Universidade Estadual do Maranhão
Departamento de Engenharia da Computação

Ana Carolina Cutrim Bessa

Desenvolvimento de uma Ferramenta para Segmentação de Petições Iniciais

São Luís - MA
2025

Ana Carolina Cutrim Bessa

Desenvolvimento de uma Ferramenta para Segmentação de Petições Iniciais

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Engenharia da Computação, pela Universidade Estadual do Maranhão.

Orientador: Prof. Dr. Antonio Fernando Lavareda Jacob Junior

São Luís - MA

2025

Bessa, Ana Carolina Cutrim
Desenvolvimento de uma ferramenta para segmentação de petições iniciais. / Ana Carolina Cutrim Bessa. – São Luis, MA, 2025.
38 f

TCC (Graduação em Engenharia da Computação) - Universidade Estadual do Maranhão, 2025.

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

1.Automação Jurídica. 2.Expressões Regulares. 3.Petições Iniciais. 4.Segmentação de Textos. I.Título.

CDU: 347.9:004.434

Elaborado por Cássia Diniz - CRB 13/910

Ana Carolina Cutrim Bessa

Desenvolvimento de uma Ferramenta para Segmentação de Petições Iniciais

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Engenharia da Computação, pela Universidade Estadual do Maranhão.

São Luís - MA, (17 de Fevereiro de 2025):

Documento assinado digitalmente
gov.br ANTONIO FERNANDO LAVAREDA JACOB JUNIO
Data: 24/02/2025 11:03:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antonio Fernando Lavareda Jacob Junior
Orientador - UEMA

Documento assinado digitalmente
gov.br HUGO ASSIS PASSOS
Data: 24/02/2025 11:34:34-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Hugo Assis Passos
Examinador Interno - UEMA

Documento assinado digitalmente
gov.br THYAGO MACHADO RODRIGUES
Data: 24/02/2025 11:17:47-0300
Verifique em https://validar.iti.gov.br

Thyago Machado Rodrigues
Examinador Interno - PECS/UEMA

Dedico este trabalho a Deus, à minha família e amigos.

Agradecimentos

Agradeço primeiramente a Deus, por estar comigo em cada passo dessa caminhada, guiando-me e fortalecendo-me ao longo dos anos.

Aos meus pais, Ana e Ulisses, pelo amor incondicional, pelo apoio e por sempre acreditarem no meu potencial. Às minhas irmãs, Lorena e Safira, pela companhia, pelas risadas e por todos os momentos felizes.

Sou grata aos meus amigos de curso e de laboratório, pelo incentivo, pelas trocas de conhecimento e pelas experiências que compartilhamos. Aos meus amigos Camila, Sarah, Samuel e Marcelo, pelo companheirismo e pela amizade durante todos esses anos. Ao Rômulo, pelo carinho, pela paciência e por estar ao meu lado em todos os momentos, compartilhando alegrias e me motivando a continuar. Sem vocês, essa conquista não seria possível.

Agradeço à Universidade Estadual do Maranhão pela oportunidade e aos professores que contribuíram para minha formação, em especial ao meu orientador Antonio Jacob, pelo acompanhamento e orientação para a realização deste trabalho, e aos professores Paulo e Leonardo, pelos ensinamentos que me ajudaram no meu crescimento profissional.

Por fim, agradeço a todos que, de alguma forma, contribuíram para a realização deste trabalho. A cada um de vocês, minha mais sincera gratidão.

“O que realmente importa na vida é o que se faz com o tempo que nos é dado.”

J. R. R. Tolkien

Resumo

O crescente volume de processos judiciais tem gerado desafios na organização e análise de documentos, tornando essencial a adoção de soluções automatizadas. Nesse contexto, este trabalho propõe o desenvolvimento de uma ferramenta para segmentação automática de petições iniciais, identificando e extraindo automaticamente suas três seções principais: Fato, Tese e Pedido (FTP). O objetivo é melhorar a triagem desses documentos, reduzindo o tempo necessário para análise e contribuindo para o descongestionamento dos processos. A ferramenta foi implementada utilizando expressões regulares (RegEx) para identificação de padrões textuais e segmentação automática das petições. Para avaliar a qualidade da extração, foram aplicados dois métodos de validação: Similaridade de Jaccard e as métricas Precisão, *Recall* e F1-*score*, comparando os trechos extraídos automaticamente com as anotações manuais realizadas por especialistas. Os resultados revelam que a segmentação automática contribui para a automação da triagem de documentos jurídicos, minimizando a necessidade de intervenção manual e agilizando a análise textual. Apesar dos desafios relacionados com a variação estrutural das petições, a ferramenta representa um avanço na modernização do sistema judiciário.

Palavras-chave: Automação Jurídica; Expressões Regulares; Petições Iniciais; Segmentação de Textos.

Abstract

The growing volume of lawsuits has created challenges in organizing and analyzing documents, making it essential to adopt automated solutions. In this context, this paper proposes the development of a tool for the automatic segmentation of initial petitions, automatically identifying and extracting their three main sections: Fact, Thesis and Request (FTP). The aim is to improve the sorting of these documents, reducing the time needed for analysis and contributing to the decongestion of cases. The tool was implemented using regular expressions (RegEx) to identify textual patterns and automatically segment petitions. To assess the quality of the extraction, two validation methods were applied: Jaccard Similarity and the metrics Accuracy, Recall and F1-score, comparing the automatically extracted passages with the manual annotations made by experts. The results show that automatic segmentation contributes to the automation of legal document sorting, minimizing the need for manual intervention and speeding up textual analysis. Despite the challenges related to the structural variation of petitions, the tool represents an advance in the modernization of the judicial system.

Keywords*: Legal Automation; Regular Expressions; Initial Petitions; Text Segmentation.*

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AED	Análise Exploratória de Dados
BERNA	Busca Eletrônica em Registros usando Linguagem Natural
BERT	*Bidirectional Encoder Representations from Transformers*
CNJ	Conselho Nacional de Justiça
CPC	Código do Processo Civil
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
FTP	Fato, Tese e Pedido
IA	Inteligência Artificial
LaBSE	*Language-agnostic BERT Sentence Embedding*
LINCProg	Laboratório de Inteligência Computacional e Programação
MPMG	Ministério Público de Minas Gerais
PLN	Processamento de Linguagem Natural
RegEx	*Regular Expression*
SVM	*Support Vector Machine*
TJMA	Tribunal de Justiça do Estado do Maranhão
UEMA	Universidade Estadual do Maranhão

Sumário

1 INTRODUÇÃO

O cenário atual é caracterizado pela geração massiva de dados. A globalização e os avanços na comunicação tornaram o compartilhamento de informações mais acessível, potencializado pela internet (OLIVEIRA et al., 2022). Segundo Souza (2023), a produção de dados continua em crescimento acelerado, com projeções apontando um aumento significativo nos próximos anos. Esse aumento expressivo na geração de informações também impacta o setor jurídico, onde o alto volume de processos prejudica a eficiência do sistema judiciário (SANTOS et al., 2022).

O Conselho Nacional de Justiça (CNJ), órgão responsável pelo controle administrativo e financeiro do Poder Judiciário brasileiro, revelou em seu último relatório, *Justiça em Números*[1], que o sistema judiciário possui aproximadamente 83,3 milhões de processos pendentes de uma solução definitiva. Dentre eles, 22% encontram-se suspensos, sobrestados ou arquivados provisoriamente, à espera de uma definição jurídica futura, o que exige um grande esforço para sua triagem, análise e gestão (PORTO, 2022).

O Poder Judiciário brasileiro, em sua busca pela modernização e integração dos serviços judiciais, tem promovido a transformação digital por meio da iniciativa Justiça 4.0 (CNJ, 2021). Lançado em fevereiro de 2021, este programa tem como objetivo facilitar o acesso à justiça no Brasil por meio da implementação de ações e estratégias que ampliem a quantidade de ferramentas de apoio jurídico (BRAGANÇA; BRAGANÇA, 2019). Além disso, a iniciativa visa tornar o sistema judiciário mais acessível e menos complexo, contribuindo para a redução da morosidade processual (SCHNEIDER; MOREIRA, 2023).

No setor jurídico, o surgimento de tecnologias baseadas em inteligência artificial (IA) tem proporcionado melhorias na análise dos processos, possibilitando tomada de decisões mais rápidas (AMARAL, 2024). No entanto, a implementação dessas soluções ainda é um desafio, tanto pela resistência cultural, quanto pela necessidade de garantir precisão e conformidade com a legislação vigente (QUEIROZ; BUENO; LISBINO, 2024).

Um documento de grande importância para um processo é a petição inicial, pois é a peça processual que inaugura uma ação judicial e apresenta três elementos essenciais: Fato, Tese e Pedido (BIZARRO, 2022). A correta identificação e separação desses trechos é importante para a compreensão da demanda. No modelo tradicional, essa atividade é realizada manualmente, demandando tempo e esforço dos profissionais do Direito. Com a automação desse processo, é possível agilizar a triagem e o encaminhamento das ações, reduzindo o tempo de análise.

