CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC (RES_020/2016 – 2024_2)	
( ) PRÉ-PROJETO (X) PROJETO	ANO/SEMESTRE: 2024/2

DESENVOLVIMENTO DE CHATBOT ESPECIALISTA EM LEGISLAÇÕES EXTRAJUDICIAIS PARA CARTÓRIOS APLICANDO TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Pedro Schumann

Prof.ª Andreza Sartori – Orientadora

1 INTRODUÇÃO

Segundo César (2019), os Cartórios do Brasil se originaram do período colonial, logo no início da colonização com as capitanias hereditárias, no qual os proprietários herdeiros das terras foram incumbidos de nomear Tabeliães. Em 1988, com o advento da Constituição Federal, ocorreu a maior mudança nas atividades cartorárias, formando serventias extrajudiciais que hoje servem de inspiração para diversos Países, devido ao grau de organização, celeridade e eficiência que alcançaram, o que levou os cartórios a serem considerados pela população a instituição mais confiável do País. Por outro lado, o cenário atual do Poder Judiciário brasileiro é caótico, completamente assoberbado com um número excessivo de processos judiciais, sendo a busca pela solução da morosidade um dos ideais perseguidos pela comunidade jurídica, para que, dessa forma, o ideal de justiça possa ser verdadeiramente perseguido. Na tentativa de solucionar a lentidão, serviços antes exclusivos do Poder Judiciário estão sendo direcionados especialmente para a atividade cartorial (Lucchesi; Teotonio; Carlucci, 2014).

A crescente demanda evidencia a importância dos cartórios para o desenvolvimento e para a segurança dos atos civis. Tais atos acompanham a população desde o nascimento até a morte, passando pela emancipação, se houver, aquisição do primeiro carro, casa, casamento, divórcio, partilha, doação e inventário (Moreira, 2022), além de práticas de conciliação e mediações, realizadas a fim de desafogar o poder judiciário. O nosso ordenamento jurídico atual comporta 5 (cinco) especialidades de serventias extrajudiciais, cada uma delas com suas atribuições, sendo essas atribuições totalmente distintas, não devendo confundi-las (César, 2019). Para cada atribuição existem legislações próprias a serem seguidas, além de emolumentos e taxas diferentes, as quais são determinadas pelo estado.

Mediante a todos os desafios enfrentados pelos cartórios, ofertar serviços de forma online se tornou necessário, principalmente pelo advento da pandemia da COVID-19, que mostrou ao mundo inteiro a possibilidade de facilitar serviços do cotidiano, dessa forma, levando o Governo Federal a anunciar a reformulação do que muitos estão chamando de Cartório Digital (Moreira, 2022). Tendo em vista as adaptações realizadas com o objetivo de tornar os serviços mais ágeis e eficazes, surge a aplicação das tecnologias atuais, com foco principal no uso de Inteligência Artificial (IA), popularizadas por ferramentas de IA generativa como o ChatGPT, pelo seu potencial tecnológico. No primeiro semestre de 2023, o ChatGPT destacou-se pela sua versatilidade em tarefas relacionadas ao Processamento de Linguagem Natural (PLN), revelando-se ser eficaz na compreensão da linguagem humana e realizando tarefas como síntese de textos, tradução e busca de informações. Esse fenômeno trouxe à tona a relevância da IA em diversos setores, inclusive o jurídico (Baumbach; Trindade, 2023).

De acordo com Baumbach e Trindade (2023), o avanço da IA é exponencial, qualitativa e quantitativamente. No âmbito jurídico existem diversos sistemas que desempenham um papel importante, classificando recursos, localizando peças processuais, resumindo e indexando autos inteiros, prognosticando o resultado de julgamentos, assessorando em situações negociais, entre outras aplicações. Entretanto, as serventias extrajudiciais ainda não foram tocadas pela IA, à exceção dos efeitos produzidos pelas aplicações genéricas, como *chatbots*, IA aplicada nas redes sociais, robôs de pesquisa, geração de texto, entre outras aplicações. Várias razões devem motivar o quadro, a exemplo da restrita oferta de tecnologia especializada e seu alto custo, o pequeno tamanho, a inclinação autárquica e ainda a sobrecarga na agenda institucional das serventias. No entanto, conforme afirmação de Baumbach e Trindade (2023), dado esse cenário geral é razoável prever que, em cinco ou dez anos, teremos em operação robôs inteligentes especificamente desenvolvidos para realizar ou apoiar a realização de atividades cartoriais, atos do ofício ou tarefas a eles intimamente ligadas.

Diante deste cenário, este trabalho visa responder a seguinte pergunta de pesquisa: “Como disponibilizar uma ferramenta que auxilie e otimize os processos cartorários?”. Dito isso, este trabalho visa disponibilizar um canal único de consulta a legislações que envolvem procedimentos cartorários. A fim de facilitar a consulta e otimizar a prática diária de atos, a disponibilização será por meio de um *chatbot* especialista, implementado com uso de técnicas de Processamento de Linguagem Natural, aplicando Web Scraping, modelos de linguagem de grande escala (Large Language Models - LLMs) de código aberto e Inteligência Artificial. Dessa forma, é possível atender uma das principais dificuldades dos cartórios, que seria o acesso rápido e confiável às informações presentes em legislação, otimizando e facilitando o serviço prestado à população.

