Instituto de Pesquisas Eldorado

Universidade Estadual de Campinas

iNuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação

Universidade de Brasília

Plataforma Computacional de Aprendizado de Máquina e Processamento de Linguagem Natural para textos jurídicos

Brasília, DF
MARÇO DE 2022

Proposta de Projeto de PD&I

Encomenda Tecnológica de Instrução Assistida por Inteligência Artificial

Gerente do Projeto
Fábio Grassiotto– Instituto de Pesquisas Eldorado

Coordenador da Equipe de Pesquisa
Luís Paulo Faina Garcia
Professor Adjunto A no Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB)

RESUMO

Este documento apresenta o projeto "Plataforma Computacional de Aprendizado de Máquina e Processamento de Linguagem Natural para textos jurídicos" e seus principais resultados almejados.

Neste projeto desenvolveremos um software que faz uso de técnicas de Inteligência Artificial (IA) para o processamento de linguagem e redação de textos jurídicos, contendo métricas de efetividade aplicadas à modelos de aprendizado de máquina e aprendizado profundo para a classificação de entidades visando: (1) a detecção de significados processuais, incluindo a identificação das alegações, exame de admissibilidade, cálculo da probabilidade de concessão de medidas cautelares; (2) a criação de um painel de jurimetria, incluindo a priorização de processos e comparação com causas anteriores; e (3) a redação de peças, incluindo a geração de comunicações aos interessados e de instruções contendo sumarização de teses e predição da análise técnica e das propostas de encaminhamento. Os produtos obtidos deverão ter seus direitos de propriedade intelectual assegurados por meio de registros junto ao INPI e serem utilizados no processamento de texto não estruturado e redação de textos aplicados ao setor jurídico, bem como pelas demais órgãos, de forma a maximizar o desempenho de análises de processos. Espera-se como resultados o desenvolvimento de uma plataforma computacional capaz de classificar e identificar similaridades semânticas entre novas peças processuais encaminhadas das denúncias e redigir, sumarizar e argumentar teses que irão subsidiar os usuários do TCU, de forma a aumentar a produtividade nas respostas das instruções contendo sumarização de teses e predição da análise técnica e das propostas de encaminhamento das denúncias. Esses benefícios corroboram para a viabilidade técnica e econômica de execução do

projeto visando obter ganhos de produtividade e redução de custos oriundos das etapas de classificação e análises dos documentos de denúncias encaminhados ao TCU.

Palavras-chave: Inteligência Artificial, Aprendizagem de Máquina, Processamento Natural de Linguagem, Jurimetria.

Duração do projeto: 36 meses

Segmento do projeto: Instrução Assistida por Inteligência Artificial

Tema de Pesquisa: Jurimetria e Instrução Assistida por Inteligência Artificial

Subtema de Pesquisa: Métricas de performance aplicadas à modelos de aprendizado de máquina e aprendizado profundo para a classificação de entidades binárias

Fase na Cadeia de Inovação: Desenvolvimento e PA – Pesquisa Aplicada

Entidades executoras:

- Universidade de Brasília (UnB)
- Universidade Estadual de Campinas (Unicamp)
- Instituto de Pesquisa Aplicada em Ciência, Tecnologia e Inovação (INuTech)
- Instituto de Pesquisas Eldorado

ÍNDICE

1. PRODUTO FINAL E BENEFÍCIOS QUANTITATIVOS DE P&D PARA O TCU

Este projeto se propõe a desenvolver uma plataforma computacional capaz de processar documentos jurídicos por meio de algoritmos de Aprendizado de Máquina para: 1) detectar o significado das peças processuais; 2) identificar similaridades semânticas entre as peças e; 3) realizar a redação de peças por meio da geração de teses e predição da análise técnica. Para a realização dessas tarefas é fundamental a construção de uma base de dados de peças processuais, além da identificação de segmentos e entidades nomeadas em cada processo. Com uma base rotulada representativa, modelos do estado da arte de Aprendizado de Máquina para classificação de texto serão induzidos. Espera-se que esses modelos, após a avaliação metodológica, sejam capazes de classificar novas peças processuais e indiquem possíveis similaridades semânticas entre as sentenças. Por fim, modelos de Aprendizado de Máquina serão induzidos para a redação de peças completando ou refutando teses. Ao final do projeto, todas as soluções de Aprendizado de Máquina poderão ser embarcadas ao processo de denúncia do TCU, de modo a auferir ganhos de desempenho, efetividade e confiabilidade. Com isto, espera-se ampliar a produção das análises e predição e consequentemente uma melhor resposta às demandas do TCU.

2. IDENTIFICAÇÃO

2.1 Entidades executoras

A equipe própria da proponente e das cooperadas participarão de todas as etapas do projeto, acompanhando todos os desenvolvimentos e entrega de produtos, bem como o cumprimento de todas as etapas previstas no cronograma. Entre as entidades executoras estão:

Razão social: Universidade de Brasília – UnB

Localização (Cidade/Estado): Brasília / Distrito Federal

Função específica no projeto: Equipe responsável pela realização da pesquisa e desenvolvimento de algoritmos, que empreguem técnicas de Inteligência Artificial, para detecção de significados processuais, incluindo a identificação das alegações, exame de admissibilidade, cálculo da probabilidade de concessão de medidas cautelares.

Experiência no tema do projeto proposto: O Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB) tem atuado no desenvolvimento tecnológico e científico em diversos tópicos, incluindo o de Mineração de Textos e Aprendizado de Máquina. Esse desenvolvimento tem gerado produtos, patentes e aplicações com diversos parceiros tecnológicos. Alguns desses, ganharam notoriedade como o projeto Victor[i] com o STF e o projeto

KnEDLe[ii] com TCDF. O portfólio da UnB em projetos envolvendo a área jurídica e IA é extenso e envolve vários departamentos e cursos (destaca-se os projetos do grupo DrIa .[iii] Além desses, diversas colaborações nacionais e internacionais têm gerado parcerias e soluções para o bem-estar e o desenvolvimento social. e do laboratório AiLab[iv]).

Razão social: Universidade Estadual de Campinas - Unicamp

Localização (Cidade/Estado): Campinas / SP

Função específica no projeto: Equipe responsável pelo levantamento e condução de pesquisa no estado da arte em algoritmos e modelos de aprendizado para processamento de linguagem natural para instrumentos jurídicos.

Experiência no tema do projeto proposto: O Instituto de Computação (IC) da Universidade de Campinas (UNICAMP) por meio de sua competência consolidada em pesquisa na Ciência da Computação tem estudado e desenvolvido metodologias e técnicas para o processamento e análise de grandes volumes de dados. As soluções têm sido aplicadas por meio de convênios com parceiros em diversos domínios, incluindo suporte às aplicações em cidades inteligentes, gerenciamento de campos de pré-sal, etc. O IC tem desenvolvido competência em soluções de aprendizado de máquina para processamento de linguagem natural aplicados a diálogos na Língua Portuguesa e textos não estruturados em plataformas de comércio eletrônico, por meio do processamento automatizado de mensagens em sistemas de perguntas e respostas. O IC/UNICAMP possui um amplo portfólio de projetos envolvendo técnicas de inteligência artificial e processamento de linguagem natural. Coordena atualmente o Hub de Inteligência Artificial e Arquiteturas Cognitivas - H.IAAC em parceria com o Instituto Eldorado, sendo uma iniciativa do Ministério da Ciência, Tecnologia e Inovações (MCTI). Projeto Viva Bem, um hub de Inteligência Artificial com foco em saúde e bem-estar, sendo uma parceria entre IC/UNICAMP, Samsung Brasil e o Instituto SiDi. Convênio com a empresa CI&T, multinacional brasileira de desenvolvimento de software, na criação do Cognitive LAB, um projeto que estuda a compreensão das máquinas sobre diálogos humanos.