Diante desse contexto, este trabalho tem como foco o desenvolvimento de uma ferramenta para identificação e segmentação automática dos trechos das petições iniciais em três categorias

[1] https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024.pdf

fundamentais: Fato, Tese e Pedido (FTP). A proposta tem como objetivo melhorar a análise desses documentos, auxiliando os usuários na organização e compreensão das petições, reduzindo o tempo gasto na leitura e interpretação, além de contribuir para o descongestionamento dos processos e a uniformização das decisões.

À vista disso, este estudo busca responder às seguintes perguntas de pesquisa:

- Como a segmentação automática de petições iniciais pode contribuir para a redução do tempo de análise e organização dos documentos jurídicos?
- Quais são os principais desafios na aplicação de expressões regulares para a segmentação de textos jurídicos estruturados?
- A ferramenta desenvolvida consegue atingir um nível de precisão suficiente para ser utilizada na triagem de petições iniciais?
- Qual a efetividade da segmentação automática em comparação com a segmentação manual, considerando métricas de similaridade como Jaccard, Precisão, *Recall* e F1-*score*?

1.1 Acordo de Cooperação Técnica

O presente trabalho está inserido no escopo do Acordo de Cooperação Técnica nº 002/2021 firmado entre o Tribunal de Justiça do Maranhão (TJMA) e a Universidade Estadual do Maranhão (UEMA). No contexto dessa parceria, esta pesquisa contribui para o desenvolvimento de uma ferramenta para identificar e separar os trechos das petições iniciais, alinhando-se às diretrizes da Justiça 4.0 para promover a modernização no segmento jurídico.

1.2 Objetivos

Com base no contexto apresentado, foram definidos os seguintes objetivos:

1.2.1 Objetivo Geral

Desenvolver uma ferramenta automatizada para identificar e segmentar trechos de petições iniciais em suas partes constitutivas: Fato, Tese e Pedido, utilizando técnicas de mineração de texto e expressões regulares.

1.2.2 Objetivos Específicos

- Coletar e organizar um conjunto representativo de petições iniciais fornecidas pelo TJMA para uso na pesquisa;

- Obter anotações manuais das petições iniciais, identificadas as seções de FTP, realizada por pesquisadores do Laboratório de Inteligência Computacional e Programação (LINCProg) com experiências em projetos de Processamento de Linguagem Natural (PLN);
- Desenvolver um algoritmo utilizando expressões regulares para realizar a segmentação automática das petições iniciais em FTP;
- Validar a eficiência da ferramenta desenvolvida por meio da comparação entre o resultado da separação automatizada e as anotações manuais obtidas utilizando o Coeficiente de Similaridade de Jaccard e as métricas de avaliação Precisão, *Recall* e F1-*score*.

1.3 Estrutura do Trabalho

Este documento está organizado em cinco seções. A Seção 2, Referencial Teórico, discute os fundamentos das petições iniciais, expressões regulares, pré-processamento e mineração de textos. A Seção 3 trata dos trabalhos correlatos sobre ferramentas desenvolvidas para segmentação/extração de informações em textos, em diferentes contextos. A Seção 4 descreve a metodologia empregada na pesquisa. Na Seção 5, Resultados e Discussões, detalha-se o a obtenção dos dados manuais, a anotação automática e a análise dos resultados obtidos. Por fim, a Seção 6 apresenta as Considerações Finais, consolidando as conclusões e propondo direções para trabalhos futuros.

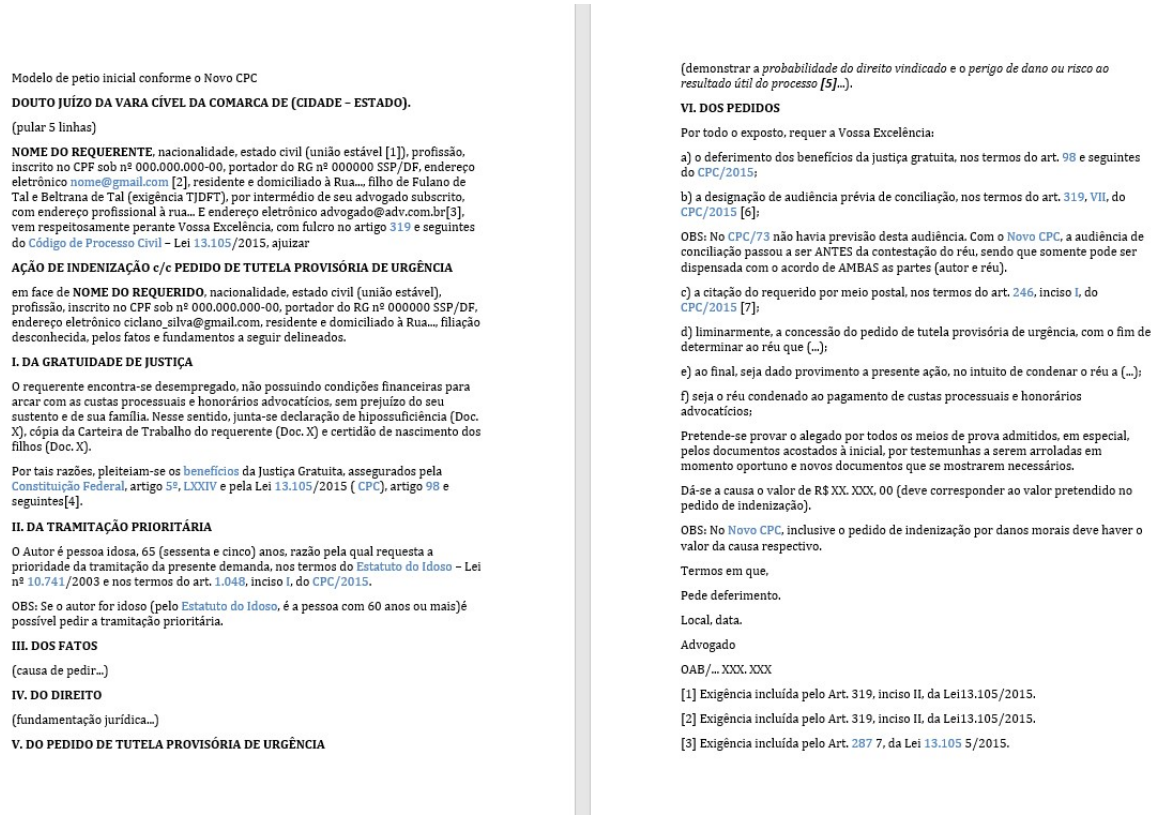
2 REFERENCIAL TEÓRICO

Esta seção apresenta o referencial teórico que fundamentou o presente estudo, abordando os principais conceitos e tecnologias relevantes para a pesquisa. A primeira subseção trata da petição inicial e sua estrutura. A segunda subseção apresenta o conceito de mineração de textos e explica o pré-processamento e suas técnicas. Por fim, a terceira subseção conceitua expressões regulares e seu uso na identificação de padrões

2.1 Petição Inicial

A petição inicial é um documento jurídico que marca o início de um processo judicial. Conforme definido pelo artigo 319 do Código de Processo Civil brasileiro, a petição inicial é o instrumento pelo qual o autor apresenta sua demanda ao juízo, expondo os fatos, os fundamentos jurídicos do pedido e as suas pretensões (Brasil, 2015). A Figura 1 exibe o modelo de petição inicial segundo o CPC de 2015.

Figura 1 – Modelo de petição inicial conforme o Novo CPC



Modelo de petio inicial conforme o Novo CPC

DOUTO JUÍZO DA VARA CÍVEL DA COMARCA DE (CIDADE – ESTADO).

(pular 5 linhas)

NOME DO REQUERENTE, nacionalidade, estado civil (união estável [1]), profissão, inscrito no CPF sob nº 000.000.000-00, portador do RG nº 000000 SSP/DF, endereço eletrônico nome@gmail.com [2], residente e domiciliado à Rua..., filho de Fulano de Tal e Beltrana de Tal (exigência TJDFT), por intermédio de seu advogado subscrito, com endereço profissional à rua... E endereço eletrônico advogado@adv.com.br[3], vem respeitosamente perante Vossa Excelência, com fulcro no artigo 319 e seguintes do Código de Processo Civil – Lei 13.105/2015, ajuizar

AÇÃO DE INDENIZAÇÃO c/c PEDIDO DE TUTELA PROVISÓRIA DE URGÊNCIA

em face de **NOME DO REQUERIDO**, nacionalidade, estado civil (união estável), profissão, inscrito no CPF sob nº 000.000.000-00, portador do RG nº 000000 SSP/DF, endereço eletrônico ciclano_silva@gmail.com, residente e domiciliado à Rua..., filiação desconhecida, pelos fatos e fundamentos a seguir delineados.

I. DA GRATUIDADE DE JUSTIÇA

O requerente encontra-se desempregado, não possuindo condições financeiras para arcar com as custas processuais e honorários advocatícios, sem prejuízo do seu sustento e de sua família. Nesse sentido, junta-se declaração de hipossuficiência (Doc. X), cópia da Carteira de Trabalho do requerente (Doc. X) e certidão de nascimento dos filhos (Doc. X).

