Ferramenta de Identificação Automatizada de Palavras-Chave para Otimização de Triagem de Processos Judiciais

Hévila S. Freitas[1], **Adrielson F. Justino** [1], **Aurelianny A. Cunha**[3], **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)

[3]Laboratório de Inovação do Tribunal de Justiça do Maranhão (Toada Lab)

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

Resumo. *A gestão de processos judiciais no Tribunal de Justiça do Maranhão enfrentava dificuldades devido à triagem e etiquetagem manuais, o que sobrecarregava os servidores e gerava risco de erros. O robô Clóvis foi desenvolvido para automatizar essa etapa, mas a configuração manual das palavras-chave ainda demandava em média 7 dias para cada setor. Para otimizar esse processo, investigou-se a aplicabilidade de uma ferramenta de identificação automática de palavras-chave usando modelagem de tópicos, baseada em Processamento de Linguagem Natural. Com isso, espera-se reduzir o tempo e minimizar os erros comuns na seleção manual, melhorando o fluxo de trabalho na gestão dos processos judiciais.*

1. Introdução

A gestão eficiente de processos judiciais é essencial para garantir maior celeridade e qualidade na prestação jurisdicional. Com a transformação digital, o Poder Judiciário brasileiro implementou a iniciativa "Justiça 4.0", buscando modernizar e integrar seus serviços por meio de tecnologias avançadas, incluindo inteligência artificial [CNJ 2021]. No Tribunal de Justiça do Maranhão (TJMA), a tarefa de triagem e etiquetagem de processos judiciais visa facilitar a classificação e o enquadramento de cada caso, otimizando a organização do fluxo de trabalho no sistema de Processo Judicial Eletrônico (PJe). No entanto, a execução manual dessas atividades era uma tarefa exaustiva e suscetível a erros humanos, comprometendo a eficiência do sistema e a agilidade na tramitação dos processos.

Apesar dos avanços trazidos pelo processo eletrônico, algumas etapas ainda são realizadas manualmente por servidores do TJMA, o que impacta negativamente na celeridade do sistema. Essas limitações no fluxo digital comprometem a eficiência pretendida com a automação do PJe [Mastella 2020]. Para mitigar esse desafio, o sistema RPA (Automação de Processos Robóticos) intitulado de robô Clóvis foi desenvolvido em parceria com o Tribunal de Justiça da Bahia, a fim de automatizar a triagem e etiquetagem de processos no sistema do PJe. No entanto, a identificação manual das palavras-chave necessárias para configurar o robô demanda, em média, 7 dias para processar um conjunto de 15 temas por setor, sendo baseada no conhecimento técnico e na experiência dos profissionais da unidade de justiça e requer refinamento contínuo.

Nesse sentido, torna-se imperativo o uso de técnicas de identificação de palavras-chave para reduzir o tempo e minimizar possíveis erros na configuração do robô. A identificação de palavras-chave é uma técnica de análise de texto que identifica automaticamente as palavras e frases mais relevantes em um documento [Gupta and Vidyapeeth 2017]. Um dos métodos com destaque na literatura é o BERTopic, o qual gera temas coesos e se mantém competitivo em diversos benchmarks, incluindo modelos clássicos e modernos de modelagem de tópicos [Marinato et al. 2024]. Neste contexto, este trabalho visa utilizar modelagem de tópicos para melhorar a identificação e definição automatizada de palavras-chave, contribuindo para aprimorar a gestão de processos judiciais no TJMA.