Razão social: Instituto NuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação

Localização (Cidade/Estado): Brasília/DF

Função específica no projeto: Apoiar as equipes do Instituto de Pesquisas Eldorado, da UnB e da UNICAMP com sua equipe de especialistas, com experiência no desenvolvimento de projetos similares envolvendo o aprendizado para processamento de linguagem natural para a análise de peças jurídicas de textos não estruturados em língua portuguesa e para aplicações de jurimetria,

com competências técnicas relacionadas ao tema, assim como com especialistas na área de domínio, nas seguintes atividades: concepção, pesquisa bibliográfica, científica e mercadológica voltada à análise de viabilidade técnica e econômica do projeto, implantação e uso de metodologia apropriada ao desenvolvimento de projetos de ciência de dados, processos e atividades de engenharia de dados, orientação à rotulagem de documentos, modelagem, criação e carga de bases de dados NoSQL para grandes volumes de dados, arquitetura de infraestrutura tecnológica em nuvem para o desenvolvimento e para a execução de aplicações de aprendizado de máquina e aprendizado profundo, e nas diversas fases e etapas do desenvolvimento da solução, com destaque para o desenvolvimento e uso de algoritmos e modelos explicáveis, análise de redes semânticas e de redes complexas, definição de métricas de performance aplicadas à modelos de classificação de entidades binárias e não binárias. **Experiência no tema do projeto proposto:** O Instituto NuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação, atuou durante os últimos quatro anos em alguns projetos de Processamento de Linguagem Natural (NLP), destacando-se: (1) Projeto Atrium com o objetivo de facilitar a mediação e conciliação judicial, apoiada por algoritmos e modelos matemáticos de processamento de linguagem natural, para a área trabalhista – http://app.atrium.eti.br/ -; (2) Projeto de Assistente de Voto para o TST com o objetivo de facilitar e otimizar as atividades voltadas ao julgamento de processos trabalhistas no TST, apoiada por algoritmos e modelos matemáticos de processamento de linguagem natural. Esses projetos tiveram como foco a análise e processamento de documentos jurídicos por meio de algoritmos inteligentes para: estimativa de risco jurídico, estimativa de prazos de tramitação, estimativa de valores de causas, apoio para a escolha de estratégia jurídica, até o apontamento de provável decisão. Como parte dos resultados, os projetos tem gerado produtos comerciais, criação de uma startup – Atrium Legaltech, artigos técnicos e capítulo de livro sobre o desenvolvimento, utilização e impactos de aplicações de inteligência artificial na área do Direito (FERNANDES, 2020).

Razão social: Instituto de Pesquisas Eldorado.

Localização (Cidade/Estado): SCS, Quadra 09, Bloco C, Torre C, Sala 503, 5º andar - Asa Sul, Brasília - DF, 70308-200

Função específica no projeto: Equipe especialista no desenvolvimento da plataforma computacional das soluções back-end, front-end e da metodologia de Ciência de Dados ágil.

Experiência no tema do projeto proposto: Entidade sem fins econômicos e com mais de 23 anos atuando ao lado de empresas de todo o mundo, o Eldorado atua fortemente no setor de PD&I

e possui um portfólio de projetos multisetoriais financiados por fontes de recursos privados ou públicos. Tem ampla experiência no desenvolvimento de projetos com recursos oriundos de benefícios fiscais e fundos de fomento governamentais como ANEEL, EMBRAPII, FINEP, BNDES, Lei da Informática e é mantido exclusivamente pelos projetos que executa. Oferece aos seus parceiros um leque completo de possibilidades de atuação conjunta em diversas áreas, tais como: P&D em Software, P&D em Hardware, P&D em Capacitação, Engenharia de Testes, Otimização de Processos, Gestão de Projetos, Assessoria em Fundos de Fomento e Certificação de Produtos Eletrônicos, Sistema de Gerenciamento de Energia Residencial, Inspeção semiautônoma de torre de linha de transmissão por drone e em Eficiência em ML. O Instituto Eldorado é referência em pesquisa e desenvolvimento de tecnologia, habituado a operar com processos e parceiros globais, sendo protagonista no cenário da inovação aberta e com experiência em trabalhar com rede de parceiros.

O Instituto Eldorado, em parceria com a Unicamp, está desenvolvendo para o MCTI por meio de projetos PPI em P&D, um projeto na área de arquiteturas cognitivas para aplicação em plataformas móveis e a disseminação do conhecimento adquirido apor meio da capacitação de profissionais em Inteligência Artificial e impulsionamento da indústria nacional de aplicações móveis. O conhecimento tecnológico será validado em aplicações piloto.

2.2 Tabela de Entidades

RAZÃO SOCIAL	FUNÇÃO NO PROJETO	CNPJ	UF
Universidade Estadual de Campinas	Executor	46.068.425/0001-33	SP
Fundação de Desenvolvimento da Unicamp - FUNCAMP	Interveniente Administrativa	49.607.336/0001-06	SP
Universidade de Brasília	Coordenador da Pesquisa / Executor	00.038.174/0001-43	DF
Fundação de Empreendimentos Científicos e Tecnológicos	Interveniente Administrativa	37.116.704/0001-34	DF

INuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação	Executor	0.164.968/0001-14	DF
Instituto de Pesquisas Eldorado	Coordenador do Projeto/Executor	02.437.460/0001-07	DF

2.3 Equipe executora

Nome	**Titulação**	**Formação Profissional**	**Função no Projeto**	**Empresa**
Luís Paulo Faina Garcia	Doutor	Professor	Coordenador	UnB
Thiago de Paula Faleiros	Doutor	Professor	Pesquisador	UnB
Vinicius Ruela Pereira Borges	Doutor	Professor	Pesquisador	UnB
10 alunos de graduação	Graduando em direito	Estudante	Bolsista	UnB
06 alunos de graduação	Graduando em ciência da computação	Estudante	Bolsista	UnB
5 alunos de pós-graduação (Mestrandos e doutorandos)	Pós-graduando em computação e áreas afins	Estudante	Bolsista	UnB
Julio Cesar dos Reis	Doutor	Professor	Pesquisador	Unicamp
2 alunos de Mestrados	Pós-graduando em computação	Estudante	Bolsista	Unicamp
2 alunos de Doutorado	Pós-graduando em computação	Estudante	Bolsista	
1 aluno de Pós Doutorado	Pós-graduando em computação	Estudante	Bolsista	
4 alunos de graduação	Graduando em ciência da computação	Estudante	Bolsista	Unicamp

Gilberto Lourenço Fernandes	Mestre	Engenheiro da dados/Analista de Sistema	Especialista	iNuTech
Ana Victória Gruginski de Carvalho Ladeira	Mestre	Cientista de Dados/Analista de Sistema	Especialista	iNuTech
Gabriel Estevam Botelho Cardoso	Graduação	Bacharel em Direito/Advogado	Especialista	iNuTech
Guilherme Albuquerque Barbosa Silva	Doutorando	Cientista de Dados/Engenheiro de Dados	Especialista	iNuTech
Sérgio Medeiros de Souza	Mestre	Engenheiro da dados/Analista de Sistema	Especialista	iNuTech
Fabio Grassiotto	Mestre	Engenheiro da dados/Analista de Sistema	Líder do Projeto /Scrum Master	Eldorado
Ricardo de Souza Leite	Graduação	Engenheiro da dados/Analista de Sistema	Dono do Produto	Eldorado
Tatiana Saldanha Tavares	Mestre	Engenheiro da dados/Analista de Sistema	Especialista	Eldorado
Priscila Batista Fayad André	Graduação	Cientista de Dados/Engenheiro de Dados	Especialista	Eldorado
Artur Henrique Brandão	Graduação	Engenheiro da dados/Analista de Sistema	Especialista	Eldorado
Mário Seiji Saiki Takafuji	Graduação	Analista de Software FullStack	Especialista	Eldorado
Valdeci Batista Caixeta	Especialista	Analista de Rede	Especialista	Eldorado
Erica Letícia Silva Serra	Especialista	Analista de Segurança	Especialista	Eldorado
André Luiz Muniz dos Reis	Especialista	Testador	Especialista	Eldorado
Juliana Fagg Menicucci	Graduação	Designer	Especialista	Eldorado