Por tais razões, pleiteiam-se os benefícios da Justiça Gratuita, assegurados pela Constituição Federal, artigo 5º, LXXIV e pela Lei 13.105/2015 (CPC), artigo 98 e seguintes[4].

II. DA TRAMITAÇÃO PRIORITÁRIA

O Autor é pessoa idosa, 65 (sessenta e cinco) anos, razão pela qual requesta a prioridade da tramitação da presente demanda, nos termos do Estatuto do Idoso – Lei nº 10.741/2003 e nos termos do art. 1.048, inciso I, do CPC/2015.

OBS: Se o autor for idoso (pelo Estatuto do Idoso, é a pessoa com 60 anos ou mais)é possível pedir a tramitação prioritária.

III. DOS FATOS

(causa de pedir...)

IV. DO DIREITO

(fundamentação jurídica...)

V. DO PEDIDO DE TUTELA PROVISÓRIA DE URGÊNCIA

(demonstrar a *probabilidade do direito vindicado* e o *perigo de dano ou risco ao resultado útil do processo* ***[5]***...).

VI. DOS PEDIDOS

Por todo o exposto, requer a Vossa Excelência:

a) o deferimento dos benefícios da justiça gratuita, nos termos do art. 98 e seguintes do CPC/2015;

b) a designação de audiência prévia de conciliação, nos termos do art. 319, VII, do CPC/2015 [6];

OBS: No CPC/73 não havia previsão desta audiência. Com o Novo CPC, a audiência de conciliação passou a ser ANTES da contestação do réu, sendo que somente pode ser dispensada com o acordo de AMBAS as partes (autor e réu).

c) a citação do requerido por meio postal, nos termos do art. 246, inciso I, do CPC/2015 [7];

d) liminarmente, a concessão do pedido de tutela provisória de urgência, com o fim de determinar ao réu que (...);

e) ao final, seja dado provimento a presente ação, no intuito de condenar o réu a (...);

f) seja o réu condenado ao pagamento de custas processuais e honorários advocatícios;

Pretende-se provar o alegado por todos os meios de prova admitidos, em especial, pelos documentos acostados à inicial, por testemunhas a serem arroladas em momento oportuno e novos documentos que se mostrarem necessários.

Dá-se a causa o valor de R$ XX. XXX, 00 (deve corresponder ao valor pretendido no pedido de indenização).

OBS: No Novo CPC, inclusive o pedido de indenização por danos morais deve haver o valor da causa respectivo.

Termos em que,

Pede deferimento.

Local, data.

Advogado

OAB/... XXX. XXX

[1] Exigência incluída pelo Art. 319, inciso II, da Lei13.105/2015.

[2] Exigência incluída pelo Art. 319, inciso II, da Lei13.105/2015.

[3] Exigência incluída pelo Art. 287 7, da Lei 13.105 5/2015.

Fonte: JUSBRASIL (2016).

Conforme exemplificado no modelo da Figura 1, a petição inicial contém elementos principais, como qualificação das partes, a narrativa dos fatos, os fundamentos jurídicos do

pedido, o pedido com suas especificações, o valor da causa, as provas com que o autor pretende demonstrar a verdade dos fatos alegados, e a opção pela realização ou não de audiência de conciliação ou de mediação (BIZARRO, 2022). Os três elementos que formam o núcleo da petição são os (*i*) **fatos**: narração clara e objetiva dos acontecimentos que fundamentam a ação; a (*ii*) **direitos/tese jurídica**: argumentação jurídica que sustenta o pedido, baseada na legislação, doutrina e jurisprudência; e o (*iii*) **pedido**: a formulação explícita do que se pretende obter com a ação judicial (JÚNIOR; CALIXTO; CASTRO, 2020).

2.2 Mineração de Textos

A mineração de texto é o processo de identificar e extrair automaticamente informações e padrões ocultos em grandes volumes de dados textuais não estruturados, como textos em linguagem natural (HASSANI et al., 2020). Esse processo identifica fatos, relacionamentos e declarações, extraindo informações que são, em seguida, transformadas em uma forma estruturada, apropriada para análises posteriores (ABDUSALOMOVNA et al., 2023).

2.2.1 Pré-processamento de Texto

Segundo Rocha (2020), o pré-processamento de dados na mineração de texto consiste em técnicas para estruturar dados brutos, tornando-os adequados para modelos de aprendizado de máquina, convertendo textos em representações numéricas compreensíveis pelo computador. Esta fase é importante para lidar com a diversidade dos textos, tornando as entradas mais padronizadas para melhorar o desempenho da segmentação. As técnicas utilizadas durante este estudo foram a remoção de números/dígitos, remoção de caracteres não alfanuméricos, remoção de espaços em branco especiais e transformação de maiúsculas para minúsculas.

Para Silva et al. (2023), os documentos jurídicos possuem características específicas, como terminologia técnica e estrutura formal, que exigem um tratamento adequado para eliminar ruídos, normalizar o texto e preservar informações relevantes. Esse processo garante que os dados extraídos sejam consistentes e apropriados para aplicações, contribuindo para a melhoria da análise jurídica e da eficiência dos sistemas automatizados (SILVA et al., 2023).

Portanto, se faz necessária a aplicação de técnicas de pré-processamento nas petições iniciais para assegurar a qualidade e a padronização dos dados, facilitando o processo de identificação e separação dos trechos dos documentos.

2.3 Expressões Regulares

As expressões regulares (RegEx) são ferramentas utilizadas em programação para a extração e manipulação de dados textuais (CHEN et al., 2023). Formuladas em uma linguagem formal, elas permitem identificar padrões específicos dentro de um texto, facilitando tarefas como

validação de dados, busca e substituição de *substrings* em *strings* (LEMOS; COELHO, 2023; TUROŇOVÁ et al., 2020).

Basicamente, uma RegEx é uma sequência de caracteres que define um padrão de pesquisa, funcionando como uma regra que determina se uma determinada entrada de dados corresponde a esse padrão. Quando um texto obedece exatamente às condições estabelecidas, considera-se que houve uma correspondência bem-sucedida (JARGAS, 2016).

Dessa forma, o uso de expressões regulares torna a manipulação de texto mais simples, permitindo localizar, extrair e processar informações de maneira automatizada e precisa.

3 TRABALHOS CORRELATOS

A pesquisa realizada por Júnior, Calixto e Castro (2020) teve como objetivo identificar e unificar automaticamente processos judiciais que compartilham o mesmo fato e tese jurídica, utilizando uma ferramenta de inteligência artificial chamada Berna. O método envolve técnicas de PLN, aprendizagem por similaridade e Redes Neurais Artificiais para analisar petições iniciais em tramitação. A aplicação desse modelo resultou na identificação de 13 petições idênticas e na constatação de que 20% dos processos nas Turmas Recursais, além de 8% nos Juizados Especiais Cíveis de Goiânia, continham similaridades significativas. Como resultados, a ferramenta apresentou uma precisão de 96% nos estudos de casos.

Tata et al. (2021) apresenta um sistema chamado *Glean*, desenvolvido por uma equipe do *Google*, voltado para a extração de informações estruturadas de documentos com formatos variáveis, como faturas e solicitações de pagamento. O *Glean* utiliza técnicas de aprendizado de máquina para lidar com a diversidade de *layouts* associados a esses documentos, permitindo a generalização em qualquer formato. São destacados três desafios principais relacionados à gestão de dados: a qualidade dos dados de referência, a geração de dados de treinamento a partir de documentos rotulados e a construção de ferramentas para desenvolver e aprimorar modelos para tipos específicos de documentos. Os resultados mostram que o método de gestão de dados adotado no *Glean* melhorou o desempenho do modelo.

O estudo feito por Constantino et al. (2022) mostra a segmentação e classificação semântica de trechos de diários oficiais do Ministério Público de Minas Gerais (MPMG), visando otimizar a extração e categorização de informações importantes para promover maior transparência. É proposta uma heurística orientada à estrutura para a segmentação de documentos PDF e utilizada uma estratégia de aprendizado ativo para realizar a classificação semântica, o que minimiza o esforço manual de rotulagem. Os resultados experimentais mostram uma acurácia de 100% na extração e de 85% na classificação.

Em Gomes, Sá e Peng (2020), os pesquisadores investigam a aplicação de técnicas de IA e mineração de texto para a classificação de sentenças judiciais brasileiras, com foco na análise da procedência dos pedidos dos autores e nas implicações para a formulação de políticas públicas. Foram coletados dados de sentenças do Tribunal Regional Federal da 2ª região e implementadas técnicas de PLN, como *bag-of-words* e TFIDF, além do uso de RegEx em classificadores para extrair informações importantes e aplicado algoritmos de aprendizado de máquina para a classificação. Os resultados do modelo de aprendizado de máquina baseado em RegEx e bigramas alcançaram uma precisão média de 94%, *recall* de 91% e F1-*score* de 92% ao classificar as sentenças.