2. Metodologia

A metodologia usada neste estudo seguiu o modelo CRISP-DM (Cross Industry Standard Process for Data Mining), adaptado para a identificação automatizada de palavras-chave em processos jurídicos [Wirth and Hipp 2000]. O processo metodológico é dividido em seis etapas principais. Primeiramente, no Entendimento do Negócio (i), foram identificadas as necessidades específicas relacionadas à identificação de palavras-chave, definindo os objetivos do projeto e as exigências dos usuários finais. Em seguida, na etapa de Compreensão dos Dados (ii), focou-se na coleta e análise dos documentos jurídicos, organizando as informações para facilitar a modelagem. A Preparação dos Dados (iii) incluiu a seleção, limpeza e tratamento inicial dos dados, utilizando um script com a biblioteca PyMuPDF para extrair e organizar o conteúdo dos PDFs, resultando em um arquivo CSV. Na fase de Modelagem (iv), foram desenvolvidas e testadas funções para a identificação de palavras-chave utilizando a linguagem Python e a biblioteca BERTopic. A escolha do modelo BERTopic foi fundamentada nos resultados de [Marinato et al. 2024], que demonstraram sua superioridade em relação a outras técnicas, como YAKE, TextRank e TF-IDF, na identificação de palavras-chave relevantes em documentos textuais. Dessa forma, o BERTopic foi implementado no pipeline proposto, seguindo as mesmas configurações descritas por [Marinato et al. 2024].

Na etapa de Avaliação (v), foi realizada uma comparação da cobertura dos termos identificados manualmente e pela ferramenta. Embora a frequência dos termos forneça uma medida quantitativa, a avaliação qualitativa das palavras-chave ainda está em andamento. Para isso, está sendo coletado o feedback dos servidores do TJMA por meio de um formulário estruturado, baseado na escala Likert, que avalia a satisfação dos usuários com o pipeline, incluindo aspectos como a organização dos dados, eficácia do pré-processamento, compatibilidade com o Robô Clóvis e sugestões de melhorias. Por fim, para a Implantação (vi) foi feito o processo de documentação e disponibilização da ferramenta para testes. Posteriormente, a ferramenta poderá ser disponibilizada no GitHub vinculado ao Laboratório de Inovação do TJMA (ToadaLab), para facilitar a manutenção e a atualização da ferramenta.

3. Resultados Preliminares

Este projeto resultou no desenvolvimento de um pipeline de identificação de palavras-chave, que gera arquivos de configuração em formato txt, como podemos observar na Figura 1, a estrutura do arquivo foi baseada no modelo do sistema do Robô Clóvis, por questões de privacidade, alguns nomes foram suprimidos.

[Palavras-chave - Tema 1]	[Palavras-chave - Tema 2]
imperatriz	decisao
salario	juizo
margem	indisponibilidade
renda	publico
cosignavel emprestimo	ministerio
efetivos	camara
fepa	estadual
reajuste	improbidade
imposto	
.	

Figura 1. Exemplo do arquivo de configuração com os tópicos gerados, baseado no modelo do Robô Clóvis.

Fonte: Elaborado pelos Autores (2024).

O pipeline desenvolvido consiste em quatro etapas principais: i) Extração de texto de documentos PDF utilizando um script com a biblioteca PyMuPDF, com duração média de duas horas; ii) Pré-processamento dos textos para limpeza e remoção de stopwords, que leva em torno de meia hora; iii) Modelagem de tópicos com a técnica BERTopic, levando cerca de uma hora e meia para a identificação das palavras-chave; e iv) Exportação dos tópicos extraídos para arquivos de configuração. Considerando o tempo médio dessas etapas para um conjunto de 15 temas, testes preliminares demonstraram que, com a implantação da ferramenta, o tempo médio para gerar o arquivo de configuração foi reduzido de aproximadamente uma semana para um dia, resultando em uma economia de tempo de 97,62%.

Exemplos de palavras-chave identificadas por meio da ferramenta incluem, no tema “Saúde”, termos como: urgência, cobertura, hospital, médico, seguro, saúde, indenização, hospitalar, plano e atendimento. Já no tema “Imperatriz”, as palavras-chave identificadas foram: justiça, serviço, adicional, imperatriz, termos, pagamento, civil, pedido, trabalho e processual. Esses exemplos demonstram a diversidade de termos relevantes para cada tema, destacando a capacidade do modelo em capturar palavras-chave relacionadas ao contexto específico de análise. A comparação da taxa de cobertura dos termos identificados manualmente e pela ferramenta pode ser observada na Figura 2.

