Revista Brasileira de Computação Aplicada, Julho, 2024

DOI: 10.5335/rbca.v16i2.15528
Vol. 16, Nº 2, pp. 88–96
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Análise de métodos de extração de palavras-chave para classificação de documentos jurídicos

Analysis of keyword extraction methods for legal document classification

Matheus S. Marinato [1], Ewaldo E. C. Santana [1], Antonio F. L. Jacob Junior [1]

[1]Programa de Pós-graduação em Engenharia da Computação e Sistemas, Universidade Estadual do Maranhão (Uema)

*marinatostm21@gmail.com; ewaldoeder@gmail.com; antoniojunior@professor.uema.br

Recebido: 25/01/2024. **Revisado:** 14/07/2024. **Aceito:** 31/07/2024.

Resumo

Diante do expressivo aumento da demanda judicial e da crescente escassez de recursos na Justiça brasileira, torna-se imperativo investir em novas aplicações tecnológicas. Uma área crítica para análise é a tramitação dos processos, que ainda envolve diversas etapas conduzidas manualmente, incluindo a triagem inicial. Nesse contexto, esta pesquisa visa analisar as técnicas de extração de palavras-chave e verificar qual técnica de extração pode auxiliar na classificação de documentos jurídicos. Para isso, foram utilizados dois bancos de dados, um contendo 86 casos e o outro com 3.543 jurisprudências, ambos pertencentes ao Tribunal de Justiça do Maranhão. Os resultados mostraram que os métodos de extração Bertopic e YAKE obtiveram melhor desempenho para os dados jurídicos. Conclui-se que esses modelos têm o potencial de serem utilizados para proporcionar maior rapidez na distribuição processual.

Palavras-Chave: Extração de palavras-chave; Classificação processual; Eficiência judicial.

Abstract

Given the significant increase in judicial demand and the growing scarcity of resources in the Brazilian legal system, it becomes imperative to invest in new technological applications. A critical area for analysis is the processing of legal cases, which still involves several manually conducted stages, including the initial screening. In this context, this research aims to analyze the techniques of keyword extraction and verify which extraction technique can help in the classification of legal documents. To this end, two databases were used, one containing 86 cases and the other with 3,543 jurisprudence, both belonging to the Judiciary Court of Maranhão. The results showed that the Bertopic and YAKE extraction methods obtained better performance for the legal data. It is concluded that these models have the potential to be used to provide greater speed in the procedural distribution.

Keywords: Keyword Extraction; Procedural Classification; Judicial Efficiency.

1 Introdução

No relatório "Justiça em Números", observa-se um aumento exponencial da demanda judicial, o qual não é acompanhado proporcionalmente pelos recursos necessários para sua resolução, resultando em um aumento no número de processos pendentes na Justiça (CNJ, 2023). Isso revela problemas de superlotação de casos nas unidades judiciais, o que impacta a eficiência no atendimento aos desejos e direitos da sociedade (Hollanda e Leite, 2020). Em termos

quantitativos, há aproximadamente 81,4 milhões de processos em tramitação, com uma média de 31,5 milhões de novos casos anuais no período de 2022.

Segundo o CNJ, não se vislumbra, a curto prazo, uma redução significativa na quantidade de processos em tramitação sem o auxílio de novas tecnologias (CNJ, 2023). Dessa forma, um dos principais focos no poder judiciário é a implementação de novas tecnologias para a resolução de problemas, visando aprimorar a prestação jurisdicional em termos de celeridade e eficiência.

Um desses problemas refere-se ao processo, cujo propósito é resolver uma disputa judicial e, como consequência, promover a paz. Resolvendo conflitos e procedimentos, o processo envolve uma série de atos coordenados para seu desenvolvimento, isto é, determina quais atos precisam ser realizados para que o processo alcance seu termo final, com a decisão judicial (Leal, 2018).

O processo atualmente é conduzido de forma eletrônica, substituindo os processos físicos e, com isso, facilitando o acesso às informações processuais (Mastella, 2020). No entanto, apesar desse avanço, ainda há etapas que são realizadas manualmente, o que impacta negativamente na celeridade do processo e na gestão otimizada dos recursos disponíveis (Mastella, 2020). Por exemplo, muito tempo é perdido no registro de um processo, pois é preenchido manualmente no sistema, podendo conter informações incorretas. Isso resulta em erros na classificação do processo, o que leva a retrabalho. Os servidores públicos responsáveis pelo registro dos processos devem ser devidamente treinados e orientados para que haja uma organização adequada, mesmo com o grande número de processos em andamento (Ramos, 2020).

Outro problema diz respeito ao fato de que, mesmo com o treinamento dos servidores, é comum registrar mais de um assunto no mesmo processo, o que gera erros em sua contabilidade; assim, as cifras apresentadas não refletem o número exato de casos arquivados (Ramos, 2020). Portanto, o agrupamento correto de casos de acordo com o tema pode ajudar a reduzir o tempo gasto na redistribuição, proporcionando ao sistema judicial um tempo de resposta mais longo para a sociedade.

Outra tarefa que pode ser facilitada pelo agrupamento dos processos é a triagem inicial. Nesse estágio, o processo precisa ser direcionado à equipe que trabalha naquele tipo de causa, para o qual vários servidores são alocados no setor jurídico. Eles realizam o agrupamento com o objetivo de identificar a natureza do processo (Leme, 2021). Para concluir a leitura e análise desses documentos, os servidores realizam uma busca superficial de palavras-chave existentes no texto, classificando assim cada arquivo de acordo com sua natureza (Inácio, 2020).