UnB - Luís Paulo Faina Garcia. Possui graduação em Engenharia de Computação (2010) e doutorado em Ciências da Computação (2016) pela Universidade de São Paulo. Em 2017 teve a tese classificada entre as melhores pela Sociedade Brasileira de Computação (SBC) e recebeu o prêmio CAPES de melhor tese em Ciência da Computação do país. Atualmente é Professor Adjunto A no Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB), em Brasília - Distrito Federal. Tem experiência na área de Ciência da Computação principalmente nos temas relacionados a Mineração de Dados e Aprendizado de Máquina, atuando nas seguintes linhas de pesquisa: detecção de ruídos, meta-aprendizado e fluxo contínuo de dados. http://lattes.cnpq.br/1607852138156562

UnB - Thiago de Paula Faleiros –Atualmente é Professor Adjunto A no Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB), em Brasília - DF. Possui graduação em Ciência da Computação pela Universidade Federal de Goiás (2007), mestrado em Ciência da Computação pela Universidade Estadual de Campinas (2010) e doutorado em Ciência da Computação pela Universidade de São Paulo (2016), com período sanduíche na University of Maryland, EUA. Tem experiência na área de Ciência da Computação, com ênfase em Inteligência Artificial, atuando principalmente nos seguintes temas: algoritmos, modelos probabilísticos de tópicos, extração de informação, teoria dos grafos, aprendizado não supervisionado e otimização combinatória. http://lattes.cnpq.br/1193412523364471

UnB - Vinicius Ruela Pereira Borges. Lattes: Atualmente é Professor Adjunto A no Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB). Obteve o grau de Bacharel (2009) e de Mestrado (2011) em Ciência da Computação na Faculdade de Computação da Universidade Federal de Uberlândia. Obteve o título de doutor (2016) em Ciências da Computação e Matemática Computacional no Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP), em São Carlos-SP. Realizou doutorado sanduíche (2014-2015) na University of California, Davis (UC Davis), em Davis, Califórnia, Estados Unidos. Foi Professor Substituto (2016-2017) no Departamento de Ciência da Computação (DCC) da Universidade Federal de Lavras (UFLA), em Lavras-MG. Suas áreas de interesse compreendem visualização da informação, aprendizado de máquina, mineração visual de dados, visão computacional (extração de características e reconhecimento de padrões em imagens) e processamento de imagens. Ademais, coordena o projeto de extensão "Maratona de Programação" no âmbito da UnB, além de ser técnico das equipes do CIC/UnB na Maratona SBC de Programação e do International Collegiate Programming Contest (ICPC). É membro da Sociedade Brasileira de Computação e da IEEE. http://lattes.cnpq.br/1841593572448050

Unicamp – Julio Cesar dos Reis - Lattes: Professor Associado no Instituto de Computação (IC) da Universidade Estadual de Campinas (UNICAMP). Possui doutorado em Ciência da Computação (2014) pela Faculdade de Ciências da Universidade de Paris-Sud XI (França); mestrado em Ciência da Computação (2011) pelo IC/UNICAMP e graduação em Tecnologia em Informática (2008) pela Faculdade de Tecnologia da UNICAMP. Tem interesse de pesquisas em Web Semântica, Engenharia de Ontologias Computacionais, Processamento de Língua Natural e Interação Humano-Computador. Investiga principalmente os seguintes temas: representação do conhecimento; grafos de conhecimento; semântica computacional; design, alinhamento e evolução de ontologias; linked data; adaptação e refinamento de mapeamentos semânticos; recuperação semântica de informação; detecção e representação de intenções declaradas por usuários; design da interação; design participativo e universal; sistemas colaborativos interativos. http://lattes.cnpq.br/2971609673131726

iNuTech - Gilberto Lourenço Fernandes. Lattes: Mestre em Ciência da Informação pela Universidade de Brasília/UnB (2014), graduação em Engenharia Eletrônica pela Universidade Federal do Rio de Janeiro (1980) e especialização em Inteligência Artificial e Redes Neurais pela COPPE/UFRJ (1992). Desde 2018 ocupa o cargo de Diretor Executivo do Instituto NuTech, atuando em pesquisas e no desenvolvimento de projetos de inovação na área de Inteligência Artificial. Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software, atuando principalmente nos seguintes temas: processos de desenvolvimento de aplicações de software, epistemologia da informação, filosofia da informação, processos

cognitivos e problemas de entendimento, Teoria do Conhecimento, fenomenologia. http://lattes.cnpq.br/2971609673131726.

iNuTech - Ana Victória Gruginski de Carvalho Ladeira. Cientista de dados, mestre em Inteligência Artificial e graduada em Data Science e Inteligência Artificial pela Universidade de Maastricht/Holanda. Experiência profissional como Product Owner de projetos de IA e Data Science, Processamento de Linguagem Natural, Deep Learning, Análise de Redes Complexas e Sistemas de Recomendação. Participou da prova de conceito do TST/iNuTech e de dois anos no Projeto Atrium/iNuTech. Trabalhou para o Centro de Gestão e Estudos Estratégicos durante 3 anos, e na Accenture durante dois anos. http://lattes.cnpq.br/9202497499998472

iNuTech - Gabriel Estevam Botelho Cardoso. Bacharel em Direito pela Universidade de Brasília (2015 - 2020). Advogado no escritório Torreão Braz Advogados, onde trabalha desde 2017 e atua majoritariamente no direito empresarial, com foco em discussões contratuais estratégicas no mercado de infraestrutura, e em execuções contra a fazenda pública, com foco no mapeamento de precatórios e na construção de acordos com a Advocacia Geral da União. Foi pesquisador sênior do LIFT Learning, programa do Banco Central do Brasil voltado ao desenvolvimento de soluções inovadoras. Pesquisador vinculado ao iNuTech (Instituto NuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação), instituição de ciência, tecnologia e inovação (ICT) que tem suas origens vinculadas ao Centro de Pesquisas em Arquitetura da Informação (CPAI) da Universidade de Brasília (UnB). Integra a equipe da startup Atrium, especializada na aplicação de ciência de dados ao Direito, em parceria com a ISG Participações e com o Ferreira & Chagas Advogados. Integrou a equipe fundadora da Legal Labs, hoje Neoway Legal, startup especializada em Big Data e Analytics aplicadas ao mercado jurídico. Foi pesquisador vinculado à Universidade Federal de Minas Gerais (UFMG), via FUNDEP/FUNDEPAR, tendo exercido o papel de agente de aceleração no programa Lemonade Brasília, voltado ao desenvolvimento de startups em fase inicial. http://lattes.cnpq.br/6904652069642388

iNuTech - Guilherme Albuquerque Barbosa Silva. Doutorando em Ciência da Computação na Universidade de Brasília, Mestrado em Sistemas Mecatrônicos pela Universidade de Brasília (2019), Graduação em Ciência da Computação pelo Centro Universitário de Brasília (2016). Experiência na área de Ciência da Computação, com ênfase em Aprendizagem de Máquina, Visão Computacional e Reconhecimento de Imagens. Atuou como cientista de dados e engenheiro de dados no Projeto Atrium/iNuTech durante dois anos. Lattes: http://lattes.cnpq.br/1626951659137199

iNuTech - Sérgio Medeiros de Souza. Mestre em Gestão do Conhecimento e da Tecnologia da Informação pela Universidade Católica de Brasília, Master's Certificate in Project Management in The George Washington University, Pós Graduado em Ciência da Informação pelo IESB-Centro Universitário, Pós Graduado em Data Base System Design pela Japan International Cooperation Agency e Graduado em Licenciatura Plena em Matemática pelo Centro Universitário de Brasília - CEUB. Atualmente é Professor do Curso de Ciência de Dados e Inteligência Artificial no IESB Centro Universitário, ministrando a disciplina Ciência da Informação e Ética. Atualmente no iNuTech - Instituto NuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação na função Diretor Administrativo e Financeiro, pesquisador e membro do Comitê Científico. Atuou como Gerente de Projetos Sr. na Capgemini, no Projeto de implantação do SAP na CEF (Caixa

