Aplicação de Modelos BERT Especializados para o Aprimoramento na Classificação de Sentenças Jurídicas

Gustavo S. Silva[1], **Gabriele S. Araújo**[1], **Antonio F. L. Jacob Jr.**[1]

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

{gustavosilva7, gabriele.20231002966}@aluno.uema.br,
antoniojunior@professor.uema.br

Resumo. *Soluções baseadas em Processamento de Linguagem Natural (PLN) são adotadas com o objetivo de otimizar a gestão processual e garantir maior uniformidade e previsibilidade nas decisões judiciais. Modelos de linguagem, como o Bidirectional Encoder Representations from Transformers (BERT), são alternativas viáveis para capturar as particularidades da linguagem legal. Este trabalho investiga a aplicação de Parameter-Efficient Fine-Tuning (PEFT) com Low-Rank Adaptation (LoRA) aos modelos LegalBert-pt e BumbaBert, com abordagens de truncamento e agrupamento dos documentos para o aumento da precisão na classificação de temas de Incidentes de Resolução de Demandas Repetitivas (IRDRs) em sentenças jurídicas.*

1. Introdução

O relatório anual "Justiça em Números 2024" do Conselho Nacional de Justiça (CNJ) sobre a movimentação processual revela que uma parte significativa do acervo judiciário brasileiro é composta por demandas de massa, muitas delas tratando da mesma matéria jurídica [Melo 2021]. Na prática, verifica-se que questões judiciais de mesmo pleito são julgadas por tribunais distintos, resultando, muitas vezes, em decisões divergentes, o que compromete a segurança jurídica e a isonomia. Com o objetivo de mitigar esse problema, o Código de Processo Civil (CPC) de 2015 (Lei nº 13.105/15) introduziu o sistema de precedentes judiciais, com o objetivo de garantir maior estabilidade e previsibilidade nas decisões. Um dos mecanismos destacados é o Incidente de Resolução de Demandas Repetitivas (IRDR), utilizado quando houver cumulativamente a repetição e controvérsias de direito em processos na área de jurisdição do Tribunal.

Abordagens promissoras para reconhecer e aplicar as regras dos precedentes IRDR em documentos legais podem ser construídas com o uso de Processamento de Linguagem Natural (PLN). Modelos de linguagem baseados no *Bidirectional Encoder Representations from Transformers* (BERT) ajustados para documentos jurídicos, têm mostrado alto desempenho na classificação de textos legais [Polo et al. 2021]. O LegalBert-pt [Silveira et al. 2023], por exemplo, traz modelos treinados em 1,5 milhão de documentos legais do ordenamento brasileiro. Outro estudo relevante apresenta o desenvolvimento do BumbaBert [Carmo et al. 2023] para a construção de modelos *embeddings* orientados ao âmbito jurídico por meio de 5,4 milhões de documentos contendo acórdãos e petições iniciais, visando alimentar aplicações na área.

Uma estratégia comum para maximizar o desempenho desses modelos em tarefas específicas é a especialização por meio de ajuste fino. A técnica *Low-Rank Adapta-*

tion (LoRA) [Hu et al. 2021], um método de ajuste eficiente de parâmetros (*Parameter-Efficient Fine-Tuning*, PEFT) baseado em reparametrização, destaca-se por permitir a adaptação de modelos pré-treinados sem a necessidade de modificar todos os parâmetros.

Diante disso, a implementação de um projeto que envolva modelos de linguagem para a classificação de precedentes judiciais torna-se não apenas uma solução tecnicamente eficaz, mas também uma contribuição relevante para o sistema judiciário brasileiro, com o potencial de reduzir a morosidade processual e aumentar a eficiência e a segurança jurídica.

2. Metodologia

A metodologia adotada neste trabalho foi o *Cross-Industry Standard Process For Data Mining* (CRISP-DM) [Wirth and Hipp 2000], por ser um dos processos mais completos para trabalhos de mineração de dados. O conjunto de dados utilizado é composto por 7.812 decisões judiciais do Tribunal de Justiça do Maranhão, relacionados aos sete primeiros temas do IRDR (exceto o 6), admitidos pelo Núcleo de Gerenciamento de Precedentes (NUGEP)[1]. Esses dados foram anotados por meio de expressões regulares, que buscavam por identificar correspondências com o nome do precedente e com os números oficiais dos temas. Os temas com maior incidência foram o tema 5 (47,67%) e o tema 1 (37,69%), seguidos pelos temas 7 (4,63%), 3 (4,45%), 8 (4,24%), 2 (0,82%) e 4 (0,50%).