1.1 OBJETIVOS

O objetivo principal desse trabalho é disponibilizar um *chatbot* para responder dúvidas do setor extrajudicial, voltado aos cartórios, utilizando técnicas de Processamento de Linguagem Natural (PLN).

Os objetivos específicos são:

a) identificar e extrair, por meio de técnicas de Web Scraping, as legislações encontradas nos principais canais web de divulgação extrajudiciais, como o Tribunal de Justiça de Santa Catarina (TJSC), que disponibiliza o Código de Normas extrajudicial, além de legislações nacionais, como a Lei 6.015, que envolve a prática de atos do registro civil, dentre outros;
b) avaliar e definir o modelo de linguagem de grande escala (LLM) de código aberto com melhor custo-benefício, analisando o retorno obtido das respostas e o custo computacional demandado;
c) realizar o pré-processamento das legislações coletadas e treinar um agente para localizá-las;
d) identificar as técnicas de Processamento de Linguagem Natural (PLN) e engenharia de *prompt* que auxiliem na associação das perguntas dos usuários ao assunto específico, a fim de gerar retornos satisfatórios.

2 TRABALHOS CORRELATOS

Nesta seção serão apresentados os trabalhos correlatos utilizados como base para o desenvolvimento do trabalho proposto. Para garantir informações mais atuais, a pesquisa abrangeu o período de 2020 a 2024. Foi utilizando o Google Acadêmico como fonte de consulta, devido à sua ampla e diversificada coleção de artigos acadêmicos e revistas relacionadas à área de ciências da computação, especialmente para aqueles voltados aos temas do presente trabalho.

2.1 REVISÃO SISTEMÁTICA

Os trabalhos selecionados foram localizados utilizando combinações dos seguintes termos de busca: “Processamento de linguagem natural”, “Linguagem natural”, “jurídico”, “PLN” (abreviação de Processamento de Linguagem Natural), “Llama 2”, “Web Scraping”, “PDF”, “chatbot” e “especialista”. O termo “Llama 2”, utilizado na busca, se refere a uma LLM e foi incluído devido ao conhecimento do autor sobre o mesmo. No Quadro 1 são apresentadas as sínteses dos trabalhos correlatos selecionados que fundamentarão o desenvolvimento do trabalho proposto.

Os trabalhos apresentados no Quadro 1 foram escolhidos por conta de sua similaridade com o trabalho proposto, trabalhando com técnicas de Processamento de Linguagem Natural (PLN), com métodos de busca de informações por meio de Web Scraping, buscando também por trabalhos que tivessem a implementação de um *chatbot*, especialmente para aqueles que utilizam tecnologias paralelas ao ChatGPT, como a LLM de código aberto Llama 2. Com relação aos métodos de busca por meio de Web Scraping, buscou-se principalmente soluções de extração de dados em sites com dados desestruturados, ou que possuem algum nível de estrutura, mas apresentam inconsistências ou variações que dificultam uma análise automatizada direta, o que se assemelha ao presente trabalho. Além disso, buscou-se encontrar soluções para a área jurídica, base para o trabalho proposto, esclarecendo os desafios e soluções desse nicho, que possui suas peculiaridades e características próprias.

Quadro 1 - Síntese dos trabalhos correlatos selecionados

Assunto	Filtro	Referência
Classificação de documentos aplicando Processamento da Linguagem Natural juntamente com técnicas de machine learning	"Linguagem natural" AND "jurídico"	Silva e Ribeiro (2023)
Classificação de texto jurídicos para reconhecimento de entidades nomeadas por meio de Processamento de Linguagem Natural	“jurídico” AND (“PLN” OR “Linguagem natural”)	Rodríguez e Bezerra (2020)
Desenvolvimento de um *chatbot* para tirar dúvidas sobre medicamentos, com base em informações coletadas por meio de Web Scraping e tratadas com técnicas de Processamento de Linguagem Natural	“Llama 2” AND "Linguagem natural"	Lin (2023)
Web Scraping em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte, disponibilizando os dados coletados por meio de um *chatbot*	"Web Scraping" AND "PDF"	Assis (2021)
Desenvolvimento de um *chatbot* especialista, voltado para perguntas frequentes de usuários da biblioteca sobre a lei de direito autoral	"chatbot" AND "especialista"	Santiago (2022)

Fonte: elaborado pelo autor (2024).

2.2 SÍNTESE DOS TRABALHOS CORRELATOS

Nesta subseção será apresentado trabalhos com características semelhantes aos principais objetivos do estudo proposto. O trabalho de Lin (2023) no Quadro 2, apresenta a implementação de um *chatbot* para sanar dúvidas referentes a utilização de medicamentos. Assis (2021), apresentado no Quadro 3, descreve uma aplicação que realiza buscas em dados governamentais. Por fim, o trabalho de Rodríguez e Bezerra (2020) apresentado no Quadro 4, se trata de um reconhecimento de entidades nomeadas em textos jurídicos.