Econômica Federal). Como função principal era gestor das equipes responsáveis pela Infraestrutura (BASIS), Arquitetura de Integração SAP-MainFrame e SAP Security. Gerente Sr. na Oi-BrasilTelecom nas áreas de Planejamento de Infraestrutura de TI, Arquitetura de Aplicações, Suporte Técnico e Operação, responsável pelo Planejamento e Execução do Orçamento de Infraestrutura de TI (CAPEX); pelas Soluções Técnicas (Infraestrutura), aquisição e especificação para implantação; responsável pelo Plano de Capacidade da Infraestrutura de TI; responsável pelo Plano de Contingência da BrT. Atuou com Professor Titular no Curso de Ciência da Computação no CEUB, ministrando aulas no Curso de Projeto de Banco de Dados ORACLE. Atuou como Chefe do Departamento de Desenvolvimento de Sistemas no SERPRO (Serviço Federal de Processamento de Dados). Participou do Projeto Atrium/iNuTech por um ano.
http://lattes.cnpq.br/9623395938256796

Eldorado - Fabio Grassiotto (Líder do Projeto /Scrum Master). Líder de desenvolvimento de soluções de inteligência artificial no Eldorado. Gerente com mais de 20 anos de experiência na liderança de equipes e no desenvolvimento de soluções para a indústria mobile. Candidato ao mestrado em Inteligência Artificial na Faculdade de Engenharia Elétrica da Unicamp, com previsão de término no primeiro semestre de 2022. Áreas de interesse: Inteligência Artificial, Autismo, Arquiteturas Cognitivas. http://lattes.cnpq.br/3571635926402375

Eldorado - Mariza Aparecida Rabelo Lira (Project Management Office-PMO). Possui graduação em Tecnologia em Processamento de Dados pelo Centro Universitário Planalto do Distrito Federal (2002). Foi Gerente do Projeto do Grupo Caixa Seguros. Tem experiência na área de Ciência da Computação, com ênfase em Gestão de Projeto. http://lattes.cnpq.br/4095139211674840

Eldorado – Ricardo de Sousa Leite (Project Owner). Graduado em Ciência da Computação pela Universidade Católica de Brasília (2007), MBA pelo Centro Universitário Unieuro (2010) e pelo Centro Universitário Senac (2019). Atualmente atuando como analista de software do Instituto de Pesquisas Eldorado - Brasília. Experiência na área de Ciência da Computação, com ênfase em Engenharia de Software. Quinze anos de experiência trabalhando na área de Tecnologia da Informação em projetos como analista de testes, passando pelas áreas de qualidade e, com base na minha atenção e curiosidade pelas necessidades dos clientes, migrei para a área de negócios, atuando desde a concepção do produto até o gerenciamento de esforço para criar produtos relevantes. Trabalho com o Agile há sete anos, aplicando e disseminando melhorias e conhecimentos nas empresas desempenhando papéis ágeis para apoio à equipe de desenvolvimento e testes (criação e manutenção de relatório e métricas, preparação, planejamento, revisão e reuniões diárias). Criação e manutenção de itens do backlog, como stories, features, improvements, além de elaborar roadmaps de entregas do projeto e alinhá-los às expectativas do cliente. Lidando com diversas indústrias (Telecomunicações, Instituições Financeiras, Conglomerados de Vendas, entre outros) com várias tecnologias como: Mobile (Android), Wearables, IoT (Sensores and Beacons), IA, ML e Cloud (PCF). http://lattes.cnpq.br/6349399372615225

Eldorado - Tatiana Saldanha Tavares (Analista de Software em Inteligência Artificial) – Mestre - Atualmente Analista de Sistemas no Instituto de Pesquisas Eldorado, docente da Pós-graduação em Inteligência Artificial e da Pós em Tecnologias Disruptivas do Instituto de Educação Superior de Brasília (IESB). doutoranda da Pós-graduação em Engenharia Elétrica (PPGEE) da Universidade de Brasília (UnB) com pesquisa nas seguintes áreas: Processamento de Sinais,

Inteligência Artificial e Engenharia Biomédica com previsão de término em 2023. Mestre e graduada em Engenharia Elétrica pela Universidade Federal de Uberlândia (UFU). Tenho sólidos conhecimentos em eletrônica, automação, processamento de sinais e imagens, inteligência artificial e algoritmos, tenho experiência como cientista de dados unindo ciência comportamental e inteligência artificial e como docente no ensino superior presencial/EaD e técnico. Meus interesses em pesquisa incluem métodos de reconhecimento de padrões para: Visão Computacional e Engenharia Biomédica tais como, Machine Learning e Deep Learning. http://lattes.cnpq.br/6918867712803013

Eldorado - Priscila Batista Fayad André (Cientista de Dados) (especialista) - Graduação em Engenharia da Computação pela faculdade IESB em 2006 e pós-graduado em Arquitetura de software com métodos ágeis pela Cruzeiro do Sul em 2020. Com 17 anos trabalhando na área de desenvolvimento de software, já atuou em todas as fases de projeto, incluindo análise de sistemas, entendimento do negócio e mapeamento de processos, modelagem de dados, arquitetura de software, desenvolvimento (front-end e back end), manutenção de banco de dados, testes unitários e de integração e configuração de pipeline de integração e entrega contínua (CI/CD). Atua como desenvolvedora de software desde agosto/2020 no Instituto de Pesquisas Eldorado, sendo responsável pela criação da estrutura de banco de dados em SQL Server e Oracle e do desenvolvimento de sistemas em .Net Core e em Python. Antes desse período, passou quase 10 anos atuando com sistemas de software de comunicação móvel de dados em empresa de gestão de frotas e logística, onde adquiriu conhecimento sobre logística e melhoria de performance dos sistemas, entre ambientes computacionais muito diversificados. Como desenvolvedor de software do projeto participará em todas as etapas técnicas. http://lattes.cnpq.br/4521736144930083

Eldorado - Artur Henrique Brandão de Souza. (Analista de Software em Inteligência Artificial): Possui graduação em Ciência da Computação pela Universidade de Brasília (2021). Desenvolveu um artigo, tendo como base o TCC apresentado como trabalho final da graduação, que ganhou a premiação como melhor artigo no WTG 2021 (SRBC2021). Atualmente atua como desenvolvedor na área de Business Intelligence e em processos ETL no Instituto de Pesquisas Eldorado. http://lattes.cnpq.br/8761390307031259

Eldorado - Mário Seiji Saiki Takafuji (Analista de Software - FullStack) - Superior - Graduação em Ciências da Computação, pelo Centro Universitário de Brasília (UniCEUB) em 2011. Atua no mercado desde 2007, período em que desenvolveu trabalhos nas seguintes áreas: Desenvolvimento e manutenção de sistemas em C# + ASP.net, JAVA (Springboot), Angular, React, manutenção da base de dados em SQL SERVER, MY SQL e ORACLE além de trabalhar com os seguintes padrões: TDD, DDD, MVC, CQRS, SOA. A última empresa em que ingressou foi o Instituto de Pesquisas Eldorado, com vínculo empregatício a partir de fevereiro de 2020, sob o cargo de Analista de Software. http://lattes.cnpq.br/5594946216950861

Eldorado - Valdeci Batista Caixeta (Analista de Rede) – Especialista– Pós -Graduação em Projeto de Redes e Computação em Nuvem, pelo Centro Universitário UDF em 2020. Tecnólogo em Análise e Desenvolvimento de Sistemas pelo Centro Universitário UDF em 2016. Atualmente atua como analista de rede no Instituto de Pesquisas Eldorado. https://www.linkedin.com/in/valdeci-caixeta-4524244b/

Eldorado - Erica Letícia Silva Serra (Analista de Segurança). Especialista – Pós Graduação em Cyber/Computer Forensics and Counterterrorism pelo Instituto de Gestão e Tecnologia da

Informação, em 2020. Graduada em Sistemas de Informação pela FAMETRO em 2015. Atualmente atua como analista de segurança de informação no Instituto de Pesquisas Eldorado. https://www.linkedin.com/in/erica-serra-b1539890/