Os modelos escolhidos para os experimentos foram o LegalBert-pt FP (ajustado do BERTimbau) e BumbaBert-*small*-SC (treinado do zero). O processamento dos dados foi realizado pelos tokenizadores nativos dos respectivos modelos. Para a classificação, foram utilizados os algoritmos *Random Forest* (RF), *k-Nearest Neighbors* (kNN) e uma Rede Neural (NN) rasa com uma camada latente, instanciados com seus parâmetros padrão. Os classificadores foram alimentados com os *embbedings* dos modelos de linguagem e suas respectivas classes.

Os experimentos foram realizados com diferentes estratégias de truncamento e agrupamento de *tokens*, com e sem a utilização da técnica PEFT para os dois modelos. Para o modelo BumbaBert, as estratégias incluíram agrupamento sem PEFT, truncamento sem PEFT e truncamento com PEFT, nomeadas BB-P-SFT, BB-T-SFT e BB-T-FT, respectivamente. Para o modelo LegalBert-pt, as mesmas estratégias foram aplicadas, nomeadas LB-P-SFT, LB-T-SFT e LB-T-FT, respectivamente. Para o ajuste fino com LoRA, o parâmetro de *rank* foi configurado em 128, com fator de escala α de 8. Além disso, foi aplicado um *dropout* de 10% para lidar com o sobreajuste.

O treinamento dos modelos foi realizado com uma taxa de aprendizagem de 5×10^{-4}, utilizando uma estratégia de ajuste com redução da taxa quando o desempenho do modelo alcançar o *plateau*. O tamanho do lote foi configurado em 16, e o treinamento foi realizado em ponto flutuante de 16 *bits*, com um decaimento de peso de 0,01 para regularização. A métrica F1 do conjunto de avaliação foi utilizada para monitorar o treinamento ao longo de 40 épocas. Para a avaliação final, foram calculadas as métricas de acurácia e *F1-macro*, uma vez que o conjunto de dados é desbalanceado.

[1]https://www.tjma.jus.br/hotsite/nugepnac/item/1992/0/irdr-admitido

3. Resultados e Discussão

A Tabela 1 fornece dados sobre o comparativo do desempenho dos diferentes modelos e classificadores com base nas métricas de acurácia e F1-*macro*.

Tabela 1. Desempenho dos modelos LegalBert-pt e BumbaBert com diferentes classificadores para as métricas de Acurácia (A), F1-*macro* (F), Precisão (P) e *Recall* (R)

Modelo	NN				kNN				RF			
	A	**F**	**P**	**R**	**A**	**F**	**P**	**R**	**A**	**F**	**P**	**R**
BB-P-SFT	0.78	0.56	0.59	0.57	0.80	0.54	0.55	0.53	0.80	0.57	0.62	0.55
BB-T-SFT	0.69	0.45	0.45	0.49	0.80	0.53	0.56	0.52	0.81	0.55	**0.73**	0.54
BB-T-FT	0.81	0.57	0.57	0.56	**0.83**	0.58	0.65	0.56	0.82	0.55	0.57	0.55
LB-P-SFT	0.78	0.49	0.53	0.50	0.78	0.51	0.54	0.50	0.78	0.53	0.64	0.52
LB-T-SFT	0.80	0.52	0.53	0.53	0.81	0.54	0.57	0.53	0.81	0.54	0.57	0.53
LB-T-FT	**0.82**	**0.62**	**0.61**	**0.62**	**0.83**	**0.61**	**0.66**	**0.59**	**0.83**	**0.58**	0.68	**0.56**
Média	0.78	0.54	0.55	0.55	0.81	0.55	0.59	0.54	0.81	0.55	0.64	0.54
Desv. Pad.	±0.05	±0.06	±0.06	±0.05	±0.02	±0.04	±0.05	±0.03	±0.02	±0.02	±0.06	±0.01

Em termos de acurácia, os modelos com PEFT (BB-T-FT e LB-T-FT) apresentam desempenho superior às suas versões não ajustadas, padrão que se assemelha aos resultados de F1-*macro*. A análise da média dos resultados de acurácia para cada classificador revela que tanto o kNN quanto o RF apresentam desempenhos ligeiramente superiores, enquanto o desvio padrão do NN sugere uma maior variabilidade nos resultados.