Quadro 2 –A utilização de modelos de LLM para geração de bulas automatizadas

Referência	Lin (2023)
Objetivos	Desenvolver uma aplicação baseada em Large Language Models (LLM) com finalidade de responder às perguntas dos usuários sobre medicamentos, gerando informações essenciais, como: efeitos colaterais, contraindicações, composição, posologia e outros dados relevantes para que possam tirar dúvidas dos usuários. Realizar uma análise comparativa de desempenho de modelos de código aberto, como *falcon* e *llama* e de modelos de código fechado, como *gpt-3.5-turbo*.
Principais funcionalidades	Geração de respostas precisas e confiáveis sobre informações relacionadas a medicamentos. É realizado um estudo em diversos modelos de linguagem, como Falcon, Llama 2 e GPT, a fim de verificar o mais eficiente em termos de custo computacional e resposta fornecida.
Ferramentas de desenvolvimento	Foram utilizadas técnicas de Scraping, realização de *embedding* e *chunking* (técnicas de divisão e representação do texto), com armazenamentos dessas representações numéricas em um banco de dados de vetores. Foi utilizada a biblioteca FAISS e modelos de LLM, como Falcon, Llama 2 e GPT, para formulação de respostas condizentes. Como métricas de qualidade de linguagem foi feito uso de BLEU Score e Rouge Score.
Resultados e conclusões	O LLM que performou melhor foi o da OpenAI, elaborando textos mais coerentes e relevantes. Entretanto, dos modelos de código aberto, o modelo falcon-7b teve também o seu resultado expressivo. Em algumas situações, os resultados gerados superaram até os modelos do OpenAI.

Fonte: elaborado pelo autor (2024).

O foco do trabalho de Lin (2023) se dá no desenvolvimento de um *chatbot* especialista em tirar dúvidas relacionadas a medicamentos. A abordagem a um tema específico, com extração de dados sendo feita por meio de Scraping é de grande relevância para o estudo do trabalho proposto, visto que, um dos principais objetivos é a geração de respostas concisas e satisfatórias. O uso de diversos modelos de linguagem, ferramentas de código aberto e métricas de avaliação aplicadas para análise de desempenho de cada uma dessas ferramentas, elucidam sobre o caminho a ser seguido.

Quadro 3 –Chat Bot Sumé: Web Scraping em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte

Referência	Assis (2021)
Objetivos	Desenvolver um método de Web Scraping capaz de transformar os dados desestruturados do portal de transparência, da Câmara Municipal de Belo Horizonte, em dados abertos. Além disso, desenvolver um protótipo de *chatbot* baseado nos resultados obtidos da extração de dados.
Principais funcionalidades	Localiza todos os vereadores da Câmara Municipal de Belo Horizonte, seu partido e seus respectivos gastos, utilizando técnicas de Web Scraping, realizando uma somatória das despesas e transformando tais informações em dados abertos e acessíveis, os quais podem ser acessados por meio de um *chatbot* implementado pela ferramenta *Dialogflow* da Google.
Ferramentas de desenvolvimento	Uso majoritário de técnicas de Web Scraping. O desenvolvimento se deu pela utilização da linguagem Python e do ambiente de desenvolvimento integrado (IDE) Pycharm. Foi aplicada a API Selenium Webdriver, que interage com o navegador para extrair informações, em conjunto com a biblioteca Beatiful Soup, para criar o método Web Scraping e extrair as informações de arquivos HTML e XML, além do uso do módulo Time para controle da extração de dados e da biblioteca Pandas para armazenar e gerenciar o DataFrame gerado. Para a implementação do *chatbot* foi utilizado a plataforma Dialogflow da Google, que facilita o design e a integração de uma interface do usuário conversacional com outros aplicativos.
Resultados e conclusões	Foi comprovada a eficácia do Web Scraping como método adotado para a extração de dados desestruturados, seguida de manipulação para transformá-los em dados abertos. Por meio da implementação do *Chat Bot Sumé*, foi possível identificar irregularidades referentes aos dados analisados na Câmara Municipal de Belo Horizonte (CMBH), onde ao somar os valores de gastos do custeio parlamentar e comparar com os dados coletados e salvos pela aplicação referente a meses anteriores, obteve-se uma divergência de aproximadamente R$ 194.489,24 (Cento e noventa e quatro mil e quatrocentos e oitenta e nove reais e vinte e quatro centavos), referente a gastos excluídos do portal de transparência sem aviso.

Fonte: elaborado pelo autor (2024).

O trabalho de Assis (2021) se aprofunda, em sua maior parte, na extração de dados utilizando técnicas de Web Scraping, além de realizar a implementação de um *chatbot* para disponibilizar dados coletados da Câmara Municipal de Belo Horizonte, referente aos gastos dos parlamentares. O uso de ferramentas de desenvolvimento como a API Selenium Webdriver em conjunto com a biblioteca Beatiful Soup otimiza a raspagem dos dados das páginas, inclusive de dados desestruturados, se assemelhando com o objetivo do trabalho proposto. O

processamento e associações realizadas com os dados obtidos é desenvolvido de forma a entregar dados abertos de fácil compreensão, principalmente pela implementação do Chat Bot Sumé, por meio da plataforma Dialogflow, que possibilita a consulta de tais informações.