Eldorado - André Luiz Muniz dos Reis (Testador) - Especialista - Graduado em análise e Desenvolvimento de Sistemas, pela Universidade Católica de Brasília em 2013. Em 2020, concluiu a pós-graduação em Gestão da Qualidade de Software pela UNIBF. Possui Certificações internacionais com foco em qualidade de software pelo BSTQB – CTFL (Foundation Level), CTFL-AT (Agile Tester), CTFL-UT (Usability Testing), CTFL-MAT(Mobile Application Testing). Também possui certificação em desenvolvimento ágil pela scrum.org – PSM I – Professional Scrum Master I. Atua desde 2012 com teste e qualidade de software, testes funcionais, não funcionais, de regressão, de carga, de integração, de homologação, testes manuais e automatizados. Desde 06/2020 atua como Analista de Software (QA) pelo Instituto de Pesquisas Eldorado. http://lattes.cnpq.br/0631004534223076

Eldorado - Juliana Fagg Menicucci (Designer) – Superior - Graduada em Psicologia pelo UniCEUB, em 2010, tem como segunda formação Design Gráfico pelo IESB (2019), atualmente cursa pós-graduação pela PUC-RS voltada para a área de design de experiência e interfaces. Atua como UX designer desde 2020, tendo ingressado no Instituto Eldorado em julho de 2021. Durante esse tempo, vivenciou experiências diversas dentro da área de design de produto, indo desde pesquisas com usuários, benchmarking, entrevistas de entendimento e aprofundamento negocial, até criação de wireframes, fluxogramas, protótipos de alta-fidelidade, testes de usabilidade e acompanhamento do trabalho de diversas equipes de desenvolvimento. http://lattes.cnpq.br/8786712018918626

2.4 Modelo de Governança Executores

O modelo de governança dos executores será composto pelo Comitê Estratégico, Eldorado, UnB, UNICAMP e iNuTech. A Figura a seguir representa o modelo proposto de governança dos executores.

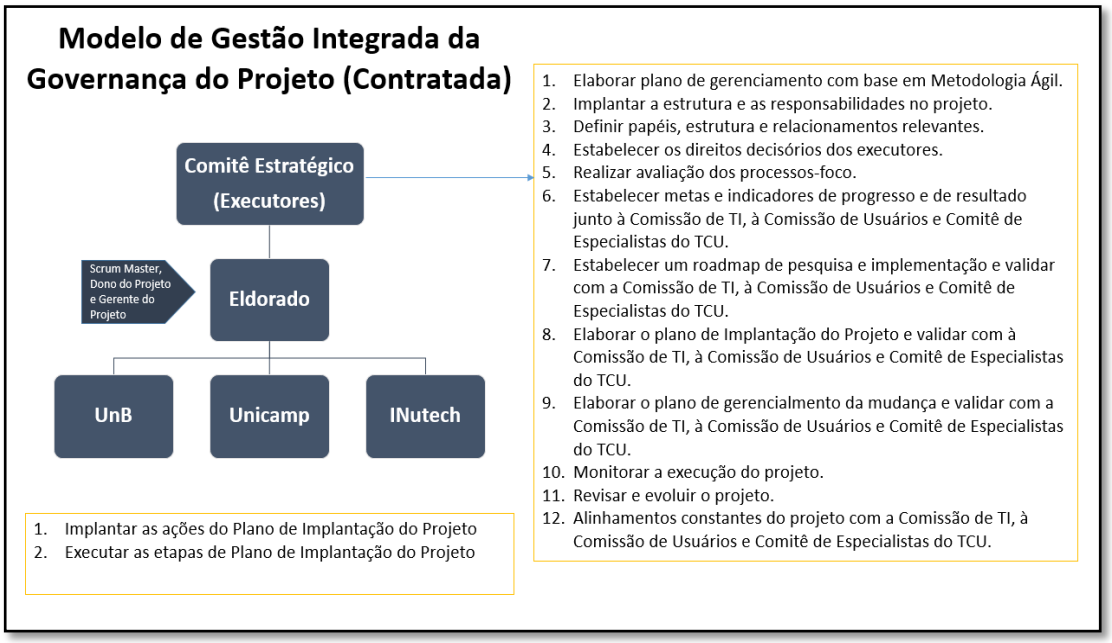

Figura 1: Modelo de gestão a ser conduzido no projeto

A organização dessas instituições deve seguir a seguinte estrutura: (1) o Comitê Estratégico ficará responsável por definir os pré-requisitos por meio das definições de papéis, estabelecer metas além de elaborar planos de implementação, gerenciamento e validação as ferramentas disponibilizadas. O Eldorado ficará responsável por intermediar essas atividades por meio do Scrum Master, Dono do Projeto e Gerente de Projeto. As instituições UnB, UNICAMP e iNuTech devem completar as demais equipes dos executores do projeto participantes do consórcio.

3. MOTIVAÇÃO

3.1 As Forças que Guiam a Transformação Digital

A inteligência artificial com suas subáreas, destacando-se aprendizado de máquina, aprendizado profundo e processamento de linguagem natural, é a principal tendência atual entre as tecnologias em fase de adoção pelo setor jurídico e pelo sistema judiciário. Em geral, as soluções envolvendo inteligência artificial (IA) vêm sendo combinadas com *big data/analytics* e com o uso da computação em nuvem. A junção combinada permite trabalhar com grande volume de dados em um ambiente seguro, realizar análises cada vez mais sofisticadas, envolvendo dados de várias fontes, texto, fala, sentimentos ou uma combinação de todos esses elementos. As soluções podem aprender com bases em dados históricos, melhorando ao longo do tempo o seu desempenho analítico.

As aplicações de interesse, que tem como base as combinações de tecnologias, métodos e

técnicas acima mencionadas, são muitas: a possibilidade de elaboração de contratos, o apoio de chatbots e assistentes virtuais nas tarefas iniciais de contato com clientes e no apoio à capacitação de pessoal em temas de interesse; uso de conselheiros especialistas em áreas chaves do setor, a possibilidade de contar com fornecedores com expertise para extração e análise de dados públicos e, os avanços possíveis, na jurimetria convencional, já incorporada ao dia-a-dia dos profissionais do Direito em nichos específicos de atuação.

As tecnologias emergentes podem turbinar as ferramentas de gestão já em uso pelo setor jurídico e sistema judiciário. Destacam-se, em especial, as novas gerações já disponíveis para gestão do ciclo de vida dos contratos, gestão de riscos legais e compliance e gestão jurídica. As legaltechs têm crescido vertiginosamente nos últimos anos. Atualmente, no Brasil, já cerca de 300 empresas que oferecem ao setor jurídico um leque diversificado de soluções para segmentos variados: jurimetria avançada, busca e gestão de documentos legais, gestão inteligente do ciclo de vida dos contratos; gestão jurídica inteligente; risco e compliance; extração e monitoração de dados públicos; inteligência artificial para o setor público; conteúdo jurídico, educação e consultoria; plataformas para relacionamento e plataformas online para resolução de conflitos. Suas ofertas estão sintonizadas com as novas tendências digitais e com a combinação das forças que estão impulsionando a nova revolução. No sistema judiciário brasileiro, onde predomina o desenvolvimento *in-house*, o PJe, sistema desenvolvido pela equipe técnica de uma das cortes, foi eleito pelo Conselho Nacional de Justiça (CNJ) como padrão para a política de unificação dos sistemas Tecnologias digitais para o setor jurídico. O sistema judiciário dos tribunais brasileiros deve ser adotado por todos os órgãos do judiciário. A sua aceitação ainda é relativamente baixa. Predominam sistemas diversos, com versões variadas que não conversam entre si, o que dificulta o trabalho dos advogados e outros usuários.