A Tabela 1 resume os trabalhos relacionados à segmentação de texto em diferentes

domínios, destacando os modelos utilizados e a descrição da pesquisa.

Tabela 1 – Resumo dos trabalhos correlatos.

Autor	**Modelo**	**Descrição**
(JÚNIOR; CALIXTO; CASTRO, 2020)	BERNA	Aplicação de inteligência artificial na ferramenta Berna para identificar conexões entre processos judiciais, com base no fato e tese jurídica.
(TATA et al., 2021)	*Glean*	Apresentação do sistema *Glean*, uma solução de extração automática de informações estruturadas de documentos templáticos.
(CONSTANTINO et al., 2022)	SVM, BERT, BERTimbau e LaBSE	Aplicação de heurística de segmentação e uma estratégia de aprendizado ativo para a classificação semântica de trechos de diários oficiais.
(GOMES; SÁ; PENG, 2020)	Regressão Logística	Aplicação de técnicas de inteligência artificial e mineração de texto para classificar sentenças judiciais brasileiras.

Fonte: De autoria própria.

Os trabalhos analisados compartilham a aplicação de técnicas de PLN e aprendizado de máquina para melhorar o tratamento de documentos. Júnior, Calixto e Castro (2020) e Constantino et al. (2022) exploram a segmentação e classificação de textos jurídicos, seja na identificação de petições semelhantes ou na extração de informações de diários oficiais. Da mesma forma, Tata et al. (2021) apresenta um sistema voltado para a extração automatizada de informações estruturadas, tratando desafios como a variação de formatos e a qualidade dos dados.

Diferente dos trabalhos citados, esta pesquisa foca na segmentação automatizada de petições iniciais, dividindo-as em FTP. Enquanto Júnior, Calixto e Castro (2020) busca identificar similaridades entre processos e Constantino et al. (2022) propõe uma heurística para segmentação de diários oficiais, este estudo estrutura petições, facilitando sua análise por modelos de IA e profissionais do meio jurídicos. Além disso, enquanto Tata et al. (2021) se concentra em documentos de natureza administrativa e financeira, este trabalho baseia-se em textos jurídicos específicos, permitindo a organização das informações processuais e contribuindo para a melhoria do fluxo de trabalho nos tribunais.

4 METODOLOGIA

Nesta seção é descrito o procedimento metodológico adotado neste estudo com o propósito de alcançar os objetivos definidos na Seção 1. O método de pesquisa utilizado foi o *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

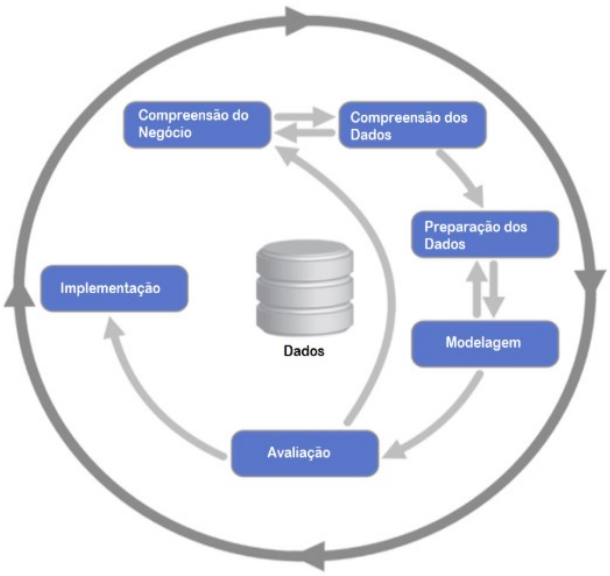
A metodologia adotada neste trabalho é de natureza experimental, pois envolve a compreensão da estrutura das petições iniciais, identificação dos padrões textuais e o desenvolvimento e a avaliação de uma ferramenta computacional para a segmentação dos trechos. A pesquisa foi conduzida por meio da implementação e testes de um modelo baseado em expressões regulares, seguido pela análise comparativa dos resultados obtidos com anotações manuais.

4.1 *Cross Industry Standard Process for Data Mining*

CRISP-DM é uma metodologia estruturada para a execução de projetos de mineração de dados. Desenvolvido em 1996 por um consórcio de líderes da indústria, o CRISP-DM é um modelo de processo não proprietário, bem documentado e disponível gratuitamente, o que facilita sua acessibilidade e aplicação em diferentes áreas (SHEARER, 2000).

Segundo Chapman et al. (2000), o ciclo de vida de um projeto de mineração de dados é composto por seis fases, como mostrado na Figura 2, onde a sequência dessas fases não é fixa, sendo comum a necessidade de avançar e retornar entre elas de acordo com os resultados obtidos em cada etapa.

Figura 2 – Fases do CRISP-DM



Fonte: Adaptado de Chapman et al. (2000).

Conforme a Figura 2, a trajetória para o presente trabalho é organizada da seguinte forma: Entendimento de negócio, Compreensão dos dados, Preparação dos dados, Modelagem, Avaliação e Implementação.

1. **Entendimento de negócio:** Identificação dos objetivos do projeto no contexto do sistema judiciário e planejamento inicial para desenvolver uma solução capaz de segmentar petições iniciais automaticamente, considerando as necessidades específicas do domínio jurídico;
2. **Compreensão dos Dados:** Coleta e exploração de petições iniciais fornecidas pelo TJMA;
3. **Preparação dos Dados:** Construção do conjunto final de dados, incluindo limpeza, normalização, remoção de inconsistências;
4. **Modelagem:** Desenvolvimento e aplicação de expressões regulares para segmentar as petições iniciais em **fato**, **tese** e **pedido**;
5. **Avaliação:** Comparação entre as anotações manuais realizadas pelos especialistas e os resultados do segmentador;
6. **Implementação:** Conclusão do processo com a apresentação de uma ferramenta funcional e documentação dos resultados obtidos para futuras melhorias e aplicações.

4.1.1 Entendimento do Negócio

De acordo com Leite (2024), o sistema judiciário brasileiro apresenta uma sobrecarga de processos, o que resulta em atrasos, impactando negativamente a eficiência dos julgamentos. Além disso, a análise manual de petições iniciais possui desafios, como a ausência de padronização, grandes volumes de texto e o uso de terminologias jurídicas específicas. Esses fatores dificultam a organização e o processamento dos documentos, causando o atraso das análises.

Diante disso, o desenvolvimento de uma ferramenta automatizada para identificar e segmentar petições iniciais em suas partes fundamentais — fato, tese e pedido —, pode contribuir para a melhoria deste cenário. Utilizando técnicas de mineração de texto e expressões regulares, é possível definir regras para buscar padrões nos documentos jurídicos e separar cada seção correspondente, facilitando a visualização das petições e reduzindo o tempo de análise.

Por tanto, a ferramenta desenvolvida neste trabalho deve identificar e segmentar automaticamente as seções de uma petição inicial, classificando-as nas categorias FTP. Para assegurar sua adaptabilidade às mudanças nas características dos documentos ao longo do tempo, a ferramenta deverá permitir ajustes dos padrões de identificação criados.

4.1.2 Compreensão dos Dados

O conjunto de dados utilizado neste projeto foi fornecido pelo TJMA e é composto por 108.907 petições iniciais de diferentes modelos. Contudo, para os experimentos deste trabalho,

foram selecionadas 100 petições iniciais, a fim de viabilizar uma análise mais focada e controlada. Para garantir a qualidade e a adequação desses dados ao objetivo do projeto, foi realizada uma Análise Exploratória de Dados (AED), permitindo identificar padrões, verificar a consistência e compreender as principais características do conjunto selecionado.

Uma das etapas fundamentais da AED foi a análise da quantidade de *tokens* presentes em cada petição inicial. A segmentação dos textos em *tokens* foi realizada utilizando o *tokenizer* da biblioteca *NLTK* (*Natural Language Toolkit*), que permite dividir o texto em unidades menores, como palavras ou pontuações. Essa escolha se deu devido à necessidade de um método para lidar com a segmentação de textos jurídicos, que possuem estruturas complexas e terminologias específicas.

A realização dessa análise é essencial, pois a distribuição do número de *tokens* por documento fornece informações sobre o tamanho médio das petições e a variabilidade dos textos no conjunto de dados. Essa informação é relevante tanto para a construção da ferramenta de segmentação quanto para o ajuste das expressões regulares, uma vez que a identificação de padrões estruturais nos documentos depende da regularidade e extensão dos textos analisados. Além disso, a análise da quantidade de *tokens* permite identificar a presença de documentos excessivamente longos ou curtos, que podem representar casos atípicos (*outliers*) e afetar a modelagem da segmentação. A distribuição dessa métrica pode ser visualizada na Tabela 2.