Quadro 4 –Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias)

Referência	Rodríguez e Bezerra (2020)
Objetivos	Criar uma plataforma web para classificação de documentos, a fim de identificar entidades nomeadas mencionadas em textos jurídicos, como nomeações, exonerações e atos.
Principais funcionalidades	Implementados classificadores que identificam entidades nomeadas em textos jurídicos, sendo realizada a análise em 10.000 registros, utilizando técnicas de PLN, como o pré-processamento dos textos, a fim de gerar uma classificação precisa. Para isso foram criados tratamentos específicos, como um dicionário de palavras a serem desconsideradas, além de trabalhar com base nas características estruturais do texto, no qual pode-se descartar trechos referente a pessoas que assinaram/autorizaram as Portarias, localizados no final do texto.
Ferramentas de desenvolvimento	Uso de biblioteca como NLTK e Scikit-Learn para implementação dos classificadores. Foram usados os seguintes classificadores referentes ao NLTK: Decision Tree, Maximum Entropy (algoritmo iis e algoritmo gis). Os classificadores utilizados provenientes do Scikit-Learn foram estes: Logistic Regression, Naive Bayes Bernoulli, Naive Bayes Multinomial, Linear Support Vector Classification, Nu Support Vector Classification, e Support Vector Classification. Sendo utilizado técnicas de PLN para melhorar a eficiência dos classificadores.
Resultados e conclusões	Os classificadores com melhores resultados foram o Linear SVC, com 100% de acurácia, o Logistic Regression, também com 100% de acurácia e Naive Bayes Multinomial, com 98,8% de acurácia. Ambos com a transformação do texto para letras minúsculas e com remoção de non letters (pontuações e caracteres especiais) e stopwords (palavras consideradas irrelevantes). Após a conclusão do processamento de 10.000 registros, foi identificada uma assertividade de 92,9% dos nomes pessoais, percentual considerado bastante aceitável para o objetivo proposto. Para os trabalhos futuros, sugere-se explorar também a biblioteca Python de PLN, spaCy, já com um modelo pronto em português.

Fonte: elaborado pelo autor (2024).

O trabalho de Rodríguez e Bezerra (2020) realiza um estudo de classificação de textos por meio de técnicas de PLN, tendo como objetivo identificar entidades nomeadas em textos jurídicos de atos administrativos. A pesquisa aborda a importância das análises morfológicas, sintáticas e semânticas, necessárias para realizar a extração das características estruturais dos textos. O pré-processamento abordado, por meio da implementação de um dicionário e a remoção de trechos específicos, baseado na estrutura do texto é necessário para a classificação correta das entidades citadas nos textos. As diversas técnicas utilizadas podem servir como base à otimização da classificação no trabalho proposto, visto que ambos os estudos abordam uma grande variedade lexical, principalmente pela sua especificação na área jurídica e por conta do regionalismo, além da similaridade na estrutura do texto.

Com base no que foi exposto, o primeiro trabalho correlato escolhido, implementado por Lin (2023) apresenta a implementação de um *chatbot*, aplicando técnicas de Processamento de Linguagem Natural como Scraping de dados. O trabalho realizado por Assis (2021) se trata de uma aplicação que realiza uma busca, por meio de Web Scraping em dados governamentais, os quais são dispersos em diversos sites e possuem uma estrutura despadronizada, disponibilizando um *chatbot* para consulta de tais informações. O trabalho de Rodríguez e Bezerra (2020) foi escolhido por reconhecer entidades nomeadas em textos jurídicos, dessa forma, integrando conhecimentos em PLN para reconhecimento de padrões juntamente com textos jurídicos, o que pode ser adaptado para reconhecer e classificar legislações, que seria parte fundamental do proposto.

3 PROPOSTA DE PROTÓTIPO

Esta seção apresenta a justificativa para a implementação do protótipo, bem como a metodologia utilizada para o desenvolvimento do mesmo.

3.1 JUSTIFICATIVA

O Quadro 5 apresenta um comparativo entre os trabalhos correlatos detalhados na subseção 2.2, onde as linhas representam as características e as colunas os trabalhos correlatos.

Quadro 5 - Comparativo dos trabalhos correlatos

Características \ Trabalhos Correlatos	Lin (2023)	Assis (2021)	Rodríguez e Bezerra (2020)
Fonte dos dados	1.050 arquivos com informações de diferentes medicamentos obtidos do GitHub	Web Scraping de gastos parlamentares realizado na Câmara Municipal de Belo Horizonte	10.000 registros extraídos do Diário Oficial da União
Pré-processamento realizado	Técnicas de Scraping para limpeza de dados e remoção de ruídos, além de *chunking* e *embedding*	Dados já estruturados no momento da coleta, por meio da API Selenium Webdriver em conjunto com a biblioteca Beatiful Soup	Criação de um dicionário de exclusão, remoção de caracteres e acentuações, substituição de palavras maiúsculas para minúsculas e extração de trechos através de expressões regulares
Forma de apresentação dos resultados	Respostas geradas pelo *chatbot*	Relação de informações disponibilizada em formato de tabela e respostas geradas pelo *chatbot*	Arquivos de textos individuais com o nome de cada entidade localizada
Plataformas para o desenvolvimento de *chatbot*	HuggingFace e LangChain	Dialogflow	-
Métricas de avaliação	BLEU Score e ROUGE para avaliação das respostas do modelo	-	Uso da acurácia para avaliação dos classificadores, com assertividade de 92,9%

Fonte: elaborado pelo autor (2024).