Várias iniciativas recentes do CNJ apontam para um esforço relevante para reverter a situação caótica da informatização no judiciário, buscando agilidade e eficiência. A aposta do órgão foi nas tecnologias de inteligência artificial. Por meio de plataformas que ofertam microsserviços acessíveis via API, será possível compartilhar e reutilizar modelos de IA, o que irá ampliar as possibilidades de desenvolvimento de ferramentas em código aberto para uso das partes interessadas. O CNJ convida os tribunais a participarem da iniciativa com seus respectivos times técnicos, formando e fortalecendo um ecossistema colaborativo de desenvolvimento. Os primeiros resultados desta e de outras iniciativas do Programa Justiça 4.0 e outros começam a surgir e parecem animadores.

No conjunto apresentado de tecnologias portadoras do futuro, é possível observar algumas

tendências norteando as pesquisas/projetos tecnológicos em andamento. Uma delas evidencia o esforço de desenvolver soluções que permitam superar desafios apresentados pelas tecnologias digitais já em fase de adoção. Vencer, por exemplo, o obstáculo de necessitar de grandes volumes de dados para realizar análises avançadas ou, ainda, obter soluções que permitam escapar da cilada da caixa preta do aprendizado profundo de máquina, que impede que se conheçam os percursos lógicos percorridos pela máquina para chegar à tal ou qual resultado. Outra tendência clara das tecnologias é rumo à busca por soluções para intensificar o processo de automação.

Os temas de pesquisa apresentam uma clara determinação de seguir o processo de tornar as máquinas cada vez mais inteligentes. Também se observa, entre as tecnologias portadoras do futuro e temas correlatos, a intenção de colocar o ser humano no centro do processo de transformação digital: seja no intuito de fornecer a cada indivíduo boas e novas experiências, seja preocupando-se com a preservação da sua privacidade, com a proteção dos seus dados pessoais. A área de inteligência artificial ainda parece ser o grande foco das pesquisas.

3.1.2 Big Data/Analytics

Big data/Analytics é outra das forças que comandam a transformação digital. O conceito de big data incorpora várias novidades. Uma delas refere-se à variedade da fonte dos dados e informações. Os dados agora virão de todos os lados: dos departamentos da empresa, de fornecedores e clientes, mas, também, das redes sociais e dos objetos. Os dados estruturados das empresas, com padrões pré-definidos antes mesmo da sua produção, serão combinados com dados desestruturados (áudios, vídeos, infográficos, etc.), flexíveis, dinâmicos e desorganizados e, portanto, de interpretação difícil.

Outras novidades dizem respeito ao volume de dados, à velocidade com que podem ser capturados e à profundidade com que podem ser analisados. A quantidade a ser armazenada e tratada crescerá de modo muito significativo. Além disso, a análise de dados torna-se mais complexa. Enquanto os relatórios tradicionais concentram-se muito na descrição e no diagnóstico das situações, as análises de big data valorizam a previsão e prevenção. A sofisticação das análises irá requerer profissionais com um perfil diferenciado, capazes de combinar conhecimentos estatísticos complexos com insights sobre relações entre fenômenos que antes não haviam sido colocados juntos.

O processamento do alto volume de dados digitais que são gerados pelas pessoas durante a realização de atividades diárias, muitos dos quais elas não se dão conta, permite identificar

padrões de comportamento com precisão, monitorá-los e influenciá-los. Os dados pessoais são o novo petróleo da Internet e a nova moeda do mundo digital.

Com a chegada do big data, questões relacionadas com gestão, propriedade, proteção, armazenamento e privacidade de dados entram na ordem do dia.

3.1.3 Automação do trabalho do conhecimento

As máquinas inteligentes estão a caminho e cada vez mais fazem parte do nosso cotidiano. Robôs com habilidades diversas já colaboram com as tarefas diárias dos trabalhadores e, em alguns casos, irão substituí-los. O ambiente de trabalho e as relações no trabalho sofrerão, ao longo dos próximos anos, mudanças profundas. Inicialmente, a substituição de humanos por máquinas acontece em atividades rotineiras e de menor complexidade. No entanto, os trabalhadores do conhecimento, ou seja, profissionais em ocupações diversas que requerem a tomada de decisão e o processamento de dados e informações complexas, também crescentemente serão afetados à medida que a automação de processos básicos (Robotic Process Automation - RPA), ou seja, de atividades repetitivas e tarefas elementares, evolua sustentada pelas aplicações de inteligência artificial (IA).

A destruição do trabalho tal como conhecido hoje será acompanhada do surgimento de novas oportunidades, em profissões com contorno ainda pouco definidos. A necessidade de cruzamento de saberes, tradicionalmente tratados por separado (por exemplo, tecnologia, direito, psicologia, sociologia e negócios), passarão a ser fundamentais para garantir a empregabilidade. Assim, por um lado, e apesar das incertezas, existe urgência em repensar o papel das escolas e dos professores e rever o conteúdo a ser ministrado. Por outro, também se faz necessário preparar as organizações do futuro e garantir a empregabilidade dos trabalhadores.

3.2 Estágio Atual de Informatização no Setor Jurídico

Uma pesquisa da Associação Internacional de Tecnologia aplicada ao setor Jurídico (International Legal Technology Association - ILTA) [4] realizada, em 2021, envolvendo 460 firmas de advocacia, sendo 134 delas com menos de 50 advogados e 38 com mais de 700 advogados, mostra o grau de avanço no uso de tecnologias digitais por escritórios do setor jurídico. Os principais achados da pesquisa são os seguintes:

- A tendência mais clara na área é a adoção rápida das soluções em nuvem, exigência imposta pela covid-19 que requereu trabalho remoto/híbrido de advogados. Quase 25%

das firmas entrevistadas afirmaram que trabalham principalmente na nuvem e 41% responderam que já estão completamente na nuvem. As soluções localizadas na nuvem incluem ferramentas para gestão do conhecimento, calendário para registros de data e hora, marketing, litígios, gestão de documentos e conteúdo, contabilidade e finanças, além de sistemas para gerenciamento de aprendizado, e-discovery e outros.

- A nuvem também ganha força para armazenamento de backup de servidor (36% das firmas), com o backup em hard disk caindo 6% e respondendo por 25% das informantes.

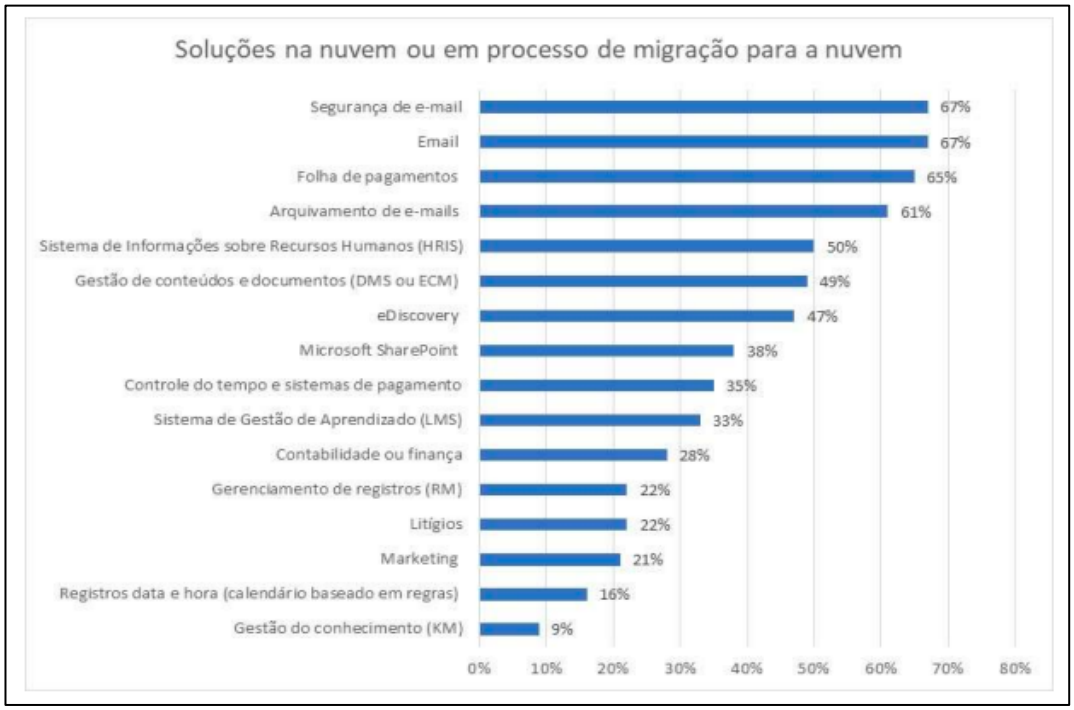

Figura 1.1: Proporção de firmas de advocacia com soluções na nuvem em processo de migração para a nuvem
Inclui respostas múltiplas.