Tabela 2 – Estatísticas da quantidade de *tokens* por petição

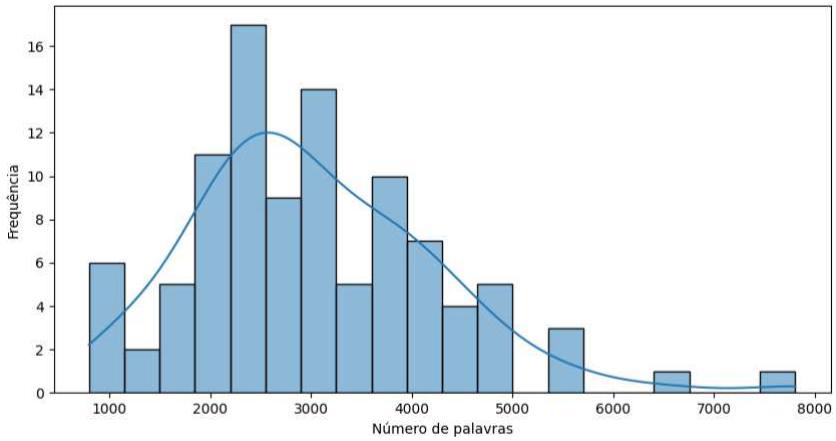
Métrica	**Valor**
Contagem (*count*)	100
Média (*mean*)	3.032,99
Desvio Padrão (*std*)	1.214,59
Mínimo (*min*)	796
1º Quartil (25%)	2.245,5
Mediana (50%)	2.906,5
3º Quartil (75%)	3.807,75
Máximo (*max*)	7.805

Fonte: De autoria própria.

Os resultados apresentados na Tabela 2 revelam uma variação no número de *tokens* por petição, com uma média de 3.032 *tokens* e um desvio padrão de 1.214, indicando diferenças consideráveis no tamanho dos documentos (Figura 3).

Além disso, observa-se que 25% das petições possuem menos de 2.245 *tokens*, enquanto 50% apresentam até 2.906 *tokens* (mediana) e 75% não ultrapassam 3.807 tokens, sugerindo que a maioria dos textos segue um padrão relativamente estável, embora existam documentos consideravelmente maiores. O maior valor identificado (7.805 *tokens*) indica a presença de *outliers*, possivelmente relacionados a petições mais extensas, anexos incorporados ou trechos

Figura 3 – Distribuição do tamanho de petições

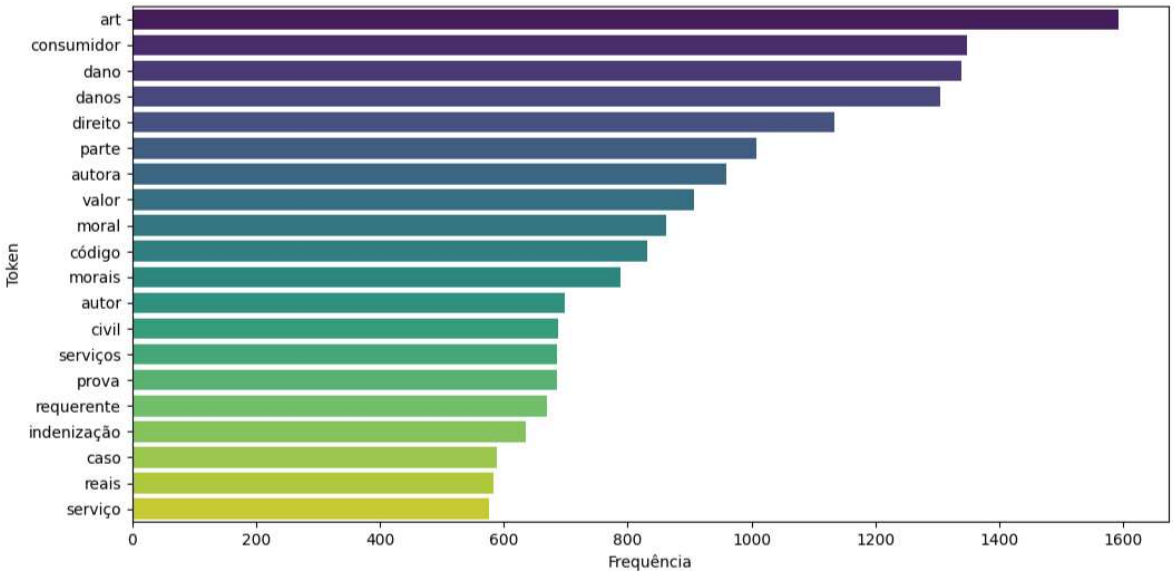


Fonte: De autoria própria.

redundantes. Além disso, a ocorrência de petições com um número reduzido de *tokens* pode indicar documentos incompletos ou registros corrompidos.

Outra análise realizada foi quanto à frequência de ocorrência das palavras no conjunto de petições, excluindo as *stopwords*. O objetivo dessa análise foi identificar os termos mais representativos dos documentos. A Figura 4 exibe o resultado obtido nesta análise.

Figura 4 – *Tokens* mais recorrentes nas petições



Fonte: De autoria própria.

A Figura 4 mostra que os termos “art”, “consumidor”, e “dano”/“danos” aparecem com alta frequência, indicando que as petições frequentemente tratam de questões relacionadas a danos e responsabilidades, com ênfase em casos envolvendo consumidores. O termo “art” está relacionado à referência a artigos legais. Os *tokens* “direito”, “parte”, “autora” e “valor” caracterizam elementos das ações jurídicas, como a parte envolvida, o direito discutido e os valores pleiteados. Já as palavras “serviço”, “indenização” e “reais” aparecem com menor

frequência, o que pode significar um foco específico em ações de indenização e reparação de danos.

4.1.3 Preparação dos Dados

A etapa de pré-processamento é essencial, pois alguns textos apresentam elementos que podem prejudicar a análise, como erros ortográficos, palavras irrelevantes e variações no formato (GRANCHAROVA; JANGEFALK, 2018). Ao padronizar e limpar os dados, o pré-processamento contribui para aumentar a precisão dos algoritmos.

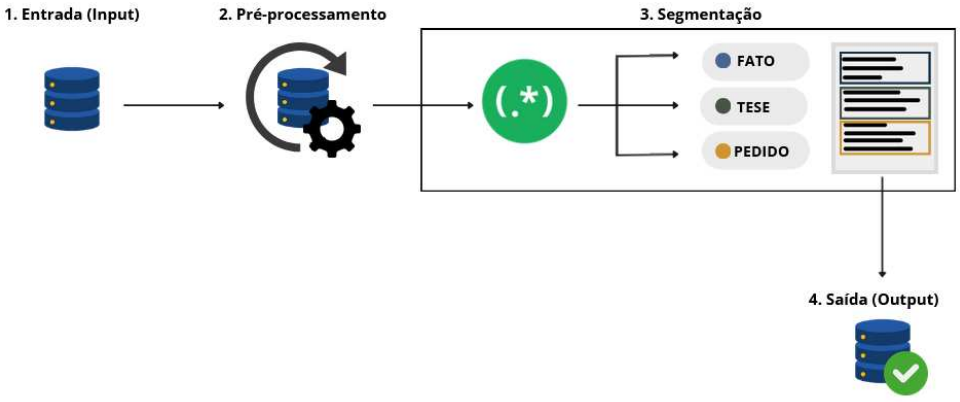
Durante esta fase, foram aplicadas técnicas de pré-processamento para tratar as petições iniciais no momento da execução da ferramenta. As principais técnicas aplicadas foram:

- **Remoção de números/dígitos:** Números isolados ou sequências numéricas foram eliminados para evitar interferências na análise semântica;
- **Remoção de caracteres não alfanuméricos:** Símbolos, pontuações e outros caracteres especiais foram filtrados para evitar ruídos no texto;
- **Remoção de espaços em branco excessivos:** Foram eliminados espaços em branco desnecessários, que poderiam comprometer a normalização e estruturação das petições;
- **Transformação de maiúsculas em minúsculas (*lowercase*):** A conversão de todo o texto para minúsculas foi realizada para evitar desigualdade entre palavras que possuem a mesma grafia, mas aparecem em diferentes formatos.

4.1.4 Modelagem

A modelagem da ferramenta seguiu um fluxograma estruturado em quatro fases principais, conforme a Figura 5.

Figura 5 – Fluxograma da ferramenta



Fonte: De autoria própria.

Inicialmente, na fase de entrada, a base de dados contendo as petições iniciais do TJMA foi carregada no ambiente de desenvolvimento. Em seguida, na fase de pré-processamento, foram aplicadas técnicas para limpeza e padronização dos textos, removendo ruídos e uniformizando a estrutura textual para facilitar a extração das informações.

A fase de segmentação consistiu na aplicação de RegEx para identificar e separar automaticamente as seções de Fato, Tese e Pedido dentro das petições. Para isso, foram projetados padrões capazes de reconhecer os elementos textuais característicos de cada seção seguindo o modelo de petição inicial do Novo CPC.

Por fim, na fase de saída, os trechos extraídos foram armazenados de forma estruturada em uma nova base de dados para utilizações posteriores. A implementação foi feita na linguagem *Python*, com bibliotecas especializadas na manipulação e organização de textos. A biblioteca *re* foi utilizada para o processamento das expressões regulares, enquanto *pandas* foi empregada para estruturar os dados extraídos em um *DataFrame*, facilitando sua manipulação e análise. O código foi desenvolvido para processar cada petição iterativamente, aplicando os padrões de segmentação definidos e armazenando os resultados de maneira organizada.