Conforme apresentado no Quadro 5, os autores utilizaram diferentes métodos de coleta de dados. Lin (2023) obteve-os no GitHub, enquanto Rodríguez e Bezerra (2020) extraiu sua base do Diário Oficial da União, sem a necessidade de utilizar técnicas de Web Scraping. Para Assis (2021), grande parte do trabalho se deu no desenvolvimento de técnicas de Web Scraping para raspagem de dados contidos no site da Câmara Municipal de Belo Horizonte. Os dados coletados por Assis (2022) são, em grande parte, semiestruturados, exigindo um processamento baseado nos rótulos presentes na estrutura das páginas da Web. Esse processo é semelhante ao do presente trabalho, cujo objetivo é coletar legislações de diversas fontes, cada uma com uma estrutura única para suas páginas, sem padrões predefinidos.

Os três autores utilizaram técnicas para filtrar e limpar os dados coletados, facilitando dessa forma a análise das informações e posteriormente, para Lin (2023) e Assis (2021), a implementação de um *chatbot*, onde o pré-processamento torna o treinamento mais rápido e as respostas mais precisas. Ademais, o uso de plataformas como HuggingFace, LangChain e Dialogflow demonstraram ser efetivas para implementação de um *chatbot*, demonstrando serem opções viáveis para utilização no trabalho proposto. Juntamente com a avaliação de diferentes modelos de linguagem realizada por Lin (2023) por meio de métricas de BLEU Score e ROUGE, pode-se assegurar a eficiência dos modelos de linguagem Falcon, Llama 2 e GPT.

Diante dessas características pode-se afirmar que, a extração e limpeza de dados realizada por meio de Web Scraping e técnicas de PLN, é possível, tornando acessível uma gama de informações em uma única base. Ao aplicar técnicas como embedding, realizando a criação de índices semânticos para cada dado coletado, pode-se obter buscas mais eficientes e melhores retornos por similaridade. A elaboração de uma plataforma que centralize todas as informações necessárias tem potencial para tornar o processo menos moroso, não havendo a necessidade de realizar diversas consultas para obter a informação desejada. A existência de diferentes plataformas para desenvolvimento de *chatbot*, torna conveniente sua utilização, por conta das facilidades empregadas, possibilitando para o presente trabalho contribuir, por meio dessas ferramentas, para as atividades cartorárias, com informações mais precisas e de fácil acesso. A raspagem de dados das principais fontes de legislações extrajudiciais disponível na internet, aliada ao pré-processamento dessas informações, de forma que as respostas geradas pelo *chatbot* sejam coerentes e satisfatórias, podem contribuir para a melhoria do serviço prestado à população.

3.2 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: realizar levantamento bibliográfico sobre os principais sites que contém as legislações extrajudiciais voltadas aos cartórios, técnicas e modelos de PLN, Web Scraping, *chatbot*, bem como trabalhos correlatos;
b) estudar tipos de dados: analisar os sites que contém as legislações a serem extraídas e estudar de que forma as informações podem ser coletadas, por meio de arquivos disponíveis e/ou estrutura HTML, conforme ocorre, por exemplo, com o Código de Normas extrajudicial, passível de aplicação de Web Scraping;
c) implementação do protótipo de Web Scraping: realizar a implementação do protótipo de Web Scraping aplicando as ferramentas e bibliotecas estudadas;
d) coleta de dados: utilizar o protótipo de Web Scraping para as legislações dos sites elencados. Inicialmente esse procedimento será realizado apenas uma vez, de forma a estruturar a base de dados e realizar o treinamento do agente, visto que possíveis alterações de legislações não impactam no desenvolvimento do presente trabalho. Posteriormente, a periodicidade de atualização da base de dados fica a critério dos especialistas extrajudiciais que acompanham as alterações e lançamentos das legislações.
e) estruturar base de dados: com base nos dados coletados, estruturar a melhor forma de armazenar essas informações;
f) pré-processamento dos dados: realizar o pré-processamento dos dados, limpando textos desnecessárias, aplicando *tokenização*, normalização, remoção de pontuações e de *stopwords*, dentre outras técnicas de Processamento de Linguagem Natural (PLN);
g) definir modelos de linguagem: realizar um estudo dentre os principais modelos de linguagem de código aberto, como Llama 2 e Falcon, que tiveram sua eficiência comprovada por Lin (2023), explorando os modelos mais recentes e definindo qual a quantidade adequada de parâmetros para o estudo aplicado;
h) analisar plataformas de criação de *chatbot*: realizar um estudo sobre as principais plataformas de criação de *chatbot*, como HuggingFace, LangChain e Dialogflow, a fim de definir a plataforma que melhor atende o propósito do trabalho;
i) implementar um protótipo de agente: aplicar os estudos realizados para a implementação de um agente, realizando o treinamento do mesmo com a base de dados criada e aplicando técnicas de engenharia de *prompt*;
j) estruturar perguntas e respostas esperadas: elencar, juntamente com um especialista do setor extrajudicial, perguntas que atendam as diversas áreas de atuação dos cartórios, envolvendo os principais assuntos de cada especialidade.
k) analisar respostas do agente: utilizar métodos de avaliação de modelo, como BLEU Score e ROUGE para avaliação das respostas geradas pelo agente;
l) coletar feedbacks: disponibilizar uma versão piloto para um especialista do setor extrajudicial, com o propósito de coletar sugestões de melhoria e identificar falhas nas respostas geradas;
m) implementar melhorias: realizar correções no agente para as devidas falhas encontradas pelos especialistas, realizando também implementação de melhorias nos *prompts* utilizados;
n) validação final: validar as melhorias e correções realizadas com o especialista do setor.

