2. Materiais e Métodos

Este estudo seguiu a metodologia *Data Science Trajectories* (DST), que proporciona flexibilidade para que cientistas de dados adaptem o fluxo de trabalho às necessidades específicas do projeto [Martínez-Plumed et al. 2019]. A trajetória escolhida neste trabalho foi organizada em etapas interdependentes, começando pelo entendimento do negócio, seguido da exploração e preparação dos dados, modelagem e, por fim, avaliação. O **entendimento do negócio** envolveu a identificação das necessidades específicas do domínio jurídico, com foco no aprimoramento do processamento de documentos longos. Em seguida, foi realizada a **exploração do valor dos dados**, utilizando um *dataset* de 5.494 petições iniciais do TJMA, categorizadas em 8 tipos de Incidentes de Resolução de Demandas Repetitivas (IRDR). Na Tabela 1 é apresentada a distribuição das classes e as estatísticas de tamanho dos documentos.

**Tabela 1. Conjunto de dados de petições iniciais do TJMA**

| Tipo de IRDR | Amostras | Incidência (%) | Média de *Tokens* (mean ± std) |
|---|---|---|---|
| 1 | 3.581 | 65,18 | 9.306,60 ± 12.427,28 |
| 2 | 27 | 0,49 | 7.132,00 ± 4.834,97 |
| 3 | 502 | 9,14 | 8.202,29 ± 6.166,80 |
| 4 | 54 | 0,98 | 6.758,00 ± 4.866,07 |
| 5 | 1.068 | 19,44 | 10.827,00 ± 13.494,84 |
| 6 | 4 | 0,07 | 3.978,50 ± 1.641,17 |
| 7 | 6 | 0,11 | 6.380,33 ± 4.996,18 |
| 8 | 252 | 4,59 | 12.235,66 ± 7.451,56 |
| TOTAL | 5.494 | 100 | 9592.79 ± 11974.39 |

A contagem média de *tokens* foi obtida por meio do tokenizador do BumbaBERT [do Carmo 2024]. Com isso, observa-se que 100% dos documentos excedem o limite de 512 *tokens*, com média geral de 9.592,79 *tokens*, reforçando a relevância do estudo, onde em modelos comuns a maioria das informações dos documentos podem ser perdidas [Kalamkar et al. 2022]. A etapa de **preparação dos dados** incluiu o pré-processamento, a divisão e estratificação do conjunto de treino, validação e teste, além da tokenização e codificação dos textos.

Para a **modelagem**, investigaram-se métodos que abordam limitações do *Transformer* em documentos longos. Entre os métodos selecionados, destaca-se o ToBERT, uma abordagem hierárquica para documentos longos proposta por [Pappagari et al. 2019].

Esse modelo divide os documentos em *chunks* de 200 *tokens* e utiliza uma camada *Transformer* sobre as representações BERT desses *chunks*. Outros modelos, como o BERT+TextRank e o BERT+Random, propostos por [Park et al. 2022], também foram analisados. Essas abordagens concatenam a representação BERT dos primeiros 512 *tokens* e selecionam outros 512 *tokens* utilizando diferentes estratégias: o BERT+TextRank usa o algoritmo TextRank para selecionar sentenças relevantes, enquanto o BERT+Random escolhe sentenças aleatoriamente. Os códigos experimentais foram disponibilizados pelos autores e adaptados para o contexto jurídico utilizando como ponto de partida o modelo BumbaBERT de [do Carmo 2024], com otimização de hiperparâmetros como tamanho de lote (3e-5 a 5e-5), número de épocas (10) e *dropout* (0.1). Por fim, a **avaliação** do desempenho dos modelos foi realizada utilizando métricas como acurácia, precisão-*macro*, *recall-macro* e *F1-macro* na tarefa de classificação de petições iniciais, conforme sugerido nos estudos de [Park et al. 2022] e [Pappagari et al. 2019].

## 3. Resultados e Discussão

A análise dos resultados obtidos na classificação de petições iniciais utilizando diferentes abordagens de processamento de textos longos revelou *insights* significativos sobre o desempenho dos modelos adaptados, conforme apresentado na Tabela 2.