4.1.5 Avaliação

A avaliação dos resultados obtidos neste estudo foi conduzida por meio da comparação entre os segmentos identificados automaticamente pela ferramenta e aqueles anotados manualmente. Para mensurar a qualidade da segmentação automática foi utilizado o Coeficiente de Similaridade de Jaccard e as métricas de Precisão, *Recall* e F1-*score*.

4.1.5.1 Coeficiente de Similaridade de Jaccard

O coeficiente de Jaccard é definido como a razão entre a interseção e a união dos elementos presentes nos segmentos comparados:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

onde:

- A representa o conjunto de palavras do trecho segmentado manualmente;
- B representa o conjunto de palavras do trecho identificado automaticamente.

Essa métrica varia de 0 a 1, onde valores próximos de 1 indicam uma alta similaridade entre as segmentações e valores próximos de 0 apontam baixa correspondência (BEZERRA et al., 2021).

4.1.5.2 Precisão, *Recall* e F1-*score*

A Precisão (*Precision*) indica a proporção de palavras corretamente extraídas pela ferramenta em relação ao total de palavras que ela identificou como pertencentes a uma determinada seção, sendo calculada da seguinte forma:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

onde:

- TP (Verdadeiros Positivos) representa as palavras corretamente extraídas pela ferramenta;
- FP (Falsos Positivos) são palavras indevidamente atribuídas à seção pela ferramenta.

O *Recall* mede a capacidade da ferramenta de recuperar corretamente as palavras que deveriam estar na seção, sendo definido por:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

onde:

- FN (Falsos Negativos) corresponde às palavras que estavam no segmento manual, mas não foram extraídas pela ferramenta.

Por fim, o F1-*score* é a média harmônica entre Precisão e *Recall*, proporcionando uma medida balanceada entre esses dois aspectos da segmentação:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

4.1.6 Implementação

A etapa de implementação representou a conclusão do processo de desenvolvimento da ferramenta. Após a definição do fluxo de segmentação e a avaliação dos resultados, a ferramenta foi finalizada em sua versão funcional, permitindo a identificação e separação automática das seções FTP nas petições analisadas.

O código-fonte da ferramenta está disponibilizado na plataforma *GitHub*[1] com a devida documentação detalhada.

[1] https://github.com/acarolinabessa/initial-petition-structure-identification.git

5 RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da segmentação de petições iniciais, considerando tanto as anotações manuais realizadas pelos pesquisadores quanto a segmentação automática gerada pela ferramenta desenvolvida. A avaliação foi conduzida comparando os trechos identificados por ambos os métodos, analisando a correspondência entre os segmentos extraídos, utilizando as métricas Precisão, *Recall* e F1-*score*.

5.1 Anotações Manuais

Para avaliar a precisão da ferramenta de segmentação, foi necessário construir um conjunto de dados anotado manualmente, servindo como referência para a validação dos trechos extraídos automaticamente. Esse processo envolveu a marcação de 100 petições iniciais por pesquisadores do LINCProg, utilizando o aplicativo *Adobe Acrobat*. Durante a anotação, foram identificadas e rotuladas as seções FTP com os seguintes rótulos: “início fato”, “fim fato”, “início tese”, “fim tese”, “início pedido” e “fim pedido”, conforme anotação da petição inicial de exemplo na Figura 6.

Figura 6 – Petição inicial de exemplo anotada manualmente

Fonte: De autoria própria.

No entanto, esse processo de anotação manual apresentou desafios. O *Adobe Acrobat*,

apesar de ser uma ferramenta utilizada para manipulação de documentos em formato PDF, não foi desenvolvido especificamente para a anotação estrutural de textos, o que dificultou a marcação dos trechos dentro das petições. Algumas dificuldades enfrentadas incluíram a limitação da seleção de trechos extensos e a necessidade de repetição da marcação devido a falhas no reconhecimento. Além disso, a interface do *software* não foi otimizada para tarefas de anotação em larga escala, tornando o processo mais demorado.

Após a etapa de anotação, foi necessário extrair essas marcações e organizá-las de forma estruturada. Para isso, os arquivos PDF foram processados utilizando a biblioteca *PyMuPDF*, permitindo a leitura dos documentos e a recuperação do conteúdo. Inicialmente, os arquivos foram carregados a partir de um arquivo compactado (ZIP) contendo as petições anotadas, sendo extraídos para uma lista que armazenava os documentos em formato manipulável.

A extração das anotações foi realizada percorrendo cada página dos documentos e identificando as marcações feitas pelos pesquisadores. Sempre que um rótulo correspondente às categorias anotadas era encontrado, seu índice dentro do texto do documento era registrado, possibilitando a posterior recuperação dos trechos delimitados.

Uma vez coletados os índices das anotações, foi implementada uma função para extrair os trechos textuais correspondentes. Essa função percorreu os índices registrados, recuperando os segmentos do texto entre os pares de marcações (exemplo: “início fato” - “fim fato”), associando cada trecho à sua respectiva categoria. A Figura 7 mostra o resultado da extração obtido para a petição de exemplo.

Figura 7 – Trechos extraídos da anotação


```
{'inicio fato - fim fato': 'No dia [data], o Autor celebrou um contrato de prestação de serviço com o Réu,
visando à execução de [descrição do serviço contratado]. No entanto, o serviço foi prestado de maneira
defeituosa e em desacordo com o estabelecido no contrato, causando prejuízos financeiros e transtornos ao Autor.
O Autor notificou o Réu extrajudicialmente por meio de correspondência enviada em [data], solicitando o reparo
do dano ou a devolução dos valores pagos, sem, contudo, obter resposta satisfatória. Diante da inércia do Réu e
do prejuízo sofrido, não restou alternativa ao Autor senão recorrer ao Judiciário para ver seus direitos
resguardados.',
 'inicio tese - fim tese': 'O presente caso encontra amparo no artigo 186 do Código Civil, que dispõe: "Aquele
que, por ação ou omissão voluntária, negligência ou imprudência, violar direito e causar dano a outrem, ainda
que exclusivamente moral, comete ato ilícito." Além disso, o artigo 14 do Código de Defesa do Consumidor
estabelece a responsabilidade objetiva do fornecedor de serviços pelos danos causados ao consumidor por defeitos
na prestação do serviço, independentemente da existência de culpa. No caso concreto, o Réu descumpriu a
obrigação contratual, causando prejuízos materiais e morais ao Autor, que devem ser devidamente reparados.',
 'inicio pedido - fim pedido': 'Diante do exposto, requer o Autor: 1. A concessão de tutela antecipada,
determinando que o Réu efetue o reparo do serviço ou reembolse o valor pago no prazo de [XX] dias, sob pena de
multa diária a ser fixada por este juízo; 2. A citação do Réu, para, querendo, apresentar contestação, sob pena
de revelia e confissão quanto à matéria de fato; 3. A condenação do Réu ao pagamento de danos materiais, no
valor de R$ [XXX,XX], devidamente corrigidos e acrescidos de juros desde a data do evento danoso; 4. A
condenação do Réu ao pagamento de danos morais, no valor de R$ [XXX,XX], a ser arbitrado por este juízo,
considerando o abalo sofrido pelo Autor; 5. A condenação do Réu ao pagamento das custas processuais e honorários
advocatícios, nos termos do artigo 85 do CPC. Protesta provar o alegado por todos os meios de prova admitidos em
direito, especialmente documental, testemunhal e pericial. Dá-se à causa o valor de R$ [XXX,XX]'}
```


Fonte: De autoria própria.

Como resultado, foi gerado um conjunto de textos devidamente segmentados para comparação com a segmentação automática realizada pela ferramenta desenvolvida.

5.2 Anotações Automáticas

A ferramenta foi desenvolvida utilizando RegEx para detectar padrões textuais que delimitam cada seção dentro das petições. O código foi estruturado para processar os documentos de maneira iterativa, aplicando regras pré-definidas e extraindo os segmentos correspondentes. O fluxo da segmentação seguiu quatro etapas principais:

1. **Carregamento dos dados:** As petições foram lidas a partir de um arquivo CSV contendo os textos brutos;
2. **Pré-processamento:** Aplicação de técnicas de normalização, como remoção de caracteres especiais, conversão para minúsculas e eliminação de espaços excessivos;
3. **Segmentação automática:** Utilização de expressões regulares para identificar os pontos de início e fim das seções Fato, Tese e Pedido;
4. **Armazenamento dos resultados:** Os trechos extraídos foram organizados e armazenados para análise posterior.

Na implementação da lógica, a estratégia adotada para identificar os trechos das petições foi baseada na detecção de padrões textuais específicos, utilizando RegEx para localizar os pontos de início e fim de cada seção. A Figura 8 ilustra as expressões regulares aplicadas na ferramenta.