4 REVISÃO BIBLIOGRÁFICA

Esta seção descreve brevemente sobre os assuntos que fundamentarão o estudo a ser realizado neste trabalho, no caso Processamento de Linguagem Natural (PLN), Web Scraping e *chatbot*.

4.1 PROCESSAMENTO DE LINGUAGEM NATURAL

De acordo com Mooney e Bunescu (2005), o Processamento de Linguagem Natural (PLN) surgiu devido à necessidade de compreensão e comunicação de forma automática do ser humano com o computador. Trata-se de um mecanismo criado não só para extrair as informações de textos, mas também para facilitar a entrada de dados nos sistemas, auxiliando na estruturação dos mesmos. De acordo com Rodríguez e Bezerra (2020), a PLN consiste em uma subárea da inteligência artificial e refere-se, de forma ampla, ao estudo e desenvolvimento de sistemas computacionais que podem interpretar a fala e o texto conforme naturalmente os seres humanos falam e o digitam, além de desenvolver melhor a forma de interpretação da linguagem humana em diferentes dispositivos.

As formas mais comuns e conhecidas de interação com o PLN se dão através do Sistema de Posicionamento Global (Global Positioning System - GPS), chatbot para atendimento ao cliente e assistentes digitais. Entretanto, também pode ser aplicado para automatizar e facilitar processos de negócio (IBM, 2024). Uma forma de aplicação seria no processo de limpeza de dados e remoção de ruídos, possibilitando que os dados estejam adequados para o aprendizado de máquina. Dessa forma o aprendizado se torna mais eficiente, por conta de uma

maior qualidade e uniformidade no processo de entendimento dos dados, melhorando a qualidade e precisão (Lin, 2023).

O PNL possui quatro etapas para o processamento de informações. A primeira consiste na análise morfológica, que é responsável por formar um dicionário através da classificação de artigos, substantivos, verbos e adjetivos. O segundo se trata da análise sintática, responsável por verificar sujeito, complementos nominais, verbais, adjuntos e apostos. A terceira etapa consiste em realizar a análise semântica que associa os morfemas trazendo um sentido real para a frase em questão. Por fim, na quarta etapa é feita a análise pragmática, que faz a junção de todo o processo, trazendo resultado de análises realizadas (Bulegon e Moro, 2010).

4.2 WEB SCRAPING

Web Scraping é o processo de extrair e organizar dados automaticamente de sites, permitindo que seja reunidos grandes quantidades de informações da web. Essas informações permitem criar conjuntos de dados que podem ser analisados e aplicados de várias maneiras (Shubladze, 2023). Como fonte primordial de acesso à informação, a internet ganha protagonismo ao armazenar inúmeros dados no contexto Web que, por sua vez, geram informação e conhecimento. Entretanto, na maioria das vezes, os dados disponíveis na Web são encontrados em formatos desestruturados (Assis, 2021).

De acordo com Hernández *et al*. (2015, p. 114, tradução nossa), "Web Scraping ou extração de dados da Web é o processo de rastreamento e download de sites de informações e extração de dados não estruturados para um formato estruturado.". Esse processo objetiva buscar dados de sites diferentes e não estruturados e transformá-los em uma estrutura compreensível. Os dados e as informações transformam-se em planilhas, banco de dados ou arquivos de valores. Com isso, a técnica Web Scraping possibilita que os dados finais sejam dados abertos, que podem ser manipulados por máquinas (Mattosinho, 2010).

Quase todos os *scrapers* de dados na Web hoje são *bots* inteligentes, qualificados para extrair o código HTML de um site e, em seguida, editá-lo em informações estruturadas. Para isso, é feito um pedido de informações por meio de protocolo HTTP para o site desejado. O servidor web trata a solicitação e, se for legítima, fornece permissão ao *bot* para ler e selecionar o HTML da página. Após isso, são selecionadas as informações necessárias, sendo salvas em variáveis definidas (IBM, 2020). Algumas organizações podem usar desse processo para monitorar sites concorrentes ou plataformas de mídia social, onde obtém-se percepções sobre o comportamento do consumidor e tendências de mercado. Outras podem usá-lo para extrair dados de catálogos de produtos online, sites de avaliação e listas de empregos para melhorar suas ofertas ou serviços (Shubladze, 2023).