Fonte: ILTA`s, 2021.

- Soluções em hardware voltam a ter destaque, com questões sobre onde imprimir documentos, quantos laptops a empresa precisa adquirir para distribuir para seus funcionários, se ainda vale a pena manter desktops na empresa, etc.
- A terceirização tem avançado na área da segurança. Uma tendência relacionada é a evolução rápida do papel da Governança da Informação, em que o conhecimento da corporação, compreensão da cultura e apreciação profunda das normas e requisitos éticos tornam-se aspectos muito relevantes.
- Sistemas de gestão de aprendizado (Learning Management System - LMS) têm se tornado cada vez mais voltados para plataformas móveis. E-learning ganha importância.

Outra tendência na área é a entrega de cursos curtos, com conteúdo reduzido, por meio das plataformas Teams ou Zoom.

- Aceitação da ideia de que as plataformas da Microsoft darão o suporte básico ao setor jurídico, sendo complementadas por aplicações específicas, a serem adotadas em cada caso.
- A busca por fornecedores de aplicações específicas para o setor jurídico tem aumentado nos últimos anos.
- Quase 70% das firmas entrevistadas usam MS-Teams para, no mínimo, atividades de chat e de colaboração e reuniões online. Conferências por áudio são usadas por cerca de 30% das firmas.
- O home office acelerou a adoção da digitalização e mudou os processos jurídicos. Com isso, cresce o emprego de assinaturas eletrônicas nos contratos e certidões. Observa-se igualmente a transformação das tecnologias e dos processos adotados pelos tribunais. Em especial, os procedimentos jurídicos que antecedem o julgamento estão muito diferentes do que eram em momento pré-pandemia.
- A automação de processos básicos (RPA) ganha impulso em algumas áreas (11% dos entrevistados possuem a tecnologia e 7% planejam utilizá-la nos próximos anos). Entre as firmas usuárias, a única ferramenta de automação com mais de 5% de respostas é a UIPath, plataforma para automação fim-a-fim que oferece soluções para automatização de tarefas de escritório repetitivas.
- A adoção de inteligência artificial/aprendizado de máquina (Machine Learning – ML) tem se mantido estável nos últimos 3 anos. Mais da metade das firmas entrevistadas não busca opções de AI/ML. Poucas contam com ferramentas de AI/ML já em funcionamento (Figura 1.2).

Figura 1.2 – Adoção de AI/ML pelas firmas de advocacia

Fonte: ILTA`s, 2021.

- A adoção de chatbots está sendo vagarosa: 14% das firmas fazem alguns testes ou têm planos para uso. Um percentual ainda elevado de firmas não adotou a tecnologia e nem tem planos de fazê-lo.
- A adoção de sistemas de Voz sobre IP (VoIP) aumentou de 10%, em 2018, para 27%, em 2021. O uso de fones tradicionais sofreu queda relevante, nos últimos anos.
- Entre as firmas de advocacia, cresce o uso do MS-Office 365 (MS-O365). A solução é utilizada para e-mails por 53% das firmas, em 2021, e 76% afirmam que MS-O365 será a sua plataforma de e-mail daqui a 12 meses.
- No que diz respeito às ferramentas de comunicação, Slack, MS-Teams e Zoom têm ganhado cada vez mais adeptos.
- Além da elaboração de documentos, as seguintes atividades foram automatizadas por 20% ou mais das firmas entrevistadas: consumo de informações, litígios, fluxo de trabalho de aprovação, captação de despesas, segurança da informação e admissão e demissão de pessoal (Figura 1.3).

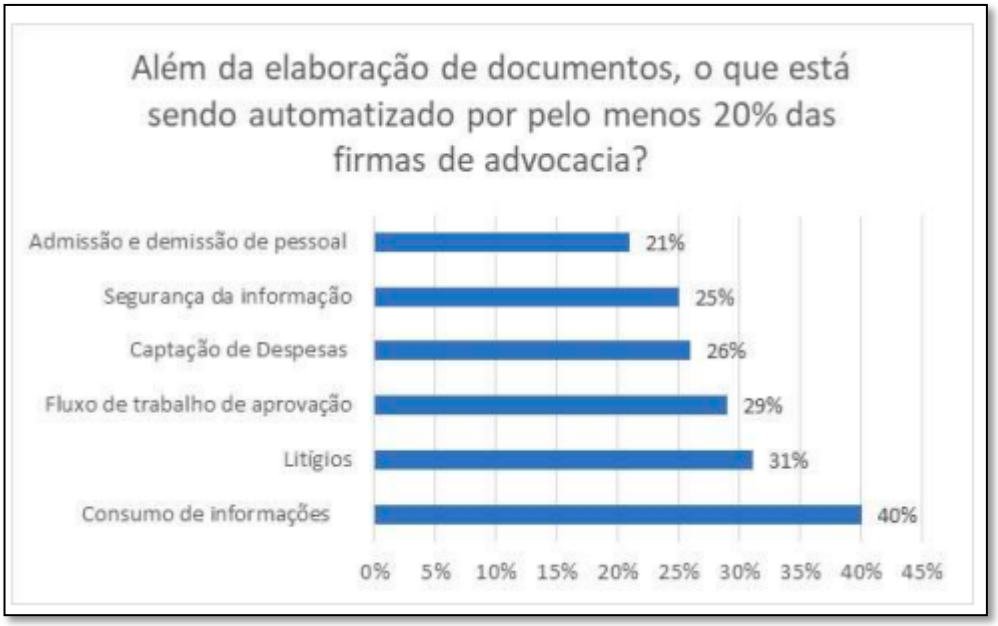

Figura 1.3 – Proporção de firmas com processos automatizados

Inclui respostas múltiplas.
Fonte: ILTA`s 2021.

- A Governança da Informação é questão relevante para as firmas de advocacia, por conta da bomba relógio escondida nas práticas de home office e em modelos híbridos de trabalho. O impacto é maior para firmas de grande porte, com questões transfronteiriças, jurisdições variadas e múltiplas. A preocupação das firmas é com a quantidade de e-mails que estão agora nas contas privadas de seus funcionários, em mídia removível, sendo impressos nas residências e não descartados de acordo com a política adotada pela organização. Além do advogado, outros residentes do domicílio podem acessar os computadores de propriedade da firma, fornecendo senhas e quebrando a segurança. As firmas também estão preocupadas com a possibilidade de perder o controle da produção de seus funcionários, vindo a responder por coisas que desconhecem, pois não constam do seu controle de registros.
- Outras preocupações de segurança relacionadas com a adoção do home office incluem: gestão do dispositivo remoto (consertos, inventário); insegurança dos WiFi usados nas residências; aprovisionamento e coleta de equipamentos fornecidos para uso nas residências; impressão de documentos jurídicos, muitas vezes sensíveis, no ambiente doméstico; uso ampliado de dispositivos pessoais pelos advogados (Figura 1.4).

Figura 1.4 - Principais preocupações das firmas de advocacia com segurança
Inclui respostas múltiplas.
Fonte: ILTA`s 2021.