Figura 8 – Expressões regulares definidas

```
fato   = re.search(r'dos fatos(.*?)(do direito|dos pedidos|$)', text, re.IGNORECASE)
tese   = re.search(r'do direito(.*?)(dos pedidos|$)', text, re.IGNORECASE)
pedido = re.search(r'dos pedidos(.*)', text, re.IGNORECASE)
```

Fonte: De autoria própria.

Conforme apresentado na Figura 8, a regra utilizada para a segmentação das petições iniciais segue um padrão estruturado para identificar as principais seções do documento. Para a seção **Fato**, a extração inicia a partir do marcador “DOS FATOS” e se estende até a ocorrência da próxima seção identificável. No caso da seção **Tese**, o padrão utilizado reconhece “DO DIREITO” como delimitador inicial, capturando o conteúdo até a transição para “DOS PEDIDOS”. Por fim, a seção **Pedido** é extraída a partir da ocorrência de “DOS PEDIDOS” até o final do documento.

O resultado da segmentação para um exemplo de petição inicial é apresentado na Tabela 3, onde é possível visualizar os trechos extraídos pela ferramenta.

Tabela 3 – Resultado da segmentação da ferramenta

Fato	**Tese**	**Pedido**
no dia data o autor celebrou um contrato de prestação de serviço com o réu visando à execução de descrição do serviço contratado no entanto o serviço foi prestado de maneira defeituosa e em desacordo com o estabelecido no contrato causando prejuízos financeiros e transtornos ao autor o autor notificou o réu extrajudicialmente por meio de correspondência enviada em data solicitando o reparo do dano ou a devolução dos valores pagos sem contudo obter resposta satisfatória diante da inércia do réu e do prejuízo sofrido não restou alternativa ao autor senão recorrer ao judiciário para ver seus direitos resguardados	o presente caso encontra amparo no artigo 186 do código civil que dispõe aquele que por ação ou omissão voluntária negligência ou imprudência violar direito e causar dano a outrem ainda que exclusivamente moral comete ato ilícito além disso o artigo 14 do código de defesa do consumidor estabelece a responsabilidade objetiva do fornecedor de serviços pelos danos causados ao consumidor por defeitos na prestação do serviço independentemente da existência de culpa no caso concreto o réu descumpriu a obrigação contratual causando prejuízos materiais e morais ao autor que devem ser devidamente reparados	diante do exposto requer o autor a concessão de tutela antecipada determinando que o réu efetue o reparo do serviço ou reembolse o valor pago no prazo de xx dias sob pena de multa diária a ser fixada por este juízo a citação do réu para querendo apresentar contestação sob pena de revelia e confissão quanto à matéria de fato a condenação do réu ao pagamento de danos materiais no valor de r xxxxx devidamente corrigidos e acrescidos de juros desde a data do evento danoso a condenação do réu ao pagamento de danos morais no valor de r xxxxx a ser arbitrado por este juízo considerando o abalo sofrido pelo autor a condenação do réu ao pagamento das custas processuais e honorários advocatícios nos termos do artigo 85 do cpc protesta provar o alegado por todos os meios de prova admitidos em direito especialmente documental testemunhal e pericial dáse à causa o valor de r xxxxx reais

Fonte: De autoria própria.

Avaliando os resultados gerados, observou-se que a ferramenta apresentou uma segmentação consistente nos casos em que as petições seguem um formato padronizado. No entanto, em alguns casos, ocorreram falhas na delimitação correta das seções, especialmente quando os documentos possuíam variações na estrutura textual, como, por exemplo, quando no lugar da seção "DO DIREITO" é usada a nomenclatura "FUNDAMENTAÇÃO JURÍDICA" e assim por diante.

5.3 Avaliação dos Resultados

Nesta seção, são apresentados e analisados os resultados obtidos a partir da comparação entre as anotações manuais e a segmentação automática realizada pela ferramenta desenvolvida. Para avaliar a correspondência entre os segmentos extraídos automaticamente e manualmente, foram utilizadas duas abordagens de avaliação: Similaridade de Jaccard e as métricas Precisão, *Recall* e F1-*score*.

5.3.1 Avaliação com a Similaridade de Jaccard

A Similaridade de Jaccard foi utilizada para medir o grau de sobreposição entre os segmentos extraídos automaticamente e aqueles anotados manualmente. Essa métrica quantifica a interseção entre os conjuntos de palavras extraídas em relação à união desses conjuntos. O resultado pode ser visto na Tabela 4.

Tabela 4 – Resultado da comparação utilizando o método de similaridade de Jaccard

Seção	Similaridade de Jaccard (%)
Fato	50,12%
Tese	48,16%
Pedido	47,07%

Fonte: De autoria própria.

Os valores médios obtidos indicam que a ferramenta apresentou uma similaridade moderada em todas as seções analisadas. Conforme apresentado na Tabela 4, a seção Fato obteve um coeficiente médio de 0,50, seguida pela seção Tese com 0,48, e pela seção Pedido, que apresentou a menor similaridade, com 0,47. Esses resultados apontam que, embora a segmentação automática tenha conseguido capturar alguns trechos corretos, ainda há variações na correspondência entre as seções identificados manualmente e automaticamente.

Portanto, esta avaliação revelou que a seção Pedido apresentou maior variação, o que pode ser atribuído à variação na estruturação dessa parte da petição. Além disso, verificou-se que algumas petições não tiveram seus segmentos extraídos corretamente pela ferramenta, resultando em conteúdos vazios. Esse problema pode estar relacionado a falhas na extração das anotações manuais utilizando o *PyMuPDF*, o que afetou a avaliação em determinados casos.

5.3.2 Avaliação com Precisão, *Recall* e F1-*score*

Além da Similaridade de Jaccard, foram aplicadas as métricas de Precisão, *Recall* e F1-*score* para uma avaliação mais detalhada do desempenho da ferramenta. Os resultados dessas métricas são apresentados na Tabela 5.

Tabela 5 – Resultado da comparação utilizando as métricas de Precisão, *Recall* e F1-*score*

Seção	**Precisão**	***Recall***	**F1-*score***
Fato	61,28%	70,64%	63,45%
Tese	61,38%	62,55%	61,21%
Pedido	59,38%	61,69%	59,72%

Fonte: De autoria própria.

De acordo com a Tabela 5, os valores médios obtidos mostram que a ferramenta apresentou um desempenho razoável na segmentação das três seções Fato, Tese e Pedido. A segmentação da seção Fato obteve a maior taxa de *recall* (0,70), indicando que a ferramenta conseguiu identificar uma grande quantidade de palavras presentes nas anotações manuais. No entanto, a precisão ficou em 0,61, sugerindo que, apesar de identificar corretamente os trechos esperados, a ferramenta também extraiu algumas informações irrelevantes.

Já para a seção Tese, observou-se um equilíbrio entre precisão (0,61) e *recall* (0,62), resultando em um F1-score de 0,61. Isso mostra que ferramenta conseguiu segmentar essa parte do texto de forma mais consistente, embora ainda existam erros devido à variação na nomenclatura e na estrutura desse trecho dentro das petições.

A extração da seção Pedido, por sua vez, apresentou a menor precisão dentre as três categorias (0,59), acompanhada de um *recall* de 0,61. Esse resultado está associado à variedade de formatos nos quais os pedidos são estruturados dentro das petições, tornando o reconhecimento automático dessa seção mais complexo.

6 CONSIDERAÇÕES FINAIS

A aplicação de técnicas de PLN no âmbito jurídico demonstra um grande potencial para otimizar o funcionamento do sistema judicial brasileiro, que constantemente enfrenta desafios relacionados ao elevado volume de processos pendentes. A digitalização e a automação de tarefas repetitivas podem contribuir significativamente para a redução do tempo de análise documental, facilitando a organização e o processamento das petições iniciais. O desenvolvimento de ferramentas especializadas voltadas para esse domínio visa aprimorar a triagem de documentos jurídicos, possibilitando uma análise mais estruturada e ágil, ao mesmo tempo em que leva em consideração as particularidades e terminologias próprias dos textos legais.

Diante dessa necessidade, este trabalho propôs a criação de uma ferramenta para a segmentação automática das petições iniciais, classificando-as em suas três principais seções: Fato, Tese e Pedido. A ferramenta desenvolvida utilizou RegEx para identificar e extrair automaticamente os segmentos correspondentes, buscando minimizar a necessidade de intervenção manual no processo de organização das petições.

Os resultados obtidos demonstraram que a segmentação automática pode contribuir para a redução do tempo de análise e organização dos documentos jurídicos, respondendo assim à primeira questão de pesquisa. A ferramenta apresentou um desempenho razoável nas métricas de Similaridade de Jaccard, Precisão, *Recall* e F1-*score*, indicando que a segmentação automática conseguiu capturar parte dos trechos corretamente, mas ainda requer ajustes para alcançar maior precisão. A seção Fato apresentou melhor *Recall*, sugerindo que a ferramenta conseguiu identificar boa parte das palavras pertencentes a essa categoria, mas a precisão da segmentação foi afetada por ruídos textuais e variações na nomenclatura utilizada nas petições.