4.3 CHATBOT

De acordo com Prisco *et al*. (2019, p. 1, tradução nossa) "Um Chatbot é um sistema de computador cuja interface é fornecida por interação por meio de um diálogo em linguagem natural, simulando uma conversa humano-humano.". Conforme Følstad *et al*. (2021), o uso de *chatbots* está se tornando comum, ao ponto de que em 2019, estimou-se que mais de 50% dos consumidores dos EUA e da Alemanha usaram chatbots pelo menos uma vez, sendo ainda mais utilizado no Reino Unido e na França. No entanto, por mais que os *chatbots* tenham se popularizado recentemente, essa tecnologia já é aplicada a décadas, surgindo com iniciativas como a do Instituto de Tecnologia de Massachusetts, onde Weizenbaum implementou o chatbot ELIZA para emular um psicoterapeuta (Abushawar; Atwell, 2015).

Existem dois tipos principais: os *chatbots* declarativos, que seguem *scripts* e são úteis para responder a perguntas simples e repetitivas; e os preditivos, que são mais sofisticados, utilizando IA para aprender e se adaptar às preferências dos usuários, como as assistentes virtuais populares, Alexa e Siri (Shweta, 2022). De acordo com a IBM (2024, tradução nossa), "Os chatbots de perguntas frequentes não precisam mais ser pré-programados com respostas para perguntas definidas: é mais fácil e rápido usar IA generativa em combinação com a base de conhecimento de uma organização para gerar respostas automaticamente para uma gama mais ampla de perguntas.". Os chatbots baseados em Modelos de Linguagem de Grande Escala (LLMs) operam ajustando parâmetros essenciais para gerar respostas, como a temperatura, que regula a variabilidade das respostas, e o número máximo de tokens, que define o limite de extensão do texto produzido. Esses ajustes influenciam diretamente a coerência e a criatividade das interações, permitindo a criação de respostas adaptadas às necessidades do usuário, sejam elas mais objetivas ou detalhadas, conforme a configuração escolhida (Torres, 2024).

Os *chatbots* podem interagir por meio de texto ou voz, ajudando não apenas a responder perguntas, mas também a registrar solicitações de serviço e, quando necessário, conectar o usuário a um agente humano, tendo como principal forma de uso a melhora na experiência do cliente, fornecendo suporte 24/7 de forma econômica. Além disso, são utilizados para compartilhar informações e automatizar tarefas rotineiras, tornando o acesso a informações mais rápido, permitindo que os usuários evitem longas pesquisas em documentos ou esperas por suporte técnico (Shweta, 2022). Os principais impulsionadores desse desenvolvimento incluem avanços em campos de Inteligência Artificial (IA), como Processamento de Linguagem Natural (PLN) e compreensão de

linguagem natural (Natural Language Understanding - NLU), bem como a maior aceitação do consumidor por plataformas que conduzem à interação conversacional (Følstad *et al.*, 2021).

REFERÊNCIAS

ABUSHAWAR, Bayan; ATWELL, Eric. **Alice chatbot: Trials and outputs**. 2015. Computacion y Sistemas. 19. 10.13053/cys-19-4-2326. Disponível em: https://www.researchgate.net/publication/289684788_ALICE_chatbot_Trials_and_outputs. Acesso em: 21 set. 2024.

ASSIS, Wendel V. **Chat Bot Sumé: Web Scraping em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte**. 2021. Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) – Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento, Fundação Mineira de Educação e Cultura, Belo Horizonte. Disponível em: https://repositorio.fumec.br/handle/123456789/858. Acesso em: 08 set. 2024.

BAUMBACH, Rudinei; TRINDADE, Alexsandro S. **Inteligência artificial e direito: perspectiva para os cartórios extrajudiciais**. 2023. Revista de Direito Notarial, v. 5, n. 2. Disponível em: http://rdn.cnbsp.org.br/index.php/direitonotarial/article/view/89. Acesso em: 07 set. 2024.

BULEGON, Hugo; MORO, Claudia M. C. **Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar**. 2010. Journal of Health Informatics, v. 2, n. 2. Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/5. Acesso em: 22 set. 2024.

CÉSAR, Gustavo S. **A função social das serventias extrajudiciais e a desjudicialização**. [S.l.] [S.n.], [2019?]. Disponível em: http://stageieptbmg.com.br/variados/serventias_extrajudiciais_desjudicializacao.pdf. Acesso em: 08 set. 2024

FØLSTAD, Asbjørn., Araujo, Theo., Law, Effie.LC. et al. **Future directions for chatbot research: an interdisciplinary** research agenda. 2021. Computing 103, 2915–2942. Disponível em: https://doi.org/10.1007/s00607-021-01016-7. Acesso em: 29 set. 2024

HERNÁNDEZ, Alexis T.; VÁZQUEZ, Edy G.; RINCÓN, César A. B.; GARCÍA, Jorge M.; MALDONADO, Adrian C.; OROZCO, Rodolfo I. **Metodologías para análisis político utilizando Web Scraping**. 2015. Research in Computing Science, v. 95, n. 1, p. 113-121, 1 dez. 2015. DOI 10.13053/rcs-95-1-9. Disponível em: https://www.rcs.cic.ipn.mx/2015_95/Metodologias%20para%20analisis%20politico%20utilizando%20Web%20Scraping.pdf Acesso em: 22 set. 2024.