Os resultados indicam que as palavras identificadas manualmente apresentaram uma taxa de cobertura variável, com melhores resultados em alguns temas e desempenho inferior em outros. Por exemplo, no tema ”Imperatriz”, a identificação manual alcançou uma taxa de cobertura de 88%, enquanto o BERTopic obteve 95%. Para o tema ”Saúde”, a taxa de cobertura manual foi de 64%, enquanto o BERTopic atingiu 96%. Esses resultados evidenciam a superioridade do BERTopic na identificação das palavras-chave, oferecendo maior cobertura nos processos analisados.

4. Considerações Finais

O projeto teve como objetivo principal desenvolver uma ferramenta automatizada para a identificação de palavras-chave relevantes de processos judiciais, visando otimizar a triagem e etiquetagem destes processos, bem como a organização, classificação e acesso aos documentos no TJMA. Com a implantação da ferramenta proposta, buscou-se reduzir o

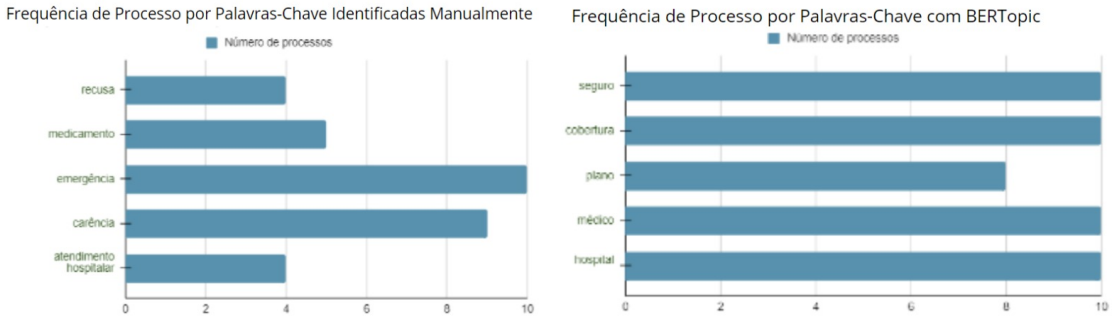

Figura 2. Verificação das palavras-chave selecionadas manualmente do Robô Clóvis e as geradas automaticamente com BERTopic, da etiqueta "Saúde" nos processos.

Fonte: Elaborado pelos Autores (2024).

tempo gasto na seleção manual de palavras-chave, minimizar erros humanos, permitir a realocação de servidores para atividades mais estratégicas e melhorar a adesão dos servidores ao uso do Robô Clóvis. A abordagem baseada na modelagem de tópicos utilizando BERTopic demonstrou ser mais eficiente do que o método tradicional do Robô Clóvis, facilitando a etiquetagem automática e reduzindo significativamente o tempo e o esforço despendidos pelos servidores do TJMA. A implementação deste pipeline de identificação automatizada de palavras-chave não só aprimorará a eficácia do robô, como também agilizará a triagem e etiquetagem de processos judiciais.

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 (2021). Resolução No 370, de 28 de janeiro de 2021: Estabelece a Estratégia Nacional de Tecnologia da Informação e Comunicação do Poder Judiciário (ENTIC-JUD). Acesso em: 15 out. 2024.

Gupta, T. and Vidyapeeth, G. (2017). Keyword extraction: a review. *International Journal of Engineering Applied Sciences and Technology*, 2(4):215–220.

Marinato, M. S., Santana, E. E., and Junior, A. F. L. J. (2024). Análise de métodos de extração de palavras-chave para classificação de documentos jurídicos. *Revista Brasileira de Computação Aplicada*, 16(2):88–96.

Mastella, J. O. (2020). Uma metodologia usando ambientes paralelos para otimização da classificação de textos aplicada a documentos jurídicos. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

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.