Com relação aos desafios da segmentação automática de textos jurídicos estruturados, observou-se que as principais dificuldades são em relação a variação das petições em termos de nomenclatura e organização textual. Algumas petições utilizam marcadores diferentes para introduzir cada seção, o que impactou diretamente o reconhecimento automático dos trechos. Além disso, a presença de trechos extensos e a ausência de delimitações padronizadas dificultaram a correta extração das seções, reforçando a necessidade de métodos mais flexíveis para lidar com essas variações.

A comparação entre a segmentação automática e a manual mostrou que, embora a ferramenta tenha sido capaz de identificar corretamente diversos trechos, algumas petições não tiveram seus segmentos extraídos com êxito, resultando em conteúdos vazios. Esse problema esteve relacionado a falhas na extração das anotações manuais utilizando a biblioteca *PyMuPDF*, necessitando de aprimoramento na obtenção dos dados de referência para a avaliação da ferramenta.

Por fim, a ferramenta desenvolvida representa um avanço na automação da análise de petições iniciais, contribuindo para a organização e padronização desses documentos. No entanto, os resultados indicam que a ferramenta pode ser aprimorada com abordagens híbridas que combinem RegEx com aprendizado de máquina, aumentando sua adaptabilidade a diferentes formatos de petições. Além disso, a expansão do conjunto de dados utilizado para treinamento e teste da ferramenta pode proporcionar uma avaliação mais abrangente de seu desempenho.

6.1 Trabalhos Futuros

Este trabalho contribui para a área de automação jurídica, oferecendo uma solução para auxiliar na organização e análise de petições iniciais. A ferramenta pode ser utilizada como um ponto de partida para o desenvolvimento de sistemas mais avançados, que combinem diferentes técnicas de PLN e aprendizado de máquina. Pesquisas futuras podem abordar:

- Explorar o uso de modelos de linguagem pré-treinados, como BERT e suas variações, para melhorar a identificação dos padrões textuais e a segmentação das petições;
- Ampliar o conjunto de dados de treinamento e avaliação, incluindo uma variedade maior de tipos de petições e estilos de escrita;
- Melhorar as expressões regulares utilizadas na ferramenta, adaptando-as para lidar com as variações na estrutura das petições;
- Desenvolver uma interface de usuário intuitiva para facilitar o uso da ferramenta por profissionais do direito;
- Realizar testes com usuários reais para avaliar a usabilidade e o impacto da ferramenta no fluxo de trabalho dos tribunais.

Referências

ABDUSALOMOVNA, T. D. et al. Text mining. *European Journal of Interdisciplinary Research and Development*, v. 13, p. 284–289, 2023. Citado na página 18.

AMARAL, F. F. do. Justiça digital: O papel da tecnologia no sistema jurídico moderno. *Revista Ilustração*, v. 5, n. 6, p. 3–25, 2024. Citado na página 14.

BEZERRA, T. L.; CORADESQUE, F. A. A.; LACERDA, M. G.; LIMA, J. V. S. A eficácia da segmentação por crescimento de regiões na identificação de pistas de pouso não pavimentadas medida pelo índice jaccard. *Revista do CIAAR*, v. 2, n. 1, 2021. Citado na página 27.

BIZARRO, J. P. S. M. A petição inicial. *Revista Española de Derecho Canónico*, v. 79, n. 192, p. 325–356, 2022. Citado 2 vezes nas páginas 14 e 18.

BRAGANÇA, F.; BRAGANÇA, L. F. da F. Revolução 4.0 no poder judiciário: levantamento do uso de inteligência artificial nos tribunais brasileiros. *Revista da Seção Judiciária do Rio de Janeiro*, v. 23, n. 46, p. 65–76, 2019. Citado na página 14.

Brasil. *Código de Processo Civil*. Brasília, DF: Planalto, 2015. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm>. Acesso em: 05 set. 2024. Citado na página 17.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. Crisp-dm 1.0: Step-by-step data mining guide. In: . [s.n.], 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:59777418>. Citado na página 22.

CHEN, Q.; BANERJEE, A.; DEMIRALP, Ç.; DURRETT, G.; DILLIG, I. Data extraction via semantic regular expression synthesis. *Proceedings of the ACM on Programming Languages*, ACM New York, NY, USA, v. 7, n. OOPSLA2, p. 1848–1877, 2023. Citado na página 18.

CNJ. *Justiça 4.0*. 2021. Disponível em: https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/. Citado na página 14.

CONSTANTINO, K.; CRUZ, V. A. L.; ZUCHERATTO, O. M.; FRANÇA, C.; CARVALHO, M.; SILVA, T. H.; LAENDER, A. H.; GONÇALVES, M. A. Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In: SBC. *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*. [S.l.], 2022. p. 304–316. Citado 2 vezes nas páginas 20 e 21.

GOMES, L. M.; SÁ, J. M. C. de; PENG, Y. *Línguas naturais e máquinas artificiais: Aplicação de técnicas de mineração de texto para a classificação de sentenças judiciais brasileiras*. [S.l.], 2020. Citado 2 vezes nas páginas 20 e 21.

GRANCHAROVA, M.; JANGEFALK, M. *Comparative study of the combined performance of learning algorithms and preprocessing techniques for text classification*. 2018. Citado na página 26.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEGI, M. R. Text mining in big data analytics. *Big Data and Cognitive Computing*, MDPI, v. 4, n. 1, p. 1, 2020. Citado na página 18.

JARGAS, A. M. *Expressões Regulares-5a edição: Uma Abordagem Divertida*. [S.l.]: Novatec Editora, 2016. Citado na página 19.

JÚNIOR, A. P. de C.; CALIXTO, W. P.; CASTRO, C. H. A. de. 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, 2020. Citado 3 vezes nas páginas 18, 20 e 21.

JUSBRASIL. *Modelo de petição inicial conforme o Novo CPC*. 2016. Acesso em: 11 nov. 2024. Disponível em: <https://www.jusbrasil.com.br/modelos-pecas/modelo-de-peticao-inicial-conforme-o-novo-cpc/390816480>. Citado na página 17.

LEITE, T. M. A. d. A. Busca por maior eficiência na execução fiscal: análise do tema nº 1.184 do supremo tribunal federal. Centro Universitário do Rio Grande do Norte, 2024. Citado na página 23.

LEMOS, D. L. d. S.; COELHO, A. Qualidade de dados em acervos do patrimônio cultural: uma avaliação diagnóstica semiautomática nos objetos culturais sob gestão do instituto brasileiro de museus. *Encontros Bibli*, SciELO Brasil, v. 28, p. e90510, 2023. Citado na página 19.

OLIVEIRA, L. V. S.; TAUCHERT, M. R.; MIRANDA, T. A.; SIQUEIRA, R.; SOUZA, R. X. de et al. Big data e data science: Um estudo sobre a nova visão jurídica tecnológica do século xxi. *Facit Business and Technology Journal*, v. 1, n. 33, 2022. Citado na página 14.

PORTO, F. R. A "corrida maluca" da inteligência artificial no poder judiciário. *Inteligência artificial e aplicabilidade prática no direito. Brasília: Conselho Nacional de Justiça*, p. 103–130, 2022. Citado na página 14.

QUEIROZ, A. M. de; BUENO, P. L. N.; LISBINO, J. K. T. O impacto da inteligência artificial na advocacia brasileira: Benefícios e desafios no setor jurídico. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, v. 10, n. 11, p. 2697–2712, 2024. Citado na página 14.

ROCHA, A. C. P. Mineração de textos para classificação de processos judiciais trabalhistas. 2020. Citado na página 18.

SANTOS, N. H. S. dos; ALENCAR, V. F.; TAUCHERT, M. R.; JÚNIOR, W. O. C.; SOUZA, R. X. de; SIQUEIRA, R. Aplicação do big data e data science no âmbito jurídico. *Facit Business and Technology Journal*, v. 1, n. 33, 2022. Citado na página 14.

SCHNEIDER, A. F.; MOREIRA, A. C. S. A justiça 4.0 como ferramenta de eficiência: O caso ambiental. *Revista da EMERJ*, v. 25, n. 2, p. 22–30, 2023. Citado na página 14.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000. Citado na página 22.

SILVA, M. V. da; SANTANA, E. E.; LOBATO, F. M.; JR, A. F. J. Preprocessing applied to legal text mining: analysis and evaluation of the main techniques used. In: SBC. *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2023. p. 1010–1021. Citado na página 18.

SOUZA, M. P. R. d. Qualidade de dados dentro do contexto de big data: uma revisão global. 2023. Citado na página 14.

TATA, S.; POTTI, N.; WENDT, J. B.; COSTA, L. B.; NAJORK, M.; GUNEL, B. Glean: Structured extractions from templatic documents. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 14, n. 6, p. 997–1005, 2021. Citado 2 vezes nas páginas 20 e 21.

TUROŇOVÁ, L.; HOLÍK, L.; LENGÁL, O.; SAARIKIVI, O.; VEANES, M.; VOJNAR, T. Regex matching with counting-set automata. *Proceedings of the ACM on Programming Languages*, ACM New York, NY, USA, v. 4, n. OOPSLA, p. 1–30, 2020. Citado na página 19.