**Tabela 2. Classificação de petições iniciais**

| Modelo | Acurácia | Precisão-*macro* | *Recall-macro* | *F1-macro* |
|---|---|---|---|---|
| BumbaBERT_Baseline | 0.81 | 0.44 | 0.40 | 0.42 |
| BumbaBERT+Random | **0.84** | 0.58 | 0.47 | 0.51 |
| BumbaBERT+TextRank | 0.82 | **0.73** | **0.54** | **0.60** |
| BumbaBERT+ToBERT | 0.83 | 0.49 | 0.40 | 0.42 |

Os resultados do modelo BumbaBERT_Baseline, ponto de referência para este estudo, foram obtidos de [do Carmo 2024] utilizando o mesmo conjunto de dados. A adaptação BumbaBERT+Random, que complementa os primeiros 512 *tokens* com sentenças aleatórias do documento demonstrou uma melhoria notável, alcançando uma acurácia de 0.84 e um *F1-macro* de 0.51, sugerindo que a inclusão de informações adicionais mesmo que selecionadas aleatoriamente pode contribuir para uma melhor compreensão do contexto global do documento [Park et al. 2022]. No BumbaBERT+TextRank, embora sua acurácia tenha sido ligeiramente inferior à do BumbaBERT+Random, o *F1-macro* alcançou um valor superior entre todos os modelos testados (0.60), indicando que a seleção de sentenças baseada em relevância conforme implementada pelo algoritmo TextRank pode ser especialmente eficaz na captura de informações úteis para a classificação de petições iniciais. Por fim, a abordagem BumbaBERT+ToBERT apresentou uma acurácia de 0.83, superando ligeiramente o *baseline*. No entanto, seu *F1-macro* foi equivalente ao do modelo base, sugerindo que, embora a abordagem hierárquica melhore a classificação geral, ela pode não ser igualmente eficaz para todas as classes.

Estes resultados corroboram a hipótese de que adaptações específicas para o processamento de textos longos podem melhorar consideravelmente o desempenho de modelos de linguagem no domínio jurídico [Park et al. 2022]. A variação observada entre acurácia e *F1-macro* nos diferentes modelos ressalta a importância de considerar múltiplas métricas ao avaliar o desempenho em tarefas de classificação multiclasse, especialmente em conjuntos de dados desbalanceados, comum em aplicações jurídicas.

## 4. Conclusão

Neste estudo foram comparados métodos para processamento de documentos jurídicos longos, focando na análise de técnicas de seleção de sentenças e abordagens hierárquicas para estender a capacidade dos modelos além do limite. Os resultados demonstram que métodos baseados em seleção de sentenças oferecem melhorias significativas na classificação de documentos jurídicos. Desse modo, este estudo contribui para a modernização do sistema judiciário, em consonância com as metas de transformação digital do programa Justiça 4.0 e do Acordo de Cooperação TJMA/UEMA. Pesquisas futuras poderão explorar a combinação de abordagens ou técnicas de Ensemble para otimizar o equilíbrio entre acurácia global e desempenho por classe, além de investigar a aplicabilidade destes métodos em outros tipos de documentos jurídicos.

## Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9; e 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

CNJ (2020). Resolução n. 332, de 21 de agosto de 2020. *Diário da Justiça do Conselho Nacional de Justiça*, Brasília, DF. Acesso em: 2 set. 2024.

CNJ and PNUD (2021). Cartilha justiça 4.0. Disponível em: https://www.cnj.jus.br/tecnologia-da-informacao-ecomunicacao/justica-4-0/cartilhas/.

de Almeida, N. D. and de Almeida Pinto, P. A. L. (2022). O uso da inteligência artificial como ferramenta de eficiência e acesso à justiça em revisão sistemática da literatura. *Research, Society and Development*, 11(11):e349111133674–e349111133674.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

do Carmo, F. A. (2024). Representações *Embeddings* Orientadas à Linguagem Jurídica Brasileira. Master's thesis, Universidade Estadual do Maranhão, São Luís - MA.

Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., and Modi, A. (2022). Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*.

Park, H., Vyas, Y., and Shah, K. (2022). Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.