Neste contexto, para auxiliar na melhoria do agrupamento dos processos, propõe-se o uso de técnicas como a extração automática de palavras-chave, que consiste na busca e identificação automática de termos que melhor representam as informações contidas em um documento (Beliga et al., 2015; Subramanian e Karthik., 2017). Portanto, esse tipo de técnica pode proporcionar um agrupamento melhor, pois seu uso pode encontrar termos mais relevantes ou úteis para categorizar o processo. E com isso, quando chega ao local de distribuição, o assunto sendo automaticamente categorizado, sua direção é facilitada, e não é necessário redistribuí-lo devido a erros de classificação, o que proporciona maior rapidez processual (Faraco, 2020).

Neste cenário, esta pesquisa concentrar-se-á nos métodos de extração não supervisionados, os quais podem ser divididos em três grupos: métodos estatísticos, baseados em grafos e baseados em incorporação. Dessa forma, este trabalho propõe a análise de técnicas de extração de palavras-chave, utilizando os métodos mais utilizados de cada tipo, com o objetivo de verificar qual técnica de extração pode auxiliar na classificação de documentos jurídicos.

2 Trabalhos Relacionados

A extração automática de palavras-chave é uma área de pesquisa que parece não ser muito explorada; no entanto, alguns estudos envolvendo a extração de palavras vêm sendo realizados. Por exemplo, em Lu Yao (2019), que utiliza textos de notícias da Sina News Corpus como objeto de pesquisa para métodos de extração de palavras-chave. Neste estudo, o TF-IDF e o TextRank foram combinados para extrair palavras-chave do texto, construindo um modelo de gráfico de palavras, contando a frequência das palavras e a frequência inversa do documento. O desempenho do algoritmo é avaliado por meio de um cálculo próprio de Recall, Precision e F-Measure.

Por outro lado, na pesquisa de Jungiewicz e Łopuszyński (2014), o algoritmo de extração de palavras-chave não supervisionado chamado RAKE é aplicado a um corpus de textos legais poloneses no campo de compras públicas. O método foi avaliado qualitativamente, verificando o grau de relevância das palavras na língua polonesa. Os autores afirmam que o RAKE é essencialmente um método independente de idioma e domínio.

Em Li (2021), a extração de palavras-chave de textos em inglês foi analisada principalmente. Primeiramente, foi utilizado o algoritmo de frequência inversa de frequência de documento (TF - IDF) e o algoritmo de extração de frases-chave (KEA). Em seguida, um algoritmo TF-IDF aprimorado foi desenvolvido para melhorar o desempenho da extração de palavras-chave. Os resultados mostraram que o algoritmo TF-IDF aprimorado teve o tempo de execução mais curto e levou apenas 4,93s para processar 100 textos; A precisão dos algoritmos diminuiu com o aumento do número de palavras-chave extraídas. F1 - score de 60,75%, quando cinco palavras-chave foram extraídas de cada artigo.

Em Moreno et al. (2023), o artigo explora a aplicação da mineração de dados em conteúdos gerados pelos usuários de alojamentos turísticos em plataformas de infomediação e redes sociais. O objetivo é apresentar um algoritmo que permita identificar características do serviço relevantes para a satisfação e confiança dos hóspedes. Na análise, é aplicada uma abordagem de classificação com o BerTopic e Zero-shot para categorizar as avaliações dos hóspedes em rótulos relacionados com a satisfação e confiança dos hóspedes. Em suma, e considerando as implicações práticas, o estudo contribui para o conhecimento sobre a economia compartilhada, fornecendo insights para o desenvolvimento de políticas de marketing e uma melhor compreensão dos serviços de hospitalidade.

No trabalho de Berry e Kogan (2010), é apresentado um método não supervisionado, independente de domínio e linguagem, para extração automática de palavras-chave. RAKE utiliza um conjunto simples de parâmetros de entrada e extrai automaticamente palavras-chave em uma única passagem, tornando-o adequado para uma ampla variedade de documentos e coleções. Para avaliar o desempenho, RAKE é testado em uma coleção de resumos técnicos utilizados nos experimentos de extração de palavras-chave relatados em Hulth (2003) e Mihalcea e Tarau (2004), principalmente com o propósito de permitir uma comparação direta com seus resultados.

Em Campos et al. (2020), é apresentado o algoritmo chamado YAKE, um método leve e não supervisionado de extração automática de palavras-chave que se baseia em recursos estatísticos de texto extraídos de documentos únicos para selecionar as palavras-chave mais relevantes de um texto. O método não precisa ser treinado em um determinado conjunto de documentos, nem depende de dicionários, corpora externos, tamanho do texto, idioma ou domínio. Para demonstrar a eficácia, foi comparado com dez abordagens não supervisionadas de última geração e um método supervisionado. Os resultados realizados demonstram que o YAKE supera significativamente outros métodos não supervisionados em textos de diferentes tamanhos, idiomas e domínios.

No artigo de Piskorski et al. (2021), é apresentado um estudo de algoritmos de extração de palavras-chave não supervisionados e linguisticamente não sofisticados, baseados em abordagens estatísticas, gráficas e baseadas em incorporação. O estudo foi motivado pela necessidade de selecionar a técnica mais apropriada para extrair palavras-chave para indexação de artigos noticiosos em um mecanismo de análise de notícias em larga escala do mundo real.

Em Korenčić et al. (2021), é apresentada uma abordagem de tópico baseada na medição da cobertura do tópico, correspondendo computacionalmente aos tópicos do modelo com um conjunto de tópicos de referência que os modelos devem descobrir. Testes são conduzidos com diferentes tipos de modelos de tópicos em dois domínios de texto distintos nos quais a descoberta de tópicos é de interesse. Os experimentos incluem a avaliação da qualidade do modelo, análise da cobertura de categorias temáticas distintas e análise da relação entre a cobertura e outros métodos de avaliação de modelos temáticos. Os experimentos levaram a descobertas sobre modelos de tópicos e outros métodos de avaliação de modelos de tópico.

O trabalho de Castano et al. (2024) apresenta o ASKE (*Automated System for Knowledge Extraction)*, o qual utiliza uma abordagem para extração de conhecimento no domínio jurídico, combinando modelos de incorporação sensíveis ao contexto e técnicas de aprendizado zero-shot em um ciclo de extração de três fases. É usada uma estrutura de dados baseada em grafos, a qual é continuamente enriquecida com classificação de fragmentos de documentos, novas terminologias e conceitos derivados. Avaliações quantitativas e qualitativas foram realizadas usando o conjunto de dados Eurlex[1] e decisões de jurisprudência italiana. Os resultados mostram que o ASKE geralmente supera o BERTopic e tem desempenho comparável ao Zero-Shot TM, sem precisar predefinir o número de tópicos alvos.

Como pode ser observado, existem trabalhos utilizando diversas técnicas de extração de palavras-chave em diversos tipos de conjuntos de dados. No entanto, os trabalhos citados, além de utilizar algoritmos tradicionais de palavras-chave, apenas comparam uma ou duas técnicas, ao invés de comparar diversas técnicas de extração de palavras. Também se percebe que existe uma lacuna na literatura, pois quase não existem trabalhos de extração aplicados a dados jurídicos. Nesse cenário, este trabalho propõe analisar algumas técnicas de extração de palavras-chave, com o objetivo de verificar qual a técnica de extração que obtém os melhores termos relevantes de documentos jurídicos.

3 Uso de Inteligência Artificial para Classificação de Textos Jurídicos

O interesse no uso de técnicas para a classificação automática de textos jurídicos tem crescido significativamente. Devido ao grande volume de informações textuais no campo do Direito, tem-se explorado o potencial das técnicas de inteligência artificial para otimizar diversas tarefas jurídicas. Um exemplo notável é o estudo de Sousa (2019), que propõe aprimorar a classificação de processos eletrônicos por meio de inteligência artificial. O objetivo é auxiliar tanto os responsáveis pelo cadastro da petição inicial quanto os responsáveis por sua análise, resultando em maior agilidade na tramitação e melhoria na qualidade das informações nos autos judiciais. A pesquisa utilizou processos judiciais e suas petições iniciais do Tribunal de Justiça do Tocantins como amostra. Metodologicamente, diversos algoritmos de aprendizado de máquina foram empregados para a classificação, sendo que o algoritmo de vetor de suporte, *Support Vector Machine - SVM*, obteve os melhores resultados nas métricas utilizadas.

Castro Júnior et al. (2019) apresentam a possibilidade de identificar e unir automaticamente grandes volumes de demandas judiciais em tramitação que possuam o mesmo fato e tese jurídica. O objetivo é criar pendências no Sistema de Processo Eletrônico, informando a ocorrência de conexão entre diferentes unidades judiciais que receberam as causas por distribuição. Isso alerta e facilita a análise pelo Julgador, permitindo que, ao tomar conhecimento das similaridades, ele analise, decida ou julgue conforme seu entendimento.

Para alcançar esse objetivo, foram aplicadas técnicas de Processamento de Linguagem Natural, aprendizado por similaridade e Redes Neurais Artificiais. Essas técnicas culminaram na construção da solução de Inteligência Artificial (IA) chamada Berna, que atualmente está em produção no Poder Judiciário Goiano. Os resultados mostraram uma precisão de 96% nos estudos de casos, demonstrando a efetividade do método. A solução identificou 13 petições iniciais idênticas nas Turmas Recursais, com números de protocolos diferentes na justiça. Além disso, revelou que 8% dos processos em tramitação nos Juizados Especiais Cíveis de Goiânia possuem o mesmo fato gerador e tese jurídica na petição inicial.

[1] https://huggingface.co/datasets/coastalcph/multi_eurlex

No Tribunal de Justiça do Amazonas (Digital, 2021), foi implementada a classificação automática de petições intermediárias, identificando a sugestão do tipo e da categoria da petição. Essa abordagem facilita a rotina dos advogados e agiliza o andamento processual. Ao protocolar uma petição intermediária, um advogado pode encontrar dificuldades em escolher uma das diversas opções de classificação. Numa rotina corriqueira, essa decisão pode consumir um tempo que o profissional não tem, resultando, muitas vezes, na classificação genérica da petição.

A classificação automática ocorre por meio da técnica de Processamento de Linguagem Natural, permitindo que um algoritmo leia o assunto da petição e interprete o seu significado. Em seguida, entra o componente de aprendizado de máquina, que realiza automaticamente a leitura da petição intermediária protocolada pelo advogado. No mesmo instante, são apresentadas quatro sugestões de tipos de petição, baseadas no assunto da mesma. Conforme uma das opções é validada, o algoritmo aprende com essa decisão e passa a oferecer sugestões cada vez mais precisas.

Araujo et al. (2020) apresentam o Dataset VICTOR, um corpus criado a partir de documentos jurídicos digitalizados do Supremo Tribunal Federal, contendo 45 mil recursos, que incluem cerca de 692 mil documentos e cerca de 4,6 milhões de páginas. O VICTOR pode ser utilizado na classificação de tipos de documentos, com seis categorias distintas; atribuição de temas; e problema multi-label com 29 tags diferentes. São realizados experimentos com o dataset empregando diversos métodos de classificação para compreender a natureza sequencial dos processos e implementar melhorias na classificação do tipo de documento.

Apesar de existirem publicações que propuseram o estudo de técnicas de classificação automática de textos em documentos jurídicos, aparentemente ainda não há um volume amplo de publicações para o domínio jurídico, comparado com trabalhos aplicados em outras áreas. Além disso, dentre esses estudos, não há análises de abordagens que demandem menos tempo e custo computacional.

4 Extração Automática de Palavras-chave

A extração de palavras-chave é uma técnica de análise de texto que extrai automaticamente as palavras e frases mais comumente usadas e importantes de um texto (Gupta e Vidyapeeth, 2017). Ajuda a resumir o conteúdo dos textos e reconhece os principais tópicos discutidos. É empregada em diversas áreas, como classificação de texto, agrupamento de texto, rastreamento, detecção de tópicos e resumos (Papagiannopoulou e Tsoumakas, 2020).

Existem métodos de extração de palavras-chave supervisionados e não supervisionados (Siddiqi e Sharan, 2015). Os métodos não supervisionados são os mais comumente aplicados porque não precisam de dados de treinamento rotulados e são independentes (Gupta e Vidyapeeth, 2017). Os métodos supervisionados, por outro lado, têm uma pontuação de precisão muito maior do que a maioria dos métodos não supervisionados. Nesse tipo de método, os dados usados para treinamento já precisam estar anotados, o que acaba sendo um problema, pois na maioria das vezes não há dados rotulados disponíveis, tornando necessária a rotulação manual e, assim, demandando muito tempo para sua aplicação (Siddiqi e Sharan, 2015). Portanto, esta pesquisa utilizará métodos não supervisionados.

Os métodos de extração não supervisionados podem ser divididos em três grupos: métodos estatísticos, baseados em grafos e baseados em incorporação. Métodos estatísticos são os mais simples para identificar as principais palavras-chave ou frases-chave em um texto. As principais abordagens desse tipo são: TF-IDF, que avalia o peso de uma palavra para um documento em uma coleção de documentos (Yao et al., 2019), e YAKE, que utiliza várias características estatísticas para extrair palavras-chave sem a necessidade de um grande conjunto de dados (Campos et al., 2020). Essas abordagens não requerem nenhum dado para extrair as palavras-chave mais importantes em um texto. No entanto, porque esses modelos se baseiam apenas em estatísticas, podem não extrair palavras-chave ou frases relevantes que são mencionadas apenas uma vez, mas ainda podem ser consideradas relevantes (Gupta e Vidyapeeth, 2017).

Nos métodos baseados em grafos, o texto é representado por um grafo, transformando os termos em vértices e as arestas sendo a ligação entre os termos (Siddiqi e Sharan, 2015). A aresta ganha maior peso se esses termos ocorrerem com mais frequência um ao lado do outro no texto. Depois de criar o grafo, os vértices podem ser classificados com base na importância. Um dos métodos mais usados é o TextRank, que classifica o grafo e extrai frases ou palavras-chave relevantes com base no peso dos vértices (Yao et al., 2019). Esse tipo de abordagem não requer nenhum dado rotulado para extrair as palavras-chave mais importantes em um texto.

Métodos de incorporação, por outro lado, exploram a união de documentos para identificar palavras-chave candidatas. O método mais amplamente utilizado é o BERTopic, que é uma técnica de modelagem de tópicos que cria aglomerados densos e de fácil interpretação e mantém as palavras mais importantes no tópico como descrição do mesmo. Ele combina um transformador pré-treinado e algoritmos de agrupamento para identificar tópicos (Grootendorst, 2022; Sleiman et al., 2022). Mais especificamente, o BERTopic gera incorporação de documentos com base em um transformador, agrupa essas incorporações e, finalmente, gera representações de tópicos com o procedimento TF-IDF baseado em classes (Grootendorst, 2022). O BERTopic gera temas coesos e permanece competitivo em uma variedade de benchmarks envolvendo modelos clássicos e aqueles que seguem a abordagem mais recente de agrupamento de modelagem de tópicos (Sleiman et al., 2022).

5 Metodologia

Nesta seção, descrevemos as etapas do framework experimental adotado para analisar os métodos de extração de palavras-chave em documentos jurídicos; esse framework foi baseado nas pesquisas relacionadas. Na implementação e execução dos testes, foi utilizado o ambiente *Google Colab* com a linguagem Python e as bibliotecas de aprendizado de máquina *scikit-learn*, *Natural Language Toolkit* (NLTK) para processamento de linguagem natural e as bibliotecas

específicas de cada método de extração. A Fig. 1 ilustra as etapas do framework experimental.

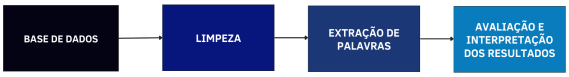

Figura 1: Etapas dos experimentos.

5.1 Base de Dados

Foram utilizadas duas bases de dados. A primeira contendo 86 processos fornecidos pelo Laboratório de Inovação do Tribunal de Justiça do Maranhão. Os processos encontram-se no formato "pdf", sendo composto principalmente por: 1) Autor - Quem move a ação judicial; 2) Registros - são os documentos que compõem o processo: pedido do autor, documentos, resposta do réu, provas, despachos e decisões. Vários tipos de processos e suas palavras-chave foram disponibilizados, a Fig. 2 mostra cada tipo de processo. Pode-se observar que os tipos mais comuns de processos são "*Ação Coletiva*" e "*Saúde*".

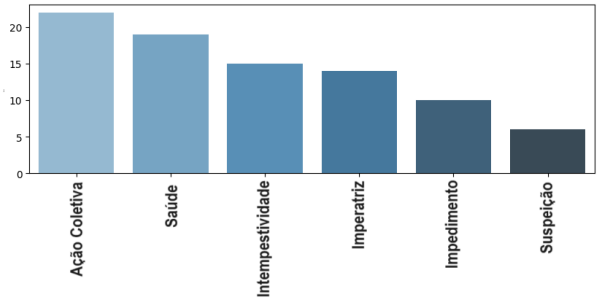

Figura 2: Distribuição dos tipos de processos.

A outra base de dados utilizada é formada por 3543 jurisprudências do Tribunal de Justiça do Estado do Maranhão. A distribuição das jurisprudências pode ser encontrada na Fig. 3.

5.2 Limpeza

Na etapa de limpeza, o primeiro passo na preparação foi ler os arquivos e criar um conjunto de dados com os dados e seus tipos; para cada base de dados, foi criado um conjunto de dados. Com o *corpus* tabulado, passamos então para a limpeza dos dados, que foi baseada em Cirqueira et al. (2018) e da Silva et al. (2023). No presente estudo, após a criação dos conjuntos de dados, todo o conteúdo foi transformado em caixa baixa, para não ocorrer distinção entre palavras iguais. Também foram removidos espaços em branco, pontuações, símbolos financeiros, caracteres especiais e *stop words*. A remoção das *stop words* é uma etapa importante, pois reduz o espaço de busca e melhora significativamente o processo de aprendizado do algoritmo Sousa et al. (2019), e até mesmo a explicabilidade Cirqueira et al. (2020).

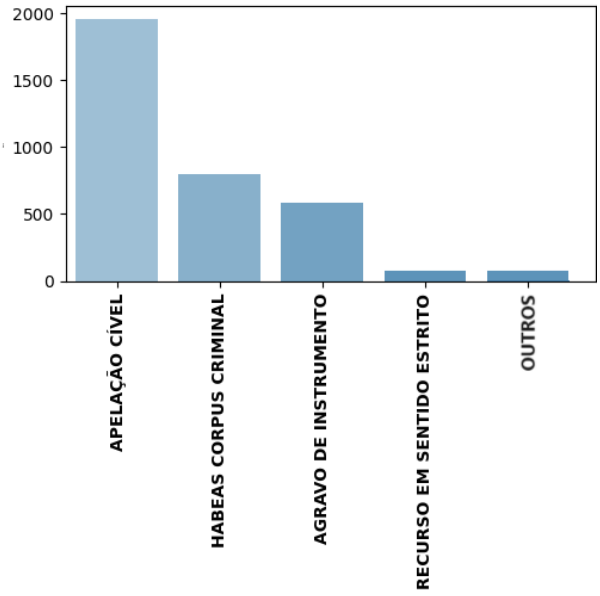

Figura 3: Distribuição dos tipos de jurisprudências.

5.3 Extração de Palavras

Na etapa de extração, foram selecionados 4 métodos de extração, representando cada tipo; esses métodos são: estáticos (YAKE e TF-IDF), baseado em grafos (TEXTRANK) e de incorporação (BerTopic). Os conjuntos de dados serão repassados para cada tipo de método, com a finalidade de gerar palavras-chave para cada tipo de documento jurídico. Os hiperparâmetros que cada método utilizou foram escolhidos empiricamente e estão apresentados na Tabela 1.

Tabela 1: Hiperparâmetros utilizados em cada modelo.

Método	Parametrização
Textrank	*pt_core_news_sm, nlp.max_length* = 13960410
YAKE	*lan* = *language, n* = *max_ngram_size, dedupLim* = *deduplication_threshold, dedupFunc* = *deduplication_algo, windowsSize* = *windowSize, top* = *numOfKeywords, features* = *None*
TF-IDF	*smooth_idf* = *True,use_idf* = *True,norm* =′ *l*2′, *sublinear_tf* = *False*
Bertopic	*verbose* = *True,top_n_words* = 20, *min_topic_size* = 10, *language* = *"portuguese"*

5.4 Avaliação e Interpretação dos Resultados

Nesta etapa, os resultados de cada método serão avaliados e analisados. Inicialmente, baseados no trabalho de Campos et al. (2020), que apresenta um cálculo baseado em uma matriz de confusão. Esse tipo de métrica faz-se

necessário o uso de palavras norteadoras para realização da análise, verificando a relevância das palavras geradas pelos métodos e quais das palavras norteadoras foram encontradas. Neste sentindo, a seguinte nomenclatura será utilizada:

- *TP*: o número de palavras-chave corretamente identificadas como relevantes;
- *TN*: a quantidade de palavras-chave corretamente identificadas como não relevantes;
- *FP*: o número de palavras-chave erroneamente identificadas como relevantes;
- *FN*: a quantidade de palavras-chave erroneamente identificadas como não relevantes.

No caso do dataset dos processos jurídicos, como apresentava um corpus de palavras-chave de cada tipo, essas palavras foram utilizadas como norteadoras. Já no caso do dataset de jurisprudências, como não há palavras-chave, foi analisado as palavras mais específicas em cada jurisprudência para serem utilizadas como palavras norteadoras.

Neste cenário, cada método foi avaliado individualmente e aplicado a cada conjunto de dados. Com os dados obtidos da avaliação, foram calculadas métricas consolidadas para avaliação de classificadores, a saber: precisão (*precision*), *recall* e *F1-score*. Para calcular essas métricas, foram aplicadas as fórmulas apresentadas nos trabalhos de Li (2021); Jacob Junior et al. (2024). A Fig. 4 mostra as fórmulas descritas.

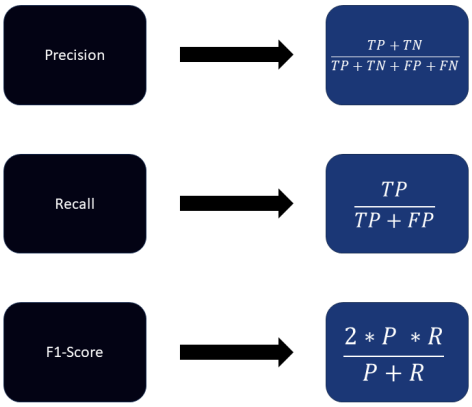

Figura 4: Métricas baseadas no trabalho de Li (2021).

6 Resultados

Os métodos foram aplicados aos dois conjuntos de dados utilizados, os resultados do conjunto de dados dos processos são apresentados na Tabela 2, exibindo uma comparação de cada método de extração nesses dados. Os melhores resultados estão **destacados**. Para fins de melhorar a exibição dos dados, os melhores resultados estão destacados em negrito e as métricas foram identificadas como P (*precision*), R (*recall*) e F1 (*F1-score*).

Conforme pode ser observado, no conjunto de dados dos processos (Tabela 2), o método Bertopic alcança as melhores métricas para quase todos os tipos de processos. Apenas nos casos dos tipos "Impedimento" e "Saúde", o extrator Textrank obteve os melhores resultados, e no caso do tipo "Imperatriz", o YAKE alcançou uma precisão maior.

Comparando os resultados, nota-se que os métodos YAKE e TF-IDF obtiveram resultados próximos aos do Bertopic, demonstrando que, em geral, o Bertopic foi o método que encontrou as palavras-chave mais relevantes para o conjunto de dados, seguido pelos métodos YAKE e TF-IDF.

No caso do conjunto de dados de jurisprudência, os resultados do conjunto de dados são apresentados na Tabela 3, também exibindo uma comparação de cada método de extração nesses dados.

Os resultados da avaliação do desempenho nos conjuntos de dados de jurisprudência oferecem insights valiosos sobre a eficácia dos métodos de extração de palavras-chave. Ao analisar as pontuações de Precision (P), Recall (R) e F1-score (F1), é possível identificar padrões e tendências significativas.

Dentre as jurisprudências com pontuações elevadas, destacam-se aquelas relacionadas à "Revisão Criminal," "Mandado de Segurança Criminal" e "Habeas Corpus Criminal." O método BERTOPIC se sobressaiu nessas categorias, atingindo pontuações notáveis em todas as métricas, como Precision de 95%, Recall de 73% e F1-score de 83% para "Mandado de Segurança Criminal."

Em jurisprudências com pontuações moderadas, como "Ação Rescisória" e "Apelação Cível," BERTOPIC manteve seu desempenho consistente, liderando em Precision, Recall e F1-score. YAKE e TF-IDF também demonstraram resultados competitivos nessas categorias.

Para jurisprudências com pontuações moderadas a baixas, como "Agravo de Instrução," "Remessa Necessária Cível" e "Mandado de Segurança Cível," BERTOPIC continuou a se destacar em todas as métricas. YAKE e Textrank apresentaram desempenhos intermediários, proporcionando resultados equilibrados entre Precision, Recall e F1-score.

Em uma análise geral, BERTOPIC mostrou ser um método robusto e eficiente em diversas jurisprudências, obtendo pontuações mais altas. A escolha entre os métodos pode depender das prioridades específicas de cada contexto de aplicação, considerando a ênfase desejada em Precision, Recall ou F1-score. Por exemplo, BERTOPIC é uma escolha sólida para precisão, enquanto YAKE pode ser preferível para abrangência na recuperação de termos relevantes.

7 Conclusões

Este artigo apresenta uma análise de métodos de extração de palavras-chave para verificar qual é a melhor técnica de extração de palavras que pode auxiliar na classificação de casos nos tribunais de justiça. Os resultados mostraram em ambos os conjuntos de dados que o método de extração BerTopic obteve melhores resultados nas métricas de avaliação, seguido pelo método YAKE.

Conclui-se, portanto, que as técnicas do tipo estatístico (YAKE) e especialmente do tipo de incorporação (Bertopic)

Tabela 2: Resultados obtidos para as medidas de desempenho adotadas no conjunto de dados dos processos.

	YAKE			Textrank			TF-IDF			BERTOPIC		
ETIQUETAS	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Interpestiv.	75%	83%	79%	25%	45%	32%	75%	88%	81%	**80%**	**89%**	**84%**
Impedimento	**70%**	83%	76%	40%	62%	48%	55%	79%	65%	**70%**	**88%**	**78%**
Saúde	**90%**	**90%**	**90%**	65%	81%	72%	85%	77%	81%	**90%**	**90%**	**90%**
Suspensão	45%	**90%**	60%	45%	53%	49%	35%	88%	50%	**55%**	73%	**63%**
Imperatriz	60%	**92%**	73%	**70%**	88%	**78%**	65%	87%	74%	55%	**92%**	69%
Exec. Indiv.	55%	**92%**	69%	55%	69%	61%	55%	58%	56%	**60%**	**92%**	**73%**

Tabela 3: Resultados obtidos para as medidas de desempenho adotadas no conjunto de dados de Jurisprudência.

	YAKE			Textrank			TF-IDF			BERTOPIC		
ETIQUETAS	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Apelação Cív.	45%	**90%**	60%	25%	42%	31%	25%	38%	30%	**90%**	78%	**84%**
Mand. Seg. Cív.	**70%**	**67%**	**68%**	40%	47%	43%	40%	**67%**	50%	40%	50%	44%
Ação Rescisória	**60%**	**80%**	**69%**	45%	75%	56%	55%	58%	56%	**60%**	57%	59%
Agravo de Instr.	50%	56%	53%	60%	57%	59%	30%	55%	39%	**75%**	**83%**	**79%**
Recurso Sent. Estr.	**75%**	**71%**	**73%**	**75%**	60%	**67%**	70%	47%	65%	55%	52%	54%
Mand. Seg. Crim.	65%	54%	59%	60%	63%	62%	60%	52%	56%	**95%**	**73%**	**83%**
Hab. Corp. Crim.	50%	50%	50%	40%	50%	44%	40%	42%	41%	**80%**	**67%**	**73%**
Rem. Nec. Cív.	40%	42%	41%	30%	38%	33%	**50%**	53%	**51%**	35%	**70%**	**47%**
Revisão Crim.	25%	29%	27%	60%	60%	60%	**75%**	56%	**64%**	**95%**	**76%**	**84%**

têm o potencial de ser auxiliares na classificação de documentos de natureza legal, proporcionando assim maior velocidade na distribuição processual. Nesse contexto, esses extratores podem ser adotados como uma ferramenta para sugerir palavras-chave para os Tribunais de Justiça, uma vez que mostraram resultados promissores e há escassez de ferramentas capazes de auxiliar nessa tarefa. O uso desse tipo de técnica, combinado com um método de busca automática pelo tipo de caso, poderia proporcionar maior precisão e velocidade na distribuição de demandas e na tomada de decisões judiciais.

Como trabalhos futuros, pretende-se aprimorar o desempenho da aplicação por meio da inclusão de outras abordagens de extração de palavras-chave e também utilizar métodos de busca para identificar automaticamente o tipo de documento, com base nas palavras-chave encontradas pelo método de extração.

Acknowledgments

Este trabalho foi parcialmente financiado pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq No. 045/2021; e pelo Acordo de Cooperação Técnica No. 02/2021 (Processo No. 38328/2020 - TJ/MA). Gostaríamos também de agradecer aos revisores por suas sugestões, que contribuíram significativamente para a melhoria do trabalho.

Referências

Araujo, P. H. L. d., de Campos, T. E., Braz, F. A. e da Silva, N. C. (2020). Victor: a dataset for brazilian legal documents classification, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1449–1458. Disponível em https://aclanthology.org/2020.lrec-1.181.pdf.

Beliga, S., Meštrović, A. e Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches, *Journal of information and organizational sciences* **39**(1): 1–20. Dispinível em https://jios.foi.hr/index.php/jios/article/view/938.

Berry, M. W. e Kogan, J. (2010). *Text mining: applications and theory*, John Wiley & Sons. https://doi.org/10.1002/9780470689646.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C. e Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features, *Information Sciences* **509**: 257–289. https://doi.org/10.1016/j.ins.2019.09.013.

Castano, S., Ferrara, A., Furiosi, E., Montanelli, S., Picascia, S., Riva, D. e Stefanetti, C. (2024). Enforcing legal information extraction through context-aware techniques: The aske approach, *Computer Law & Security Review* **52**: 105903. https://doi.org/10.1016/j.clsr.2023.105903.

Castro Júnior, A. P. d., Calixto, W. P. e de Castro, C. H. A. (2019). Aplicação da inteligência artificial na identificação de conexões pelo fato e tese jurídica nas petições iniciais e integração com o sistema de processo eletrônico, *CNJ* p. 9.

Cirqueira, D., Almeida, F., Cakir, G., Jacob, A., Lobato, F., Bezbradica, M. e Helfert, M. (2020). Explainable sentiment analysis application for social media crisis management in retail, *4th International Conference on Computer-Human Interaction Research and Applications - Volume 1: WUDESHI-DR*. http://dx.doi.org/10.5220/0010215303190328.

Cirqueira, D., Fontes Pinheiro, M., Jacob, A., Lobato, F. e Santana, Á. (2018). A literature review in preprocessing for sentiment analysis for brazilian portuguese social media, *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. https://doi.org/10.1109/WI.2018.00008.

CNJ (2023). *Conselho Nacional de Justiça. Justiça em números 2023*, CNJ, Brasília.

da Silva, M., Santana, E., Lobato, F. e Jr., A. J. (2023). Preprocessing applied to legal text mining: analysis and evaluation of the main techniques used, *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, SBC, Porto Alegre, RS, Brasil, pp. 1010–1021. https://doi.org/10.5753/eniac.2023.234555.

Digital, J. (2021). Tjam automatiza classificação de petições intermediárias no portal e-saj. Disponível em https://justicadigital.com/tjam-ia-peticionamento/.

Faraco, F. M. (2020). *Modelo de conhecimento baseado em tópicos de acórdãos para suporte à análise de petições iniciais*, Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure, *arXiv preprint arXiv:2203.05794* . https://doi.org/10.48550/arXiv.2203.05794.

Gupta, T. e Vidyapeeth, G. (2017). Keyword extraction: a review, *International Journal of Engineering Applied Sciences and Technology* **2**(4): 215–220. Disponível em https://ijeast.com/papers/215-220,Tesma204,IJEAST.pdf.

Hollanda, Y. R. d. e Leite, F. G. d. F. (2020). Petição Inicial: uma análise à luz de teorias bakhtinianas, *Macabéa-Revista Eletrônica do Netlli* **9**(4): 292–308. https://doi.org/10.47295/mren.v9i4.2677.

Inácio, L. G. V. d. S. (2020). Classificação de documentos jurídicos através de reconhecimento óptico de caracteres e expressões regulares: estudo de caso em uma empresa prestadora de serviços com o uso de uma ferramenta rpa. Instituto Federal de São Paulo, São Paulo, SP, Brasil.

Jacob Junior, A. F. L., do Carmo, F. A., de Santana, A. L., Santana, E. E. C. e Lobato, F. M. F. (2024). Evoimp: Multiple imputation of multi-label classification data with a genetic algorithm, *PLOS ONE* **19**(1): 1–23. https://doi.org/10.1371/journal.pone.0297147.

Jungiewicz, M. e Łopuszyński, M. (2014). Unsupervised keyword extraction from polish legal texts, *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9*, Springer, pp. 65–70. https://doi.org/10.1007/978-3-319-10888-9_7.

Korenčić, D., Ristov, S., Repar, J. e Šnajder, J. (2021). A topic coverage approach to evaluation of topic models, *IEEE Access* **9**: 123280–123312. https://doi.org/10.1109/ACCESS.2021.3109425.

Leal, R. P. (2018). *Teoria geral do processo*, Fórum; 14ª edição (1 janeiro 2018), Belo Horizonte.

Leme, B. (2021). *Classificação automática de documentos de características econômicas para defesa jurídica*, Master's thesis, Universidade de São Paulo, São Paulo, SP, Brasil.

Li, J. (2021). A comparative study of keyword extraction algorithms for english texts, *Lecture Notes in Computer Science*. https://doi.org/10.1515/jisys-2021-0040.

Lu Yao, Zhang Pengzhou, Z. C. (2019). Research on news keyword extraction technology based on tf-idf and textrank, *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. https://doi.org/10.1109/ICIS46139.2019.8940293.

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

Moreno, M. R., Sánchez-Franco, M. J. e Tienda, M. D. l. S. R. (2023). Examining transaction-specific satisfaction and trust in airbnb and hotels. an application of bertopic and zero-shot text classification, *Tourism & Management Studies* **19**(2): 21–37. https://doi.org/10.18089/tms.2023.190202.

Papagiannopoulou, E. e Tsoumakas, G. (2020). A review of keyphrase extraction, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(2): e1339. https://doi.org/10.1002/widm.1339.

Piskorski, J., Stefanovitch, N., Jacquet, G. e Podavini, A. (2021). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up, *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 35–44. Disponível em https://aclanthology.org/2021.hackashop-1.6.pdf.

Ramos, J. D. A. (2020). *Protótipo de um software para a classificação de processos, conforme as tabelas processuais unificadas do conselho nacional de justiça*, Master's thesis, Universidade Federal do Tocantins, Tocantins.

Siddiqi, S. e Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review, *International Journal of Computer Applications* **109**(2). https://doi.org/10.5120/19161-0607.

Sleiman, R., Tran, K.-P. e Thomassey, S. (2022). Natural language processing for fashion trends detection, *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, IEEE, pp. 1–6. https://doi.org/10.1109/ICECET55527.2022.9872832.

Sousa, G. N. d., Almeida, G. R. e Lobato, F. (2019). Social network advertising classification based on content categories, *International Conference on Business Information Systems*, Springer, pp. 396–404. https://doi.org/10.1007/978-3-030-36691-9_33.

Sousa, R. N. d. (2019). Minerjus: Solução de apoio à classificação processual com uso de inteligência artificial.

Subramanian, L. e Karthik., R. (2017). Keyword extraction: A comparative study using graph based model and rake., *Int. J. of Adv.Res.* **39,1**: 1133–1137. https://doi.org/10.21474/IJAR01/3616.

Yao, L., Pengzhou, Z. e Chi, Z. (2019). Research on news keyword extraction technology based on tf-idf and textrank, *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, IEEE, pp. 452–455. https://doi.org/10.1109/ICIS46139.2019.8940293.