IBM. **O que é processamento de linguagem natural (PLN)?**. [S.l.], [2024?]. Disponível em: https://www.ibm.com/br-pt/topics/natural-language-processing. Acesso em: 22 set. 2024.

IBM. **Web Scraping basics:** what you need to know. [S.l], 2020. Disponível em: https://community.ibm.com/community/user/ai-datascience/blogs/bruce-wilson1/2020/02/06/web-scraping-basics-what-you-need-to-know. Acesso em: 21 set. 2024.

IBM. **What is a chatbot?**. [S.l] [S.d], Disponível em: https://www.ibm.com/topics/chatbots#:~:text=A%20chatbot%20is%20a%20computer. Acesso em: 14 set. 2024.

LIN, Felipe J. T. **A utilização de modelos de LLM para geração de bulas automatizadas. 2023**. Trabalho de Conclusão de Curso (Engenharia da Computação) - Universidade Federal de Pernambuco, Recife, 2023. Disponível em: https://repositorio.ufpe.br/handle/123456789/52694. Acesso em: 07 set. 2024.

LUCCHESI, Erika. R.; TEOTONIO, Luis. A. F.; CARLUCCI, Juliana. H. **Desjudicialização do poder judiciário, função social dos cartórios e cartorização dos serviços**. 2014. Revista Reflexão e Crítica do Direito, [S. l.], v. 1, n. 1, p. 87–98, 2014. Disponível em: https://revistas.unaerp.br/rcd/article/view/350. Acesso em: 08 set. 2024.

MATTOSINHO, Felipe. J. A. P. **Mining product opinions and reviews on the web**. 2010. [S.l] [S.n.], Technische Universitat Dresden. Disponível em: https://www.rn.inf.tu-dresden.de/uploads/Studentische_Arbeiten/Masterarbeit_Mattosinho_Felipe.pdf. Acesso em: 21 set. 2024.

MOONEY, Raymond J.; BUNESCU, Razvan. **Mining knowledge from text using information extraction**. 2005. ACM SIGKDD Explorations Newsletter, v. 7, n. 1, p. 3–10, 1 jun. 2005. Disponível em: https://dl.acm.org/doi/10.1145/1089815.1089817. Acesso em: 21 set. 2024.

MOREIRA, Taynara S. C. **A importância dos cartórios na sociedade**. Iporá: UNISA, 2022. Projeto Integrador (Tecnólogo em Serviços Jurídicos, Cartorários e Notariais) — Universidade Santo Amaro, Iporá. Disponível em: http://dspace.unisa.br/handle/123456789/1631. Acesso em: 15 set. 2024.

PRISCO, A. et al. **A Facebook chat bot as recommendation system for programming problems**. 2019 IEEE Frontiers in Education Conference (FIE) 1-5. 10.1109/FIE43999.2019.9028655. Disponível em: https://ieeexplore.ieee.org/document/9028655. Acesso em: 21 set. 2024.

RODRÍGUEZ, Marcia. M.; BEZERRA, Byron. L. D. **Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias)**. Revista de Engenharia e Pesquisa Aplicada, v. 5, n. 1, p. 67–77, 26 abr. 2020. Disponível em: http://revistas.poli.br/~anais/index.php/repa/article/view/1204. Acesso em: 08 set. 2024

SANTIAGO, Augusto Z. F. S. **Desenvolvimento de um sistema de chatbot para perguntas frequentes sobre a lei de direito autoral**. 2022. Trabalho de Conclusão de Curso (Bacharelado em Engenharia da Computação) – Centro de Ciências Exatas e Tecnológicas, Universidade Regional do Maranhão, São Luís. Disponível em: https://rosario.ufma.br/jspui/handle/123456789/7627. Acesso em: 07 set. 2024.

SHUBLADZE, Sandro. **Web Scraping: what it is and how companies can leverage it**. Forbes, 3 jan. 2023. Disponível em: https://www.forbes.com/councils/forbestechcouncil/2023/01/03/web-scraping-what-it-is-and-how-companies-can-leverage-it/. Acesso em: 22 set. 2024.

SHWETA. **What is a Chatbot? Everything You Need To Know**. [S.l.]: Forbes, [2022]. Disponível em: https://www.forbes.com/advisor/business/software/what-is-a-chatbot/. Acesso em: 14 set. 2024.

SILVA, Gabriel S.; RIBEIRO, Leandro. **Processamento da Linguagem Natural para análise de documentos jurídicos**. 2023. Trabalho de Conclusão de Curso (Curso Superior de Tecnologia em Informática para Negócios) – Faculdade de Tecnologia de São José do Rio Preto, São José do Rio Preto, 2023. Disponível em: http://ric.cps.sp.gov.br/handle/123456789/21385. Acesso em: 07 set. 2024.

TORRES, Gustavo. **Controlando seu LLM: Parâmetros de Inferência**. [S.l.], [2024]. Disponível em: https://brains.dev/2024/controlando-seu-llm-parametros-de-inferencia/?utm_source=chatgpt.com. Acesso em: 23 nov. 2024.