Sem a utilização de PEFT, o modelo BumbaBert com agrupamento (BB-P-SFT) apresentou desempenhos iguais ou superiores em relação ao modelo com truncamento (BB-T-SFT) em quase todos os classificadores e métricas avaliadas, exceto na acurácia com o RF. Por outro lado, o modelo LegalBert-pt exibiu desempenhos inferiores com agrupamento em comparação com os modelos com truncamento.

Comparando os dois modelos, o BB-P-SFT teve um desempenho superior em relação ao LB-P-SFT em todos os classificadores. Em relação a técnica de truncamento, o LegalBert-pt obteve resultados melhores com PEFT (LB-T-FT) e sem PEFT (LB-T-SFT) em comparação aos modelos BumbaBert (BB-T-FT e BB-T-SFT). Esses resultados indicam que, no contexto deste trabalho, a técnica de truncamento tende a ser mais eficaz para a classificação das sentenças jurídicas conforme os temas do IRDR. É apresentado na Figura 1 o impacto do PEFT para o melhor modelo e classificador pela matriz de confusão.

Percebe-se que o ajuste fino reduziu o impacto das classes com maior incidência no conjunto de dados (temas 1 e 5), reduzindo os erros de classificação dos outros temas nessas instâncias. Com exceção dos temas 2 e 4, que possuem a menor quantidade de amostras, o PEFT aumentou a quantidade de acertos do modelo, especialmente para o tema 3.

4. Considerações Finais

O uso de modelos de linguagem na área jurídica demonstra potencial para aprimorar a eficiência no sistema judicial brasileiro. Assim, este trabalho descreveu o processo de treinamento e avaliação dos modelos de linguagem LegalBert-pt e BumbaBert ajustados por meio da técnica de reparametrização LoRA para o domínio legal, bem como diferentes abordagens de processamento, incluindo o agrupamento e truncamento, com foco na

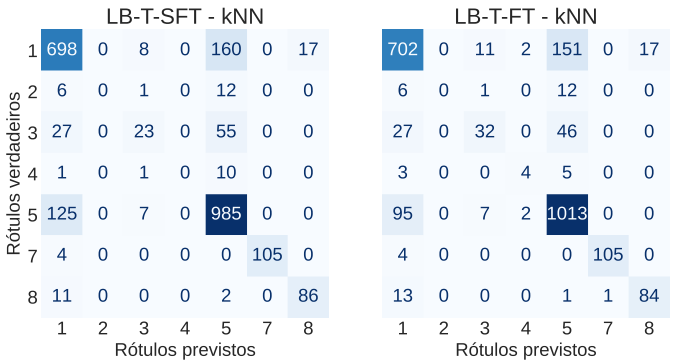

Figura 1. Comparativo sem PEFT e com PEFT para o modelo LegalBert-pt com classificador kNN usando truncamento

classificação de IRDRs em sentenças jurídicas. Os resultados demonstraram que o ajuste fino com PEFT é uma abordagem eficaz para melhorar o desempenho de modelos em tarefas dentro do contexto jurídico.

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

Carmo, F. A. D., Serejo, F., Jacob Junior, A. F. L., Santana, E. E. C., and Lobato, F. M. F. (2023). Embeddings Jurídico: Representações Orientadas à Linguagem Jurídica Brasileira. In *Anais Do XI Workshop de Computação Aplicada Em Governo Eletrônico (WCGE 2023)*, pages 188–199, Brasil. Sociedade Brasileira da Computação.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models.

Melo, J. (2021). Sistema de precedentes garante segurança jurídica e decisões ágeis. https://www.cnj.jus.br/sistema-de-precedentes-garante-seguranca-juridica-e-decisoes-ageis/.

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., Lima, L. M. P. D., Maia, A. C. D. A., and Vicente, R. (2021). LegalNLP - Natural Language Processing methods for the Brazilian Legal Language. In *Anais Do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*, pages 763–774, Brasil. Sociedade Brasileira de Computação.

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 Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, volume 14197, pages 268–282. Springer Nature Switzerland, Cham.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining.