UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE ENGENHARIA DA COMPUTAÇÃO

UMA METODOLOGIA USANDO MINERAÇÃO DE TEXTO PARA AGRUPAMENTO DE DOCUMENTOS JURÍDICOS

MATHEUS SERRÃO MARINATO

SÃO LUÍS - MA
2024

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE ENGENHARIA DA COMPUTAÇÃO

MATHEUS SERRÃO MARINATO

UMA METODOLOGIA USANDO MINERAÇÃO DE TEXTO PARA AGRUPAMENTO DE DOCUMENTOS JURÍDICOS

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Antonio Fernando Lavareda Jacob Junior e co-orientação do Prof. Ewaldo Eder Carvalho Santana.

SÃO LUÍS - MA
2024

Marinato, Matheus Serrão.

Uma metodologia usando mineração de texto para agrupamento de documentos jurídicos. / Matheus Serrão Marinato . – São Luís (MA), 2024.

49p.

Dissertação (Mestrado em Engenharia de Computação e Sistemas/PECS) Universidade Estadual do Maranhão - UEMA, 2024.

Orientador: Prof. Dr. Antonio Fernando Lavareda Jacob Junior.

Elaborado por Luciana de Araújo - CRB 13/445

MATHEUS SERRÃO MARINATO

UMA METODOLOGIA USANDO MINERAÇÃO DE TEXTO PARA AGRUPAMENTO DE DOCUMENTOS JURÍDICOS

Dissertação apresentada ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Antonio Fernando Lavareda Jacob Junior e co-orientação do Prof. Ewaldo Eder Carvalho Santana.

Aprovada em: 27/02/2024.

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br ANTONIO FERNANDO LAVAREDA JACOB JUNIO
Data: 23/05/2024 18:08:26-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Me. Antônio Fernando Lavareda Jacob Junior
Orientador - PECS UEMA

Documento assinado digitalmente
gov.br EWALDO EDER CARVALHO SANTANA
Data: 22/05/2024 18:12:26-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Ewaldo Eder Carvalho Santana
Co-orientador- PECS UEMA

Documento assinado digitalmente
gov.br OMAR ANDRES CARMONA CORTES
Data: 22/05/2024 17:40:51-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Omar Andres Carmona Cortes
Membro Interno - PECS IFMA

Documento assinado digitalmente
gov.br RICARDO MARCONDES MARCACINI
Data: 23/05/2024 17:18:58-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Ricardo M Marcacini
Membro Externo - USP

SÃO LUÍS - MA

2024

RESUMO

É expressivo o aumento da demanda judicial e da escassez de recursos, prejudicando o atendimento e a agilidade do sistema jurídico brasileiro. Dessa forma, é evidente a necessidade de se investir em novas aplicações tecnológicas para a garantia do bom andamento processual. Diante disso, vem sendo explorado o potencial das técnicas de mineração de texto em otimizar a realização de muitas tarefas do âmbito jurídico, principalmente voltado para classificação automática. Uma delas é a tramitação do processo, pois nela existem ainda muitas etapas que são realizadas de forma manual, como a determinação da classe do processo dentro do sistema do Processo Judicial Eletrônico (PJE). Diante disso, esta pesquisa tem o objetivo de realizar uma análise de técnicas de extração de palavras chaves que auxilie na classificação de documentos jurídicos, proporcionando maior celeridade processual. 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: Mineração de Texto; Processo; Jurídico; Agrupamento

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	User Experience Bidirectional Encoder Representations from Transformers
CJF	Conselho da Justiça Federal
CNJ	Conselho Nacional de Justiça
CSJT	Conselho Superior da Justiça do Trabalho
IA	Inteligência Artificial
LDA	Alocação de Dirichlet Latente
LSA	Análise Semântica Latente
PJE	Sistema judicial eletrônico
pLSA	Análise Semântica Latente Probabilística
STJ	Superior Tribunal de Justiça
TF-IDF	Frequência do termo–inverso da frequência nos documentos
TJBA	Tribunal de Justiça da Bahia
TJMA	Tribunal de Justiça do Maranhão
TJPE	Tribunal de Justiça de Pernambuco
TPUs	Tabelas Processuais Unificadas
TSE	Tribunal Superior Eleitoral
TOADA LAB	Laboratório de Inovação do Tribunal de Justiça do Maranhão
YAKE	Yet Another Keyword Extractor

SUMÁRIO

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), ocasionando superlotação nas unidades judiciais. Tal fato impacta a eficiência no atendimento aos desejos e direitos da sociedade (Hollanda e Leite, 2020). Em termos quantitativos e aproximados, só em 2022 houveram 81,4 milhões de processos em tramitação, com 31,5 milhões de novos casos.

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, as quais visam 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, determinar quais atos precisam ser realizados para que o mesmo alcance seu termo final, com a decisão judicial (Leal, 2018).

Os processos atualmente são conduzido de forma eletrônica, substituindo os físicos o que facilita o acesso às informações (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, tome-se a triagem inicial. Nesse estágio, vários servidores são alocados no setor jurídico para analisar o processo, com a finalidade de identificar a natureza do processo (Leme, 2021) e com isso agrupando os processos de acordo com o seu tipo, para que estes sejam direcionados à equipe que trabalha naquele tipo de causa.

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). Se, ao chegar neste local, o tipo já estiver automaticamente categorizado, o seu direcionamento fica facilitado não sendo necessário a sua redistribuição por erros de classificação o que dá maior celeridade processual. (Faraco,2020). Assim, para auxiliar na melhoria da triagem dos processos, os tribunais de justiça vem procurando utilizar técnicas de inteligência artificial, que auxiliem na melhoria da tramitação do processo. Como o caso do Tribunal de Justiça do Maranhão, que implementou o robô Clóvis, que trabalha 24 horas por dia, fazendo triagens e etiquetando os processos dentro do sistema do Processo Judicial Eletrônico. O Clóvis realiza a triagem de processos com base na busca de palavras-chave. Para isso, existe um arquivo de texto, em que as palavras-chave são adicionadas. Ao identificar as palavras, ele verifica qual a natureza do processo e etiqueta o processo no PJE. Apesar do robô realizar a triagem corretamente, ainda existem algumas deficiências no processo. Por exemplo, a criação do arquivo de palavras chaves, utilizado pelo robô, atualmente é feita manualmente. Cada palavra chave é selecionada pelos servidores do TJMA, o que ocasiona lentidão na criação do arquivo atrasando a etiquetação dos processos.

Neste sentido, o uso de técnicas como mineração de texto podem auxiliar na melhoria do robô. Inclusive, alguns estudos envolvendo essa técnica vêm sendo realizados no âmbito jurídico. Por exemplo, Almeida Neto et al. (2018) agrupam processos da justiça estadual como procedentes ou improcedentes. Leme (Leme, 2021) classifica o rito do processo, em sumário ou ordinário. Por sua vez, Stemler (Stemler, 2019) desenvolve uma ferramenta empregando modelagem de tópicos para identificação automática de Precedentes Judiciais semelhantes. Em termos de implementação efetiva em ambiente de produção, pode-se citar o Tribunal de Justiça do Amazonas que realiza a classificação de petições intermediárias (Justiça Digital, 2021).

Como se pode observar, aplicações de técnicas de mineração de texto vêm sendo aplicadas nos tribunais de justiça com o objetivo de acelerar a tramitação dos processos. No entanto, os trabalhos utilizam abordagens que requerem tempo e custo computacional. Com isso, percebeu-se uma lacuna, na bibliografia, de estudos que trabalham técnicas mais simples. Como o uso de extração automática de palavras chaves, que consiste na busca e identificação automática de termos, que melhor representam as informações contidas em um documento (Subramanian and Karthik, 2017). Existem métodos de extração de palavras-chave supervisionados e não supervisionados (Subramanian and Karthik, 2017). Essa pesquisa irá focar nos métodos de extração não supervisionados, que podem ser divididos em três grupos:

métodos estatísticos, os baseados em grafos e os baseados em incorporação. O uso desse tipo de técnica pode auxiliar na melhoria do robô Clovis, fazendo com que o seu arquivo de busca não seja criado de forma manual e que na busca do assunto do processo, seja utilizado palavras relevantes e mais precisas para identificar melhor a natureza do processo.

Dessa forma, este trabalho propõe a análise de técnicas de extração de palavras chaves, utilizando os métodos mais utilizados de cada tipo, com o objetivo de verificar qual técnica obtém os termos mais relevantes, auxiliando o robô Clóvis na classificação de documentos jurídicos.

1.1. OBJETIVOS

1.1.1 Objetivo Geral

Realizar análise de técnicas de extração de palavras chaves que auxilie na classificação de documentos jurídicos, com o objetivo de proporcionar celeridade processual.

1.1.2 Objetivos Específicos

- Verificar técnicas para agrupamento dos processos.
- Estudo de técnicas de extração de palavras
- Verificar as técnicas mas atuais e as utilizadas na literatura
- Realizar testes para verificar a eficácia dos métodos de extração para auxiliar na na classificação de processos

1.2. Estrutura do documento

Este trabalho está estruturado da seguinte forma: no capítulo 2 é apresentado o Referencial Teórico; o Capítulo 3 expõe a Metodologia do Trabalho; em seguida, no Capítulo 4 é apresentado os Resultados Obtidos; e por fim, o capítulo 5 apresenta a Conclusão e os trabalhos futuros.

2. REFERENCIAL TEÓRICO

2.1. MINERAÇÃO DE TEXTO

A mineração de texto é um processo semi-automatizado para extração de informações significativas de dados não-estruturados (Tan, 2014). Isto pode ser caracterizado como o processo de analisar o texto para extrair informações que são úteis para um propósito específico. Comparado com o tipo de dados armazenados em bancos de dados, o texto é ambíguo e difícil de processar, no entanto, o texto é a forma mais comum de troca formal de informações (Cohen and Hunter, 2008).

A mineração de texto foi utilizada pela primeira vez durante a década de 1980, tendo o seu estudo voltado para o desenvolvimento de várias técnicas matemáticas, estatísticas, linguísticas e de reconhecimento de padrões que permitem a análise automática de informações não estruturadas, realizando uma extração de alta qualidade e dados relevantes (Aranha and Passos, 2006). Essas informações podem ser extraídas através de quatro etapas: coleta de documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados (Tan, 2014). Na figura 1 pode-se ve essas etapas.

Figura 1 – Etapas da Mineração de Texto (Tan, 2014)

A coleta de documentos visa à aquisição de documentos relacionados ao tipo de informação que pretende-se extrair. Pode-se utilizar de várias fontes, como livros, e-mails etc. Existem técnicas como o Processamento de Linguagem Natural e Recuperação de Informação que podem ser utilizadas nesta etapa (Hassani et al, 2020). Já na etapa de pré-processamento os dados adquiridos na coleta passam por uma "limpeza" e estruturação padronizada, sem perder suas características naturais, para que os algoritmos utilizados sejam capazes de manipular todos os dados da mesma maneira. É nesta etapa também que é definida a forma utilizada para a extração de um pequeno conjunto de dados que representará os termos presentes no conjunto de textos (Pezzini, 2017).

Na etapa de extração do conhecimento, aplica-se os algoritmos de extração automática de conhecimento para buscar informações desconhecidas, mas que possam ser úteis para identificação da informação (Cohen and Hunter, 2008; Pezzini, 2017). Estes algoritmos buscam agrupar termos similares através de uma medida de proximidade, ao mesmo tempo em que tentam formar grupos com características dissimilares entre si. Por último, na etapa de avaliação e interpretação dos resultados, utiliza-se um especialista na área para descobrir se os resultados obtidos são satisfatórios (Hassani et al, 2020).

Nesta pesquisa será abordado técnicas de extração de palavras e modelagem de tópicos.

2.1.1. **Extração de Palavras Chave Automáticas**

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, métodos 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 e YAKE. O TF-IDF atribui um peso a cada palavra ou frase com base na frequência com que ela aparece no texto e na raridade em uma coleção maior de documentos (Yao et al., 2019). O método vetoriza e pontua uma

palavra multiplicando a frequência do termo (TF) da palavra pela frequência inversa do documento (IDF) (Bafna et. al,. 2016). O TF representa o número de vezes que o termo aparece em um documento dividido pelo número total de palavras no documento. Já o IDF reflete a proporção de documentos no corpus que contém o termo. Palavras exclusivas de uma pequena porcentagem de documentos (por exemplo, termos de jargão técnico) recebem valores de importância mais elevados do que palavras comuns a todos os documentos (por exemplo: a, com, de) (Bafna et. al,. 2016). O TF-IDF de um termo é calculado multiplicando as pontuações TF e IDF. À equação 1 apresenta este cálculo.

$$TF - IDF = TF * IDF$$

Equação 1 – Equação do TF-IDF

Já o YAKE utiliza várias características estatísticas para extrair palavras-chave sem a necessidade de um grande conjunto de dados (Campos et al., 2020). Para o método coletar as palavras-chave candidatas, cada palavra-chave candidata receberá então um S (kw) final, de modo que quanto menor a pontuação, mais significativa será a palavra-chave (Campos et. al.,2020). A equação 2 formaliza isso.

$$\mathrm{S(kw)} = \frac{\prod_{w \in kw} S(w)}{\mathrm{TF(kw)} * \left(1 + \sum_{w \in kw} S(w)\right)}$$

Equação 2 – Equação do YAKE

Onde S(kw) é a pontuação de uma palavra-chave candidata com tamanho máximo de 3 termos, determinado pela multiplicação (no numerador) da pontuação da primeira candidata à palavra-chave pelas pontuações subsequentes dos termos restantes, de modo que quanto menor for a multiplicação, mais significativa será a palavra-chave. Isso é dividido pela soma das pontuações S(w) em relação à médias do comprimento da palavra-chave, de modo que os n-gramas mais longos não se beneficiam apenas porque têm um n mais alto. (Campos et. al.,2018) O resultado é ainda dividido pela frequência do termo da palavra-chave. A etapa final na determinação de palavras-chave candidatas adequadas é eliminar palavras semelhantes. Para isso, utiliza-se a distância de Levenshtein, que mede a semelhança entre duas strings. Para as strings consideradas semelhantes (no nosso caso, duas strings são consideradas semelhantes se

sua distância de Levenshtein estiver acima de um determinado limite) será mantida aquela que obtiver a menor pontuação S(kw). Finalmente, o método irá gerar uma lista de potenciais palavras-chave relevantes, de modo que quanto menor for à pontuação (kw) mais importante será a palavra-chave (Campos et. al.,2018).

Essas abordagens não precisam de dados para identificar as palavras-chave mais importantes em um texto. No entanto, como esses modelos dependem apenas de estatística, podem não capturar palavras ou frases mencionadas apenas uma vez, mas que são 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). Este método pode capturar as relações semânticas entre as palavras e identificar frases-chave que consistem em várias palavras. O modelo pode ser expresso como um gráfico direcionado ponderado G = (V,E). O gráfico consiste em um conjunto de pontos V e um conjunto de arestas E, resultando em um subconjunto de V* V . O peso da borda de dois pontos arbitrários i e j são W_{ij}. Para um determinado ponto V_i, é representado os conjuntos de pontos que apontam para ele (InV_i) e que apontam deste para outros ($OutV_i$). $TR(V_i)$ representa a pontuação do ponto V_i obtido pelo modelo TextRank (Li and Zhao,2016). Na equação 3 encontra-se a fórmula do TextRank.

$$TR(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} TR(V_j)$$

Equação 3 – Equaçao ɑo ɪextrank

Nesta fórmula, d é a probabilidade da palavra ser escolhida. Ao usar o algoritmo TextRank para calcular a pontuação dos pontos do gráfico, é necessário especificar o valor inicial de qualquer ponto e então calcular recursivamente a pontuação do mesmo até a convergência. Após cada ponto ser convergente, a pontuação final representa a importância do ponto no gráfico.(Li and Zhao,2016)

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, que produz vetores do mesmo comprimento para cada entrada (uma frase ou um parágrafo). Pode-se alterar a incorporação, podendo treinar um próprio modelo de incorporação para diferentes finalidades. Após isso, agrupa-se essas incorporações e, finalmente, gera-se representações de tópicos com o procedimento TF-IDF baseado em classes (Grootendorst, 2022). Na equação 4 temos o cálculo do TF- IDF baseado em classes.

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t})$$

Equação 4 – Equação do Bertopic

O TF-IDF baseado em classe altera a entrada de um documento para um cluster, combinando todos os documentos no mesmo cluster como entrada de texto longo e calcula a pontuação do TF-IDF. Uma modificação no IDF dividindo o número médio de palavras por classe (A) pela frequência do termo t em todas as classes (Sleiman et al., 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).

2.1.2. **Trabalhos Correlatos de Extração de Palavras**

A extração automática de palavras-chave é uma área de pesquisa não muito explorada, mas observa-se que alguns estudos sobre o tópico vêm sendo realizados. Por exemplo, Lu Yao (Lu Yao, 2019) utiliza textos de notícias da Sina News Corpus para pesquisar 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 sua vez, Jungiewicz e Łopuszyński (2021) desenvolveram um algoritmo de extração de palavras-chave não supervisionado chamado RAKE, o qual 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. 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.

Moreno et al. (2023), 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.

Berry e Kogan (Berry and Kogan, 2010) apresentam um método, chamado RAKE, não supervisionado, independente de domínio e linguagem, para extração automática de palavras-chaves. 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.

Já Campos et. al.(2020) apresenta 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 trabalho de Piskorski et al. (2021) é apresentado um estudo de algoritmos de extração de palavras-chave não supervisionados e linguísticamente 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.

No Trabalhos de Campos et. al. (2018) é proposto uma abordagem para extração e classificação de palavras-chave chamada YAKE, que seleciona as palavras-chave mais importantes de um único documento. Foi comparado os métodos RAKE, TextRank e a linha de base TF.IDF, em quatro coleções diferentes para ilustrar a abordagem proposta. Os resultados experimentais sugeriram que a extração de palavras-chave de documentos utilizando o método YAKE, resulta em uma eficácia superior quando comparada a abordagens semelhantes.

Já 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 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.

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.

2.1.3. **Modelagem de Tópicos**

A modelagem de tópicos é uma técnica de mineração de texto não supervisionada capaz de utilizar um conjunto de documentos, para detectar padrões de palavras ou frases e agrupar automaticamente (Souza, 2019). Cada padrão encontrado vira um tópico, que é constituído por um conjunto de palavras importantes, extraídas automaticamente dos documentos contidos em um ou mais conjunto de documentos (corpus), criando-se grupos de palavras e expressões semelhantes que melhor caracterizam um conjunto de documentos (Vayansky and Kumar, 2020).

Os algoritmos de modelagem não possuem informações sobre os assuntos contidos nos documentos, como: títulos, palavras-chave ou áreas de concentração . Desta forma, se torna necessário a análise dos tópicos para verificar o seu grau de relevância, uma vez que não existe garantia da interpretação dos resultados (Kherwa and Bansal, 2019). Alguns algoritmos usados para tarefas de Modelagem de Tópicos são: Alocação de Dirichlet Latente (LDA), Análise Semântica Latente (LSA) e Análise Semântica Latente Probabilística (pLSA).

O LDA é um modelo estatístico e gráfico usado para obter relacionamentos entre vários documentos em um corpus (Jelodar et al., 2019). Este modelo segue o conceito de que cada documento pode ser descrito pela distribuição probabilística de tópicos e cada tópico pode ser descrito pela distribuição probabilística de palavras. Assim, pode-se ter uma visão muito mais clara de como os tópicos estão conectados (Guo et al., 2021). Já o LSA baseia-se na hipótese distributiva, que afirma que a semântica das palavras pode ser apreendida observando os contextos em que as palavras aparecem (Dumais, 2004). Calculada a frequência com que as

palavras ocorrem nos documentos, assume-se que documentos semelhantes conterão aproximadamente a mesma distribuição de frequências de palavras para determinadas palavras (Sidorov, 2019).

O pLSA é uma variante do LSA utilizado para modelar informações sob uma estrutura probabilística, nos quais os tópicos são tratados como variáveis latentes ou ocultas (Gaussier and Goutt, 2005). Ele analisa uma coleção de documentos de texto (um corpus), em seguida tenta identificar esses tópicos (em termos de palavras) e sua importância relativa para cada tópico, fornecendo um número pré-especificado de tópicos em cada documento (Ba, 2019). A modelagem de tópicos demonstra ser uma abordagem mais relevante do que outras técnicas, como à clusterização, pois fornece tópicos mais específicos e importantes, proporcionando tópicos que descrevem detalhadamente o documento. (Yuan et,. al, 2021)

2.2. PODER JUDICIÁRIO

2.2.1. **Tramitação do Processo**

De forma geral, um processo jurídico é um instrumento legal para eliminar conflitos entre sujeitos através da aplicação da lei. Ele se inicia com um pedido do autor (seja pessoa física ou jurídica) e busca dirimir o conflito através de entendimentos dos fatos apresentados no processo (Leal, 2005). Após o conflito ser reconhecido, pode haver a necessidade de determinar a execução ou obrigação pela parte que foi reconhecida como a responsável por reparar esta situação (Leal,2005). A figura 2 ilustra as etapas do processo.

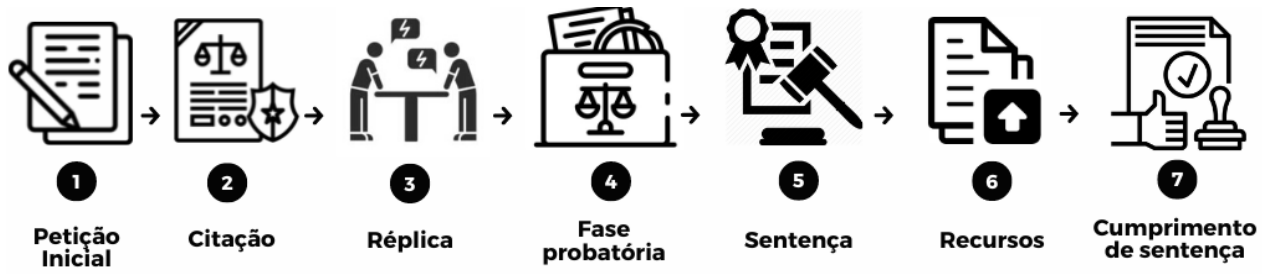


Figura 2 – Etapas da Tramitação do Processo. (Leal,2005; Camargo and Cardoso,2015)

O processo se inicia a partir do protocolo de uma petição ao juiz de primeira instância de primeiro grau. Na petição deve constar o pedido do processo, expondo todos os fatos

ocorridos de forma detalhada, trazendo os motivos pelos quais o autor está ajuizando a ação e quais dos seus direitos estão sendo prejudicados.

Caso a petição inicial esteja de acordo, o processo vai para a fase de citação, na qual o réu é informado por meio de um mandado para comparecimento em audiência de conciliação. Se as partes não entrarem em acordo, o acusado deve manifestar seus argumentos a partir de um documento denominado "contestação", no qual questiona ou desmente os fatos apresentados e formula um pedido a seu favor (Camargo and Cardoso, 2015). Com a contestação, entra-se na fase de réplica, onde, após a defesa do acusado, o autor do processo contrapõe os argumentos da defesa na contestação, apresentando novos pontos a seu favor. Desta forma entra-se na fase probatória, na qual a parte acusatória apresenta provas concretas (documentos, áudios, fotos e vídeos) e testemunhas para depor. Caso seja necessário, pode ser feita uma perícia do caso. Por sua vez, o réu apenas precisa se defender e invalidar as provas contra si (Camargo and Cardoso,2015).

Após a apresentação dos fatos, entra-se na fase de sentença, no qual o juiz responsável toma uma decisão, através de todas as provas e depoimentos. Além de sentenciar o resultado do processo, o magistrado também incube a parte perdedora da ação de pagamento de taxas e honorários com advogados. A parte perdedora pode entrar na fase de recurso, recorrendo à decisão e apresentando um recurso de apelação que deve ser julgado por desembargadores de um tribunal. Caso a decisão seja mantida, a parte perdedora ainda pode direcionar os recursos, primeiro, ao Supremo Tribunal Federal e, depois, se necessário, ao Superior Tribunal de Justiça, sendo possível reverter a decisão de forma integral ou parcial (Camargo and Cardoso, 2015).

Assim que todos os recursos forem esgotados, a sentença deve ser colocada em prática, sendo que o processo só é encerrado após o cumprimento da decisão judicial. Em casos que se tem indenização, o lado vencedor deve apresentar os cálculos ao devedor.

2.2.2. **Classificação do Processo**

Atualmente, o processo é tramitado de forma eletrônica, através do sistema judicial eletrônico (PJe), que atende às necessidades dos diversos segmentos do Poder Judiciário brasileiro (Justiça Militar da União e dos Estados, Justiça do Trabalho e Justiça Comum, Federal e Estadual). O principal objetivo do PJe é tornar virtual todos os processos que tramitam no Poder Judiciário, simplificando todas as fases de cada ação em tramitação e com isso reduzindo

as atribuições aos servidores, tendo como resultado a celeridade processual. (Falcão et. al. , 2018)

Para fazer uso do PJe é necessário que se obtenha o certificado digital. Somente com essa ferramenta é possível transmitir dados e informações de forma segura, sem o perigo de violação. O PJe mantém o trâmite processual, mas todas as petições, decisões e demais documentos que integram o processo são disponibilizados em meio eletrônico, sem a utilização de papel, sendo os dados mantidos virtualmente. (Falcão et. al. , 2018)

Apesar do poder Judiciário utilizar o processo eletrônico (PJe) para se ter celeridade processual, ainda existem muitas etapas que são realizadas de forma manual, uma delas é na determinação da classe processual. Na abertura do processo é informado manualmente pelo advogado os dados iniciais no PJE, inserindo o assunto, a classe processual, a petição inicial e outros documentos necessários (Manual do Advogado, 2022). É através da classe processual informada que o processo irá tramitar de acordo com a natureza da petição inicial (Mastella, 2020). A classe determina qual procedimento judicial ou administrativo é mais adequado ao pedido.

Antigamente não havia uma padronização das classes, fazendo com que cada instância ou esfera tivesse a sua nomenclatura. Isso fazia que um mesmo processo, ao tramitar de uma instância para outra, fosse classificado de forma diferente. Para sanar esse problema o Conselho Nacional de Justiça (CNJ), a fim de padronizar as classes, os assuntos e os movimento processuais, criou, através da Resolução CNJ n. 46, de 18 de dezembro de 2007, as Tabelas Processuais Unificadas - TPUs.

Apesar disso, ainda é necessário que ocorra a correta interpretação das TPUs para, dependendo do entendimento do processo, classificá-lo de forma acertada. Se, no ato da criação do processo, o advogado classificá-lo de forma errônea, ocasionará a distribuição do mesmo para um órgão que não trata da natureza daquele processo, gerando um retrabalho, pois o mesmo deverá ser enviado a uma central de reclassificação objetivando sua tramitação para o órgão competente. Na figura 3, tem-se uma ilustração dessa fase inicial.

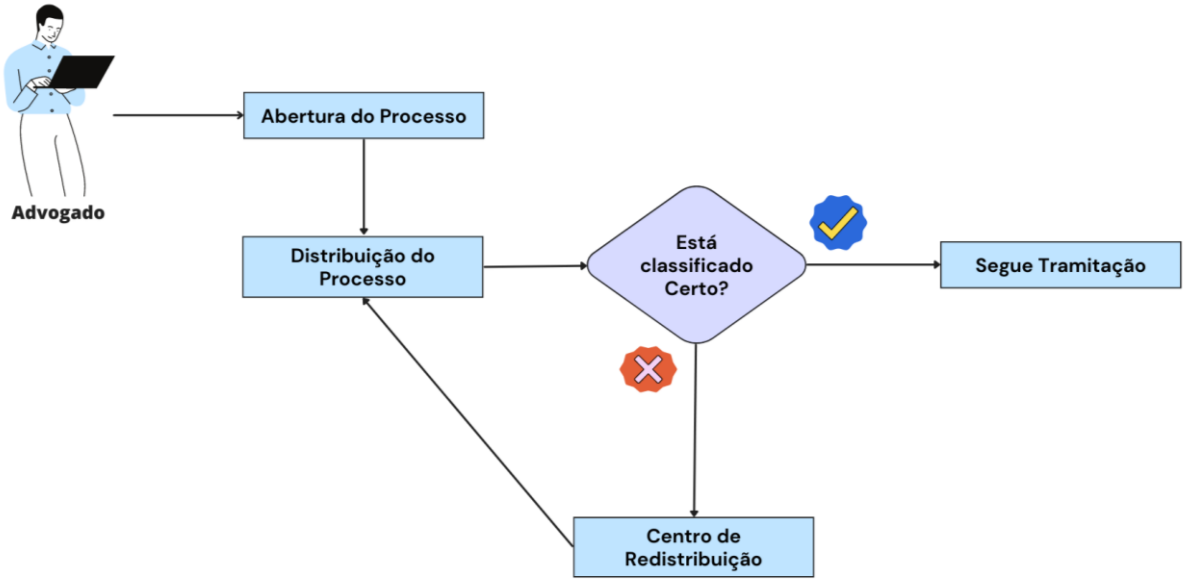


Figura 3 – Etapas da criação do Processo. Adaptada de Mastella (2020).

Muitas vezes, passa-se muito tempo antes que o processo comece a andar, o que faz com que seja importante classificá-lo automaticamente para evitar erros de preenchimento pelos advogados. Além disso, ao automatizar esse processo, também poderíamos reduzir o tempo necessário para o processo avançar

2.2.3. **Justiça 4.0**

O Programa Justiça 4.0 tem como objetivo tornar os serviços do sistema judiciário brasileiro mais eficientes e acessíveis à sociedade, através do uso de novas tecnologias e inteligência artificial, promovendo soluções que automatizam as atividades dos tribunais, garantindo mais produtividade, celeridade, governança e transparência dos processos. (Araújo, 2022)

O Programa foi criado no auge da pandemia da Covid-19 em 2020 como resultado da parceria entre o Programa das Nações Unidas para o Desenvolvimento (Pnud), o Conselho Nacional de Justiça (CNJ) e o Conselho da Justiça Federal (CJF), contando com o apoio do Tribunal Superior Eleitoral (TSE), do Superior Tribunal de Justiça (STJ) e do Conselho Superior da Justiça do Trabalho (CSJT). Atualmente o Justiça 4.0 tem adesão dos conselhos, tribunais superiores, tribunais federais e trabalhistas do país. Quase todos os tribunais estaduais já aderiram. No caso da Justiça Eleitoral, mais da metade dos tribunais integra a iniciativa e a

adesão nos tribunais militares atingiu um terço do total (Araujo, 2022). O Programa atua em quatro eixos: Soluções para transformar o Judiciário e melhorar a prestação de serviços a toda a sociedade; Gestão de informações e políticas judiciárias; Formulação, implantação e monitoramento de políticas judiciárias com base em evidências; Fortalecimento de capacidades institucionais, com base na transferência de conhecimento. (Araújo, 2022)

Desde o início do Programa Justiça 4.0, o CNJ vem firmando Termos de Cooperação com os diversos segmentos de tribunais com o objetivo de desenvolver o uso colaborativo dos produtos, projetos e serviços criados pelo programa. O programa realiza a transferência integral dos conhecimentos e das soluções aos tribunais parceiros, auxiliando na implantação e na criação de estratégias de sustentabilidade (Araújo, 2022).

2.3. 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.

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.

No trabalho de Melo et. Al.(2021) é apresentado o projeto Toth que tem como finalidade facilitar o uso interno, no sistema PJe, auxiliando na recomendação de classes e assuntos de processos do Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT). No momento em que é enviado o arquivo da petição, o PJE aciona o Toth que realiza mineração no texto

protocolado, gerando a classe e o assunto do processo. Essas informações geradas são armazenadas na base de dados do Toth.

A recomendação ocorre quando o analista processual vai avaliar o processo, pois quando ele acessar a petição inicial, o PJe enviará uma requisição ao Toth. Caso exista recomendação de classe e assuntos em sua base de dados, o Toth responderá ao PJe informando quais são as suas sugestões para o processo em questão. Essa recomendação é baseada no treinamento supervisionado de algoritmos de classificação, que utilizou como base assuntos e classes informadas pelos usuários do sistema PJe. Para comprovar sua eficácia, foi realizado um teste em uma vara piloto. O Toth foi capaz de realizar predições de classe e assuntos para a vara em 67% e 64% dos casos analisados. Suas recomendações foram compatíveis às classificações realizadas pelo analista processual em 81% dos casos de classes e 77% dos casos de assuntos.

O Toth está em fase de expansão, analisando dados vindos de 80 varas de diversas competências distintas, tendo como resultado 81% dos casos da escolha da classe igual ao escolhido pelo advogado e em 78% dos casos igual ao do analista processual do PJe. Demonstrando que o projeto não só facilita o uso do PJe por meio da sua recomendação, mas também auxilia na melhoria da qualidade dos dados que são enviados ao CNJ via DataJud. (Melo et. Al., 2021).

Araújo et al. (2020) apresenta 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.

3. METODOLOGIA

Esta seção apresenta qual é a metodologia deste trabalho. Inicialmente, solicitou-se ao Tribunal de Justiça do Maranhão (TJMA) a disponibilização dos códigos dos projetos Clóvis (antigo Triador) e ELIS para verificação de como o agrupamento dos processos estava sendo realizado e a partir daí sugerir-se melhorias.

3.1. PROJETOS DE IA NO TJMA

3.1.1. **Robô ELIS**

Desenvolvido pela equipe do Tribunal de Justiça de Pernambuco (TJPE), o robô ELIS surgiu para auxiliar o trabalho dos servidores e dos magistrados da Vara de Executivos Fiscais. Sendo desenvolvido na linguagem PYTHON, ele foi criado em um prazo de 60 dias, para atender demandas específicas de uma Vara do TJPE, levando-se em consideração os principais gargalos encontrados no transcurso dos processos que tramitam no juízo. (Brito, 2020).A figura 4 mostra as etapas aplicadas pelo ELIS.

Figura 4 – Estrutura do ELIS.

Ele foi programado para proceder a triagem inicial dos processos, conferindo a petição inicial, realizando a triagem dos documentos acostados(documentos anexados ou adicionados a um processo judicial.), conseguindo classificação de 69.351 processos em apenas 15 dias, tendo uma acurácia de 96% de acerto no classificador de prescrição; 94% de acerto no classificador de dados cadastrais divergentes; 98% na classificação de CDA com erro; e 99% nas classificações de incompetência do Juízo. O robô também faz análise de prescrição,

competência, possíveis erros na Certidão de Dívida Ativa e divergências de dados cadastrais, inserindo minutas no sistema e até mesmo assinar despachos, se determinado pelo magistrado. (Brito, 2020)

O robô Elis foi adaptado para o TJMA, com a finalidade de ser utilizado nas varas de execução fiscal do Poder Judiciário do Maranhão. A iniciativa partiu do TOADA LAB (Laboratório de Inovação do Tribunal de Justiça do Maranhão), que após contato instituiu um grupo de trabalho que adaptou o robô para o contexto da Justiça maranhense (Limeira, 2021).

3.1.2. **Robô Clovis**

O robô Clóvis (antigo robô Triador) é desenvolvido na linguagem JAVA, criado através da cooperação técnica entre o Tribunal de Justiça da Bahia (TJBA) e o Tribunal de Justiça do Maranhão (TJMA), utilizando técnicas de automação e sendo executado de forma independente (Limeira, 2022). Ele realiza triagens e etiqueta processos dentro do sistema do Processo Judicial Eletrônico no Judiciário maranhense. Antes do seu uso, os processos eletrônicos eram separados, um por um, para serem identificados e etiquetados no sistema PJE (Limeira, 2022).

O Clóvis faz a mesma tarefa repetitiva, otimizando o trabalho e o tempo, permitindo que atividades mais relevantes sejam executadas e, também, contribui para prevenir doenças funcionais como lesão por esforço repetitivo (LER). Na figura 5 vê-se as etapas de funcionamento do Clóvis. (Limeira, 2022)

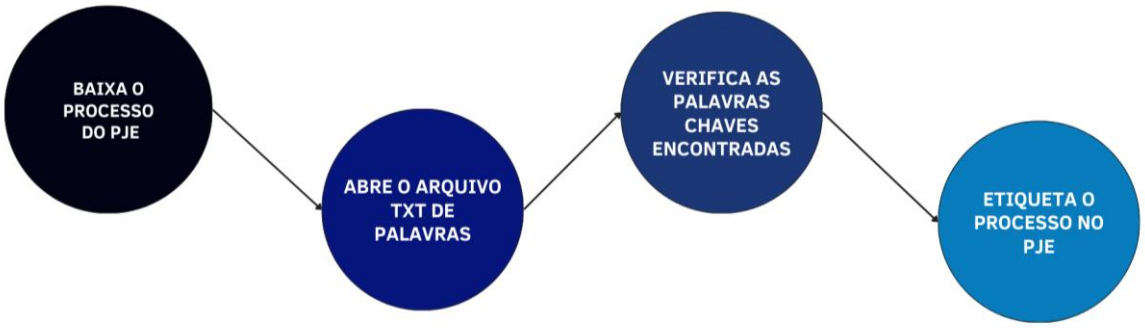


Figura 5 – Estrutura do Clóvis (Triador).

O robô Clóvis realiza a triagem de processos aplicando uma busca de palavras-chaves que são pré-definidas anteriormente pelo usuário. Para isso, existe um arquivo do tipo texto, o

qual o robô acessa, fazendo uma análise entre as palavras chaves desse arquivo e os documentos do processo que ele pode vir a etiquetar. Ao identificar estas palavras, ele aplica a etiqueta do assunto procurado. Ele analisa um processo completo em apenas 30 segundos, em que as palavras-chave são adicionadas. Ao agrupar processos de acordo com os temas inseridos, é possibilitado a realização do julgamento por matéria proporcionando maior celeridade para a prestação jurisdicional. (Limeira, 2022)

3.1.3. **Análise do Robô Clovis**

Durante o processo de análise e estudo do funcionamento do robô, surgiram diversas indagações que demandaram a realização de reuniões com membros do TOADA LAB para esclarecê-las. Durante esses encontros, identificaram-se deficiências no projeto, as quais apresentaram potencial para melhorias. Uma dessas lacunas estava associada à criação do arquivo de palavras-chave utilizado pelo robô para categorizar os processos. Notou-se que estas palavras são determinadas e inseridas manualmente pelos servidores do TOADA, o que resulta em considerável demanda de tempo no desenvolvimento do arquivo e consequente atraso no agrupamento dos processos. Além disso, observou-se que o robô não possuía capacidade para distinguir palavras flexionadas, requerendo, assim, a inclusão de múltiplas formas da mesma palavra no arquivo, a fim de que fosse detectada corretamente nos processos. Diante dessas constatações, surgiu a necessidade de desenvolver um pipeline composto pela implementação de um método de extração automática de palavras visando aprimorar a capacidade de busca do Clóvis. Os detalhes desse procedimento estão descritos a seguir.

3.3. EXTRAÇÃO AUTOMÁTICA DE PALAVRAS CHAVES

Nesta fase, avalia-se o impacto da aplicação de técnicas de extração de palavras-chave e geração automatizada de termos para cada etiqueta, com o objetivo de determinar se o emprego de abordagens automáticas para identificação de palavras-chaves pode resultar na identificação de termos mais pertinentes, os quais representariam de forma mais eficaz cada etiqueta, contribuindo, consequentemente, para aprimorar o processo de busca e agrupamento dos processos judiciais. Na figura 6 vê-se, de forma visual, as etapas propostas a serem realizadas neste estudo.

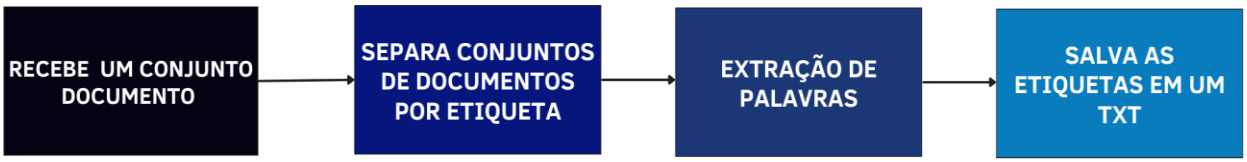


Figura 6 – Etapas da Extração de palavras

Primeiramente, um conjunto de documentos previamente etiquetados é apresentado, com cada etiqueta representando uma categoria. Esses documentos são submetidos à leitura e subsequentemente armazenados em um dataframe, acompanhados de suas respectivas etiquetas. Posteriormente, procede-se à aplicação de técnicas de pré-processamento, visando a eliminação de termos irrelevantes. Com tal refinamento, o método selecionado para extração das palavras-chave de cada etiqueta é aplicado. Por fim, o resultado é registrado em um arquivo de texto, contendo as palavras-chave identificadas e suas respectivas etiquetas correspondentes.

4. RESULTADOS

4.1 ANÁLISE E VERIFICAÇÃO DOS DADOS

Para análise e testes do método proposto, inicialmente foi solicitado ao TOADA, processos já agrupados e o arquivo com as palavras chaves (por tema) utilizado pelo Clóvis. Foi disponibilizado um dataset contendo 253 processos extraídos manualmente pertencentes às etiquetas:

- Saúde;
- Ação Coletiva 14.440;
- Ação Coletiva 6542;
- Impedimento tipo 1;
- Impedimento tipo 2;
- Impedimento tipo 3;
- Impedimento tipo 4;
- Imperatriz;
- Improbidade Administrativa;

- Insalubridade;
- Intempestividade;
- João Lisboa;
- Magistério;
- Pasep;
- Posse e Propriedade;
- Promoção e Remoção PM;
- Revisão Contratual;
- Serasa;
- Suspeição tipo 1;
- Suspeição tipo 2;
- Tarifa Bancária.

Cada processo estava disponível no formato PDF e compreendia principalmente os seguintes elementos: 1) Autor - a parte que instaura a ação judicial; 2) Registros - os documentos que constituem o processo, incluindo a petição inicial do autor, documentos anexados, resposta do réu, evidências apresentadas, despachos judiciais e decisões proferidas.

Conjuntamente aos processos disponibilizados, foi requisitado ao TOADA, o arquivo TXT contendo as palavras-chave empregadas pelo Clóvis para categorizar esses processos. Uma análise desse arquivo revelou a ausência de palavras-chave para diversos tipos de processos disponibilizados.

Diante disso, tornou-se imperativo restringir a utilização apenas aos processos que dispunham de etiquetas associadas, a fim de estabelecer uma ambiente de teste congruente com o aplicado pelo Clóvis em condições reais.

Neste escopo, foram obtidos 130 processos, distribuídos entre as seguintes categorias: Saúde; Ação Coletiva; Imperatriz; Intempestividade; Impedimento tipo 3 e Suspeição tipo 2. Na Figura 7 vê-se a distribuição de frequências desses processos. Observa-se que as categorias mais frequentes são “Ação Coletiva” e “Saúde”.

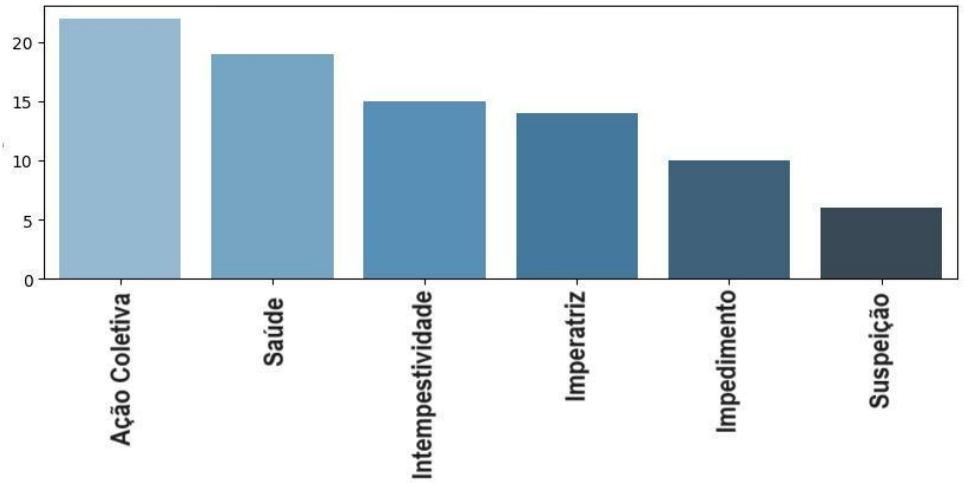


Figura 7 – Distribuição dos tipos de Processos

Com os processos selecionados, desenvolveu-se um código no ambiente *Google Colab* utilizando a linguagem de programação Python para realizar a extração das palavras-chave. Este experimento teve como propósito investigar e analisar as palavras-chave presentes nos documentos, avaliando sua pertinência às respectivas categorias.

Inicialmente, o experimento realizava uma leitura do arquivo com as palavras chaves aplicadas pelo robô Clóvis e as salvava em um dicionário. Após isto, era feita uma busca simples de contagem de palavras em cada documento, para identificar a qual etiqueta correspondia. O código verificava qual tipo de etiqueta possuía o maior número de palavras chaves no documento e com isso informava a etiqueta à qual o processo pertencia. A tabela 1 mostra um exemplo de alguns dos resultados dessa análise.

Tabela 1 Resultado da busca de palavras alguns dos processos

Processo	**Etiqueta Fornecida**	**Etiqueta proposta pelo Método**	**Palavras**
Processo 1	Saúde	Saúde	● Medicamento ● Urgência ● Saúde
Processo 2	Saúde	Imperatriz	● Pagamento ● Imperatriz
Processo 3	Saúde	Saúde	● Atendimento médico ● Medicamento

			●Urgência ●Saúde
Processo 4	Saúde	Imperatriz	●Pagamento ●Adicional
Processo 5	Saúde	Consignado	●Contrato
Processo 6	Saúde	Imperatriz	●Pagamento ●Adicional ●Servidor
Processo 7	Saúde	Imperatriz	●Pagamento ●Imperatriz
Processo 8	Intempestividade	Tema1142	● execução ●cumprimento de sentença ● execuções ● honorários advocatícios
Processo 9	Saúde	Saúde	●Recusa ●Atendimento hospitalar ●Atendimento médico ●Medicamento ●Carência ●Emergência ●Urgência ●Saúde ●Plano de saúde
Processo 10	Imperatriz	Imperatriz	●pagamento ●Imperatriz
Processo 11	Saúde	Saúde	●recusa ●atendimento hospitalar ●atendimento médico ●carência ●emergência ●urgência ●saúde
Processo 12	Saúde	Saúde	●recusa ●atendimento médico ●medicamento ●carência ●saúde
Processo 13	Imperatriz	-Tema-972/STJ	●seguro
Processo 14	Saúde	Saúde	●Recusa ●Medicamento

			●Emergência ●Urgência ●Saúde
Processo 15	Saúde	Saúde	●Emergência ●Urgência ●Saúde
Processo 16	Saúde	Consignado	●Contrato
Processo 17	Saúde	Responsabilidade Civil	●Responsabilidade civil ●Estado ●Município
Processo 18	Ações Executivas	Saúde	●Recusa ●Atendimento Médico ●Medicamento ●Carência ●Saúde
Processo 19	Ações Executivas	Ações Executivas	●Execução Individual ●Sentença Coletiva ●Título Executivo ●Título Coletivo ●Liquidação ●Líquida
Processo 20	Ações Executivas	Escritório PR	●Ilegitimidade ●Legitimidade ●Associação ●Sindicato
Processo 21	Ações Executivas	Ações Executivas	●Liquidação ●Homologação
Processo 22	Ações Executivas	Saúde	●Recusa ●Atendimento Médico ●Medicamento ●Carência ●Saúde
Processo 23	Ações Executivas	Responsabilidade Civil	●Responsabilidade Civil ●Estado ●Município
Processo 24	Ações Executivas	Responsabilidade Civil	●Estado
Processo 25	Ações Executivas	Imperatriz	●Imperatriz
Processo 26	Ações Executivas	Administrativo	●Improbidade

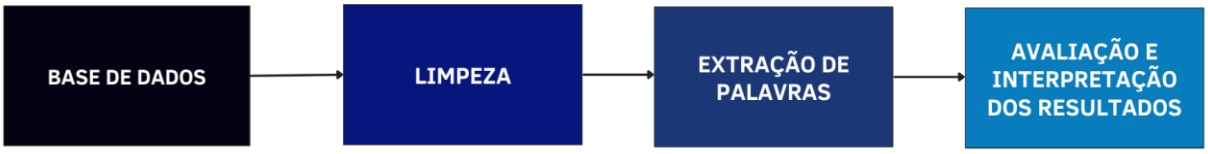
			Administrativa
Processo 27	Ações Executivas	Responsabilidade Civil	●Estado
Processo 28	Ações Executivas	Consignado	●Contrato
Processo 29	Ações Executivas	Consignado	●Contrato ●Empréstimo

Analisando os resultados, pode-se perceber que foram encontrados vários processos que realmente pertenciam à etiqueta informada pelo TOADA. O resultado também expôs processos que foram enquadrados em outras etiquetas às quais não pertenciam, isto ocorreu, pois não apresentavam um número suficiente de palavras ou nenhuma palavra chave da etiqueta informada pelo TOADA.

Também, houveram alguns processos que continham palavras de outras etiquetas e nenhuma da qual pertencia, provavelmente estes processos foram etiquetados de forma manual, ocasionando na etiquetação equivocada. Nesse cenário, foi constatado que o robô Clovis aplicava algumas palavras genéricas demais para agrupar a etiqueta, fazendo com que houvesse uma certa dificuldade para detecção de qual grupo o processo pertencia. Isto levanta a hipótese, que o uso de técnicas de extração de palavras automáticas pode apresentar palavras mais específicas e de cunho mais relevante.

4.2 FRAMEWORK EXPERIMENTAL

Após essas análises foi elaborado um framework experimental para averiguar os métodos de extração de palavras chave em documentos jurídicos. Esse framework foi baseado



nas pesquisas relacionadas. Na implementação e execução dos testes utilizaram-se o ambiente *Google Colab*, com a linguagem Python, e as bibliotecas Scikit-Learn, Natural Language Toolkit (NLTK) e as bibliotecas específicas de cada método de extração. Na Figura 8 vê-se as etapas do framework experimental.

Figura 8 – Etapas do Framework Experimental

4.2.1 **Base de Dados Utilizada**

Foram utilizadas duas bases de dados. A Primeira contendo 130 processos fornecidos pelo Laboratório de Inovação do Tribunal de Justiça do Maranhão, TOADA LAB. A outra foi uma base livre com 3543 jurisprudências do Tribunal de Justiça do Estado do Maranhão. A distribuição de frequências das jurisprudências vê-se na Figura 9.

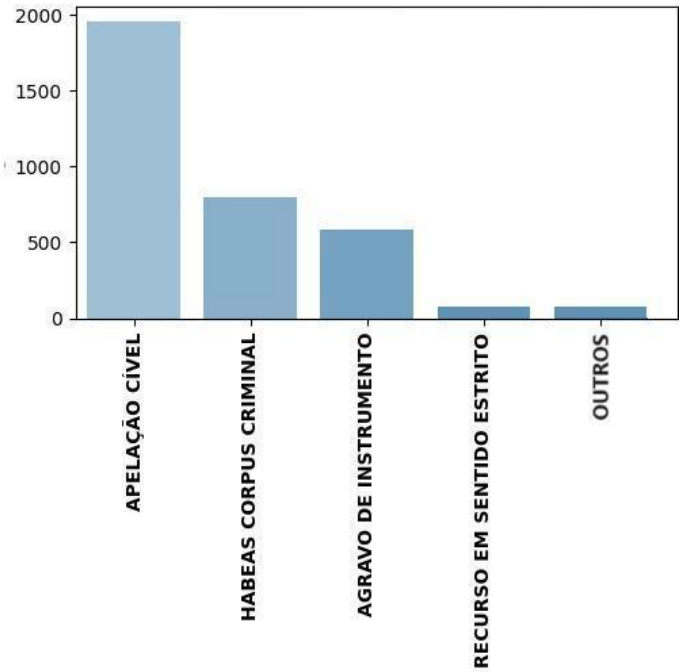


Figura 9 – Distribuição das Jurisprudências

Pode-se observar que os tipos mais comuns de jurisprudência são as de “Apelação Civil”" e a de “Habeas Corpus Criminal”".

4.2.2 LIMPEZA

Durante a fase de “limpeza” dos dados, o procedimento inicial na etapa de preparação envolveu a leitura dos arquivos e a estruturação de conjuntos de dados contendo informações e seus respectivos tipos. Para cada conjunto de dados, foi criado um corpus tabulado. Em seguida,

seguimos para a limpeza dos dados, utilizando como base a metodologia proposta por Cirqueira et al. (2018).

Neste estudo, ao lidar com os dados, todos os conteúdos foram convertidos para minúsculas a fim de evitar distinções entre palavras idênticas. Além disso, procedeu-se à remoção de espaços em branco, pontuações, símbolos financeiros, caracteres especiais e *stopwords*.

Destaca-se a importância da eliminação das *stopwords*, uma vez que este procedimento reduz o espaço de busca e aprimora significativamente o processo de aprendizagem do algoritmo (Sousa et al., 2019), contribuindo também para a sua explicabilidade (Cirqueira et al,, 2020)

4.2.3 **Extração de Palavras Chaves**

Na etapa de extração, foram selecionados quatro métodos baseados nos trabalhos apresentados na literatura, quais sejam: estáticos (YAKE e TF-IDF), baseado em grafos (TEXTRANK) e de incorporação (BerTopic) . Os *datasets* foram repassados para cada tipo de método, com a finalidade de gerar palavras chaves para cada tipo de documento jurídico. Os hiperparâmetros que cada método utilizou, foram empiricamente escolhidos e podem ser vistos na Tabela 2.

Tabela 2 Hiper-parâmetros dos Modelos Utilizados.

Algoritmo	**Parametrização**
Textrank	• Idioma='portugues' : Defina o idioma do texto de entrada • Nlp.max_length = 13960410 : Aumenta o tamanho maximo do NLP
YAKE	• Lan=”pt” : Define o idioma • Max_ngram_size=3 : Máximo de N gramas que uma palavra-chave deve ter. • WindowsSize=1: Tamanho da janela para coocorrência • Top=20 : número de palavras chaves geradas
TF-IDF	• Norm='l2 : Cada linha de saída terá norma unitária. • Smooth_idf=True : Suavize os pesos do IDF adicionando um às frequências do documento. • Use_idf=True : Reponderação de frequência de documento inversa • Sublinear_tf = false : Aplica escala tf sublinear

BerTopic	• Verbose=True : Altera o detalhamento do modelo • Top_n_words=20 : número de palavras por tópico • Min_topic_size=10 : O tamanho mínimo de um tópico • Language = "portuguese" : Define a linguagem

4.2.4 Avaliação e Interpretação dos Resultados

Nesta etapa, os resultados de cada método serão avaliados e analisados, tomando-se como partida o trabalho de Campos et al. (2020), que apresenta um cálculo baseado na matriz de confusão com a seguinte nomenclatura:

- 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 que concerne ao banco de dados dos processos jurídicos, o qual já incluía um corpus de palavras-chave para cada tipo de processo, essas palavras foram empregadas como orientações principais. Por outro lado, em relação ao conjunto de dados de jurisprudências, que não continha palavras-chave definidas, foram identificadas e analisadas as palavras mais específicas em cada jurisprudência, a fim de servirem como diretrizes.

Neste cenário, cada método foi avaliado individualmente e aplicado a cada conjunto de dados. Com os dados obtidos da avaliação, a partir da matriz de confusão, foram calculadas métricas consolidadas para avaliação de classificadores, a saber: precisão (*precision*, P), *recall* (R) e *F1-score*. Para calcular essas métricas, foram aplicadas as fórmulas apresentadas no trabalho de Li (2021). Na Figura 10 mostra-se as fórmulas descritas.

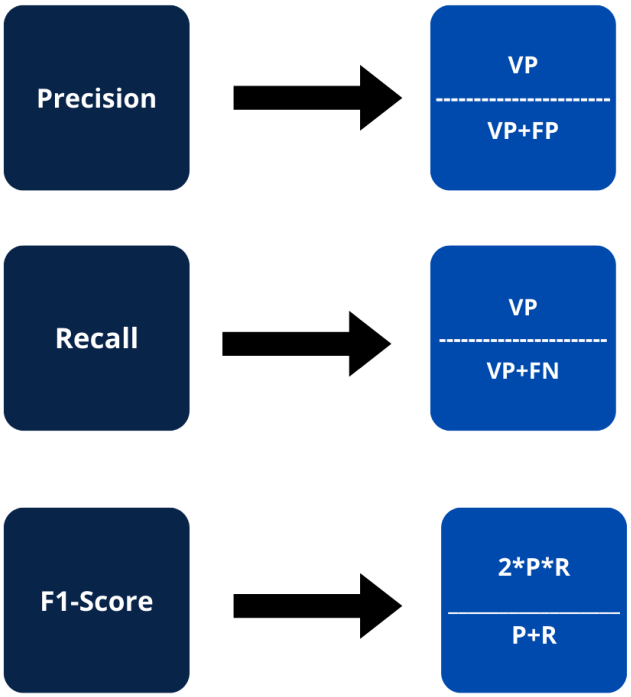


Figura 10 – Fórmulas Apresentadas no Trabalho de Li (2021).

A implementação desenvolvida neste trabalho será disponibilizada em um repositório criado no GitHub[1], visando a replicabilidade do estudo. O repositório incluirá os dados utilizados, o arquivo de testes e um documento inicial explicando cada etapa da implementação, juntamente com instruções sobre como utilizá-la. Devido à natureza sensível dos dados, por questões de segurança e privacidade, este repositório será configurado como privado. O acesso será concedido somente mediante autorização do TJMA.

4.2.5 Resultados do Framework Experimental

Os métodos foram aplicados aos dois conjuntos de dados e os resultados do conjunto de dados dos processos são apresentados na Tabela 3, exibindo uma comparação de cada método de extração nesses dados. Para fins de melhorar a exibição dos dados, os melhores resultados estão destacados em negrito e as métricas foram: P (*Precision*), R (*Recall*) e F1 (*F1-score*).

[1] https://github.com/jacobjr/jus-keywords

Tabela 3 Resultados Obtidos para as medidas de desempenho adotadas no dataset de Processos.

ETIQUETAS	YAKE			Textrank			TF-IDF			BERTOPIC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Intempestividade	75%	83%	79%	25%	45%	32%	75%	88%	81%	**80%**	**89%**	**84%**
Impedimento Tipo 3	**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%**
Suspeição Tipo 2	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%
Ações Coletivas	55%	**92%**	69%	55%	69%	61%	55%	58%	56%	**60%**	**92%**	**73%**

Pode ser observado, no conjunto de dados dos processos (Tabela 3), que o método BerTopic alcança as melhores métricas. Apenas nos casos do tipos Suspeição Tipo 2 e Imperatriz esse método tem métricas com valores menores.

No caso do conjunto de dados de jurisprudência, os resultados do conjunto de dados são apresentados na Tabela 4, também exibindo uma comparação de cada método de extração nesses dados.

Tabela 4 Resultados Obtidos para as medidas de desempenho adotadas no dataset de Jurisprudências..

ETIQUETAS	YAKE			Textrank			TF-IDF			BERTOPIC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Apelação Cível	45%	**90%**	60%	25%	42%	31%	25%	38%	30%	**90%**	78%	**84%**

Mandado De Segurança Cível	**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 Instrumento	50%	56%	53%	60%	57%	59%	30%	55%	39%	**75%**	**83%**	**79%**
Recurso Em Sentido Estrito	**75%**	**71%**	**73%**	**75%**	60%	67%	70%	47%	56%	55%	52%	54%
Mandado De Segurança Criminal	65%	54%	59%	60%	63%	62%	60%	52%	56%	**95%**	**73%**	**83%**
Habeas Corpus Criminal	50%	50%	50%	40%	50%	44%	40%	42%	41%	**80%**	**67%**	**73%**
Remessa Necessária Cível	40%	42%	41%	30%	38%	33%	**50%**	53%	**51%**	35%	**70%**	47%
Revisão Criminal	25%	29%	27%	60%	60%	60%	75%	56%	64%	**95%**	**76%**	**84%**

Nesta situação, mais uma vez, percebe-se um melhor desempenho do BerTopic . 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 das métricas utilizadas, 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 sobressai nessas categorias, atingindo pontuações notáveis em todas as métricas.

Em jurisprudências com pontuações moderadas, como “Ação Rescisória” e “Apelação Cível”, o 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*.

Além da análise, da *Precision*, *Recall* e *F1-score*. Também foi avaliado o tempo em que cada método levou para gerar as palavras-chaves.

Tabela 5 Tempo de execução de cada método extrator de palavras..

YAKE	TF-IDF	TEXTRANK	BERTOPIC
1min 20s	3.44 s	4min 56s	28.3 s

Analisando o tempo de execução, percebe-se que o método que levou mais tempo foi o Textrank com 4 mim e 56s, o TF-IDF e o YAKE obtiveram tempos medianos. O melhor método foi o BerTopic com 28.3 segundos. Em uma análise geral, BerTopic mostrou-se um método rápido, robusto e eficiente em diversas jurisprudências e processos, obtendo pontuações mais altas. A escolha entre os métodos pode depender das prioridades específicas de cada contexto de aplicação, seja a agilidade do método ou considerando a ênfase desejada em *Precision*, *Recall* ou *F1-score*. Por exemplo, o Textrank apesar de ser um método bastante utilizado no campo de extração de palavras, apresenta uma baixa eficácia no domínio jurídico e apresentou um tempo de resposta muito demorado. Já o BerTopic é uma escolha sólida para precisão e rapidez, enquanto o YAKE e o TF-IDF apesar de serem métodos que apresentam um tempo de resposta mediano, pode ser preferível para abrangência na recuperação de termos relevantes.

5. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou uma abordagem que tem como objetivo desenvolver um pipeline que agrupa os processos conforme o seu tema e com isso dando celeridade ao fluxo processual. Os resultados mostraram em ambos os conjuntos de dados que o método de extração BerTopic obteve melhores resultados nas métricas de avaliação e foi o método mais veloz, seguido pelo método YAKE que apesar de ter um tempo mediano de resposta, obteve valores significativos nas métricas de avaliação.

Conclui-se, portanto, que as técnicas do tipo estatístico (YAKE) e especialmente do tipo de incorporação (Bertopic) têm potencial para serem auxiliares na classificação de documentos de natureza legal, que dependendo do foco de aplicação no jurídico, pode ser usado tanto o método YAKE que apesar de não ser tão veloz, pode proporcionar resultados significativos, como o Bertopic que é um método robusto e rápido. Ambos podem proporcionar maior velocidade na distribuição processual. Destaque-se, porém, o melhor desempenho do Bertopic.

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.

REFERÊNCIAS

Almeida Neto, M., Moura, V., Bandeira, J., Freitas, P., and Fagundes, R. (2018). Um modelo de inferência para a classificação de resultados processuais da justiça estadual. Macabe a-Revista Eletroˆnica do Netlli, 3(3).

Aranha, Christian, and Emmanuel Passos. "A tecnologia de mineração de textos." Revista Eletrônica de Sistemas de Informação 5.2 (2006).

Araújo, Pedro Henrique Luz de, et al. "VICTOR: a dataset for Brazilian legal documents classification." Proceedings of The 12th Language Resources and Evaluation Conference. 2020.

Araújo, Valter Shuenquener, Anderson de Paiva Gabriel, and Fábio Ribeiro Porto de. "Justiça 4.0: a transformação tecnológica do poder judiciário deflagrada pelo CNJ no biênio 2020-2022." Revista Eletrônica Direito Exponencial-DIEX 1.1 (2022): 1-18.

Ba, Sileye. "Discovering topics with neural topic models built from PLSA assumptions." arXiv preprint arXiv:1911.10924 (2019).

Bafna, Prafulla, Dhanya Pramod e Anagha Vaidya. "Agrupamento de documentos: abordagem TF-IDF." Conferência Internacional de 2016 sobre Técnicas Elétricas, Eletrônicas e de Otimização (ICEEOT) . IEEE, 2016.

Berry, M. W. e Kogan, J. (2010). Text mining: applications and theory, John Wiley & Sons.

Brito, Bruno. TJPE usará inteligência artificial para agilizar processos de execução fiscal no Recife. Tribunal de Justiça de Pernambuco , Pernambuco, 25 de agosto de 2020.

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.

Camargo, Francielle Dolbert, and Oscar Valente Cardoso. "COMENTÁRIOS AO ARTIGO 330 DO NOVO CPC E AS CONTROVÉRSIAS NÃO RESOLVIDAS DO ART. 285-B DO CPC/73." Revista do CEJUR/TJSC: Prestação Jurisdicional 1.3 (2015): 181-191.

Castro Júnior, Antônio Pires, Wesley Pacheco Calixto de, and Cláudio Henrique Araujo de Castro. "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: 9.

Cohen, K. Bretonnel, and Lawrence Hunter. "Getting started in text mining." PLoS computational biology 4.1 (2008): e20.

CNJ (2022). Conselho Nacional de Justiça. Justiça em nu´meros 2022. CNJ, Brasília.

Dumais, Susan T. "Análise semântica latente." Annu. Rev. Inf. ciência Tecnol. 38.1 (2004): 188-230.

Falcão, Joaquim, et al. "Políticas públicas do Poder Judiciário: uma análise quantitativa e qualitativa do impacto da implantação do processo judicial eletrônico (PJe) na produtividade dos tribunais." (2018).

Gaussier, Eric, and Cyril Goutte. "Relation between PLSA and NMF and implications." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005.

Guo, Chonghui, Menglin Lu, and Wei Wei. "An improved LDA topic modeling method based on partition for medium and long texts." Annals of Data Science 8.2 (2021): 331-344.

Gupta, Tanya. "Keyword extraction: a review." International Journal of Engineering Applied Sciences and Technology 2.4 (2017): 215-220.

Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).

Hassani, Hossein, et al. "Text mining in big data analytics." Big Data and Cognitive Computing 4.1 (2020): 1.

Hollanda, Y. R. d. and Leite, F. G. d. F. (2020). Petição Inicial: uma análise à luz de teorias bakhtinianas. Macabéa-Revista Eletroˆnica do Netlli, 9(4):292–308.

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.

Jelodar, Hamed, et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." Multimedia Tools and Applications 78.11 (2019): 15169-15211.

.Justiça Digital (2021). Tjam automatiza classificação de petições intermediárias no portal e-saj, https://justicadigital.com/tjam-ia-peticionamento/, visitado em 28/11/2021.

Jungiewicz, M. e Łopuszyński, M. (2014). Unsupervised keyword extraction from polish legal texts, Advances in NaturalLanguageProcessing: 9thInternationalConference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9, Springer, pp. 65–70.

Kherwa, Pooja, and Poonam Bansal. "Topic modeling: a comprehensive review." EAI Endorsed transactions on scalable information systems 7.24 (2019).

Korenčić, D., Ristov, S., Repar, J. e Šnajder, J. (2021). A topic coverage approach to evaluation of topic models, IEEE Access 9: 123280–123312.

Leal, Rosemiro Pereira. Teoria geral do processo. IOB Thomson, 2005.

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.

Li, Wengen e Jiabao Zhao. "Algoritmo TextRank explorando a Wikipedia para extração de palavras-chave de texto curto." 2016 3ª Conferência Internacional sobre Ciência da Informação e Engenharia de Controle (ICISCE) . IEEE, 2016.

Limeira, Danielle. Robô organiza processos judiciais eletrônicos em 49 unidades do Judiciário. Tribunal de Justiça do Maranhão, São Luís, 30 de setembro de 2022.

Limeira, Danielle. TJMA utilizará robô Elis para triagem de dados em processos de execução fiscal. Tribunal de Justiça do Maranhão, São Luís, 20 de outubro de 2021.

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).

MANUAL, DO ADVOGADO. "PJE Disponível em https://www.pje.jus.br/wiki/index.php/Manual_do_Advogado (acesso em outubro de 2022).

Mastella, Juliana Obino. Uma metodologia usando ambientes paralelos para otimização da classificação de textos aplicada a documentos jurídicos. MS thesis. Pontifícia Universidade Católica do Rio Grande do Sul, 2020.

Mathur, Ayushi, and M. Suchithra. "Application of Abstractive Summarization in Multiple Choice Question Generation." 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES). IEEE, 2022.

Melo, Jairo Simão Santana, Verônica Ferreira Nascente, and Luiz Eduardo dos Santos. "TOTH: solução inteligente preditora de classe e assuntos para processos autuados no PJe." Sistema e-Revista CNJ 5.2 (2021): 77-85.

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.

Papagiannopoulou, Eirini, and Grigorios Tsoumakas. "A review of keyphrase extraction." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.2 (2020): e1339.

Pezzini, Anderson. "Mineração de textos: conceito, processo e aplicações." Revista Brasileira De Contabilidade E Gestão 5.10 (2017): 58-61.

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.

Ramos, Jefferson David Asevedo. "Protótipo de um software para a classificação de processos, conforme as Tabelas Processuais Unificadas do Conselho Nacional de Justiça." (2020).

Ravichandiran, Sudharsan. Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd, 2021.

Siddiqi, Sifatullah, and Aditi Sharan. "Keyword and keyphrase extraction techniques: a literature review." International Journal of Computer Applications 109.2 (2015).

Sidorov, Grigori. "Latent semantic analysis (LSA): Reduction of dimensions." Syntactic n-grams in Computational Linguistics. Springer, Cham, 2019. 17-19.

Skala, Daniel. Multi-Document Keyphrase Extraction. Diss. 2022.

SOUZA, Marcos de, and Renato Rocha SOUZA. "MODELAGEM DE TÓPICOS: Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina." Múltiplos Olhares em Ciência da Informação-ISSN 2237-6658; Vol. 9 No. 2 (2019): PPGGOG-Discentes 24.2 (2019).

Sousa, R. N. d. (2019). Minerjus: solução de apoio à classificação processual com uso de inteligência artificial. Master's thesis, Universidade Federal do Tocantins, Palmas, TO, Brasil.

Sleiman, Rita, Kim-Phuc Tran, and Sébastien Thomassey. "Natural Language Processing for Fashion Trends Detection." 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET). IEEE, 2022.

Stemler, Igor Tadeu Silva Viana. "Identificação de precedentes judiciais por agrupamento utilizando processamento de linguagem natural." (2019)

Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. Vol. 8. 1999.

Vayansky, Ike, and Sathish AP Kumar. "A review of topic modeling methods." Information Systems 94 (2020): 101582.

Yuan, Meng, Pauline Lin, and Justin Zobel. "Document Clustering vs Topic Models: A Case Study." Proceedings of the 25th Australasian Document Computing Symposium. 2021.