- As soluções para gestão de documentos mais citadas são iManage, Net Documents e Intapp.
- As firmas do setor jurídico iniciam o processo de quebra dos silos entre produtos e fontes dispersas de dados, um requisito importante para adoção de big data. Atualmente, muitas empregam soluções de business intelligence (BI) e software para análises financeiras, mas não foi identificado fornecedor dominante.
- Observa-se um movimento desde o desenvolvimento de soluções in-house para a adoção de soluções de terceiros. Nessa linha, as firmas do setor jurídico parecem mais interessadas em adquirir soluções que incorporam AI/ML como um componente do que buscar opções de AI/ML disponíveis no mercado.
- A pandemia causou uma mudança abrupta no modo como os advogados interagem com os clientes e realizam serviços jurídicos. Advogados que estão trabalhando em domicílio tendem a adotar tecnologias/soluções que os tornam mais autossuficientes. Isso inclui, por exemplo, soluções para criar seus próprios templates, para comparar documentos e apoio para criação automática de documentos.

Os sócios das firmas começam a compreender que o seu foco precisa mudar. Antes estavam muito direcionados para a busca de documentos – precedentes, exemplos de acordos, resumos legais, etc. – e agora se preocupam mais com a aquisição de dados.

3.3 Estágio Atual de Informatização no Sistema Judiciário Brasileiro

As políticas judiciárias caminham para tornar os processos exclusivamente eletrônicos. Em 2019, a digitalização de processos judiciais alcançou a marca de 90% do total de processos existentes no país. Apesar dos esforços recentes e das várias iniciativas em andamento, a situação atual do uso das tecnologias digitais pelo poder judiciário ainda é caótica, como apontado em estudo da Insper de 2020, sobre a informatização judicial e os seus efeitos sobre a eficiência da prestação jurisdicional e o acesso à justiça.

No poder judiciário brasileiro, há uma grande multiplicidade de sistemas independentes entre si, muitas vezes não comunicáveis uns com os outros (Quadro 1.1). Sistemas de mesmo nome, e supostamente idênticos, podem também diferir na versão em uso e em virtude das adaptações implementadas localmente para atender a necessidades específicas, o que é uma prática disseminada.

Quadro 1.1 – Sistemas informatizados adotados pelos tribunais federais e estaduais de justiça no Brasil – 2019

SISTEMA	TRIBUNAIS DE JUSTIÇA
e-STF	Supremo Tribunal Federal (STF)
e-STJ	Supremo Tribunal de Justiça (STJ)
i-STJ	STJ
e-SAJ	Tribunais de Justiça (TJ) de SP; SC; BA; RN; CE; AC; AM; AL e MS
Themis	TJ de PI e MG
Tucurujis	TJ de AP
E-proc	Tribunal Regional Federal da 2ª (TRF-2) e 4ª (TRF-4) regiões. TJ de RS e TO
Projudi	TJ de GO; ES; RR; PR e RJ
EJUD	TJ de ES
PJe (1º grau)	TRF -1; TRF 3; TRF -5; TJ de RJ; PE; RN; RO; MG; MT; MA; PB; BA; CE; PI; DFT (Distrito Federal e Territórios); ES e PA
PJERJ	TJ de RJ
Apolo	TRF-2 (migrando para E-proc)

Fonte: Yeung, L. et al. "Informatização Judicial e Efeitos sobre a Eficiência da Prestação Jurisdicional e o Acesso à Justiça". Insper. Relatório Final, dezembro 2020.

4. FUNDAMENTAÇÃO TEÓRICA

4.1 *CRISP-DM*

A sistemática adotada nesse desenvolvimento terá como base a metodologia ágil para projetos de ciência de dados, intitulada *CRISP-DM iNuTech*. Esse método adota uma série de boas práticas utilizadas na indústria para entregar soluções inteligentes de maneira eficiente, de

modo a também promover melhorias nas entregas parciais e definir as responsabilidades dos membros do time.

A metodologia CRISP-DM iNuTech, tem como base o CRISP-DM 1.0, incorpora características de metodologias ágeis, introduz e detalha algumas fases atividades extras, tornando mais explícitas cada etapa metodológica seus e objetivos.

A apresenta as nove fases que compõe a metodologia *CRISP-DM iNuTech*. Ao final de cada fase ou iteração, deve ser possível definir uma das seguintes sequências para o projeto:

- Os objetivos e índices de qualidade previamente definidos foram alcançados e deve-se avançar para a fase seguinte;
- Os objetivos e índices de qualidade previamente definidos não foram totalmente alcançados, mas existem indícios suficientes para indicar o prosseguimento do projeto. Nesse caso, deve-se repetir o ciclo ou a fase em questão;
- Os objetivos e índices de qualidade não foram alcançados, e existem indícios suficientes para indicar a inviabilidade dos objetivos e o cancelamento do projeto.

A seguir, descreve-se os objetivos e atividades de cada uma das nove fases da metodologia:

- **Entendimento do Negócio**

  O entendimento profundo das áreas de atuação e áreas de negócio da instituição, e de seus respectivos processos organizacionais, serão os insumos para a formulação das perguntas estratégicas que o projeto pretenderá responder. Para a obtenção de resultados relevantes, é necessário que se consiga fazer as perguntas certas. E, para isso, torna-se necessária a participação ativa de especialistas nas áreas de domínio em estudo, com competência para identificar, selecionar e priorizar os problemas a serem tratados.

  Os profissionais mais envolvidos nessa fase são o analista de negócios, especialistas na área de domínio e o cientista de dados, com o apoio eventual do engenheiro de dados. O objetivo dessa fase é estabelecer uma compreensão geral sobre a área de negócio em estudo e sobre o problema a ser resolvido, explicitando a(s) pergunta(s) que serão encaminhadas à próxima fase. Essa fase faz parte de dois microciclos iterativos, de um lado com as fases de mapeamento de processos e de modelagem de negócio, e de outro com a fase de Entendimento de Dados, sendo constituída pelas seguintes macro-atividades:

  - Entendimento do problema alvo;

- Identificação de diretrizes organizacionais que norteiem o problema e a possível solução;
- Identificação de eventual legislação pertinente;
- Definição da estratégia de mineração que será desenvolvida;
- Pesquisa Bibliográfica:

  Nessa fase o iNuTech introduz uma atividade inicial de Pesquisa Bibliográfica, não prevista originalmente do CRISP-DM, com o objetivo de se conhecer mais profundamente as características do negócio, dos problemas que se pretende resolver e eventuais soluções ou pesquisas similares, reduzindo desse modo os riscos do projeto e também o número de microciclos iterativos das fases de Entendimento do Negócio e Entendimento dos Dados. O CRISP-DM original prevê na presente fase a obtenção de um *background* do negócio, mas que não intenciona o mesmo rigor aqui pretendido com a proposta do iNuTech.

- **Mapeamento e entendimento dos processos organizacionais**
  - Mapeamento e entendimento das atividades que se pretende automatizar;
  - Mapeamento dos processos organizacionais nos quais se encontram as atividades que serão afetadas pelo projeto.

- **Modelagem de negócios e da possível solução**
  - Discussão e descrição das possíveis soluções, apontando os benefícios que serão oferecidos pela solução;
  - Descrição de diferenciais das possíveis soluções em relação a soluções e produtos existentes no mercado;
  - Determinar o Problema de Negócio que se pretende tratar;
  - Avaliação e confirmação das perguntas de negócio, inicialmente levantadas na Fase de Entendimento do Negócio.

- **Entendimento dos Dados**

  Uma vez definidas as perguntas de negócio que serão o objeto do projeto, cientistas de dados, em conjunto com os profissionais de TI, deverão transformar essas questões de negócio em perguntas de dados.

  Nessa fase são examinadas as possíveis fontes de dados que serão usadas na estratégia de mineração. Muitas hipóteses levantadas podem evoluir ou serem descartadas já nesta fase, sendo comum, em função dos resultados, a necessidade de retorno à fase

anterior de Entendimento de Dados. Esse filtro de eliminação de hipóteses é um dos principais objetivos nesse momento, evitando a continuidade do trabalho por caminhos inadequados, ou até mesmo indicando a inviabilidade do projeto. Essa fase faz parte de um microciclo iterativo com a fase anterior de Entendimento do Negócio, sendo constituída pelas seguintes macro-atividades:

- Identificação, catalogação e qualificação das possíveis fontes de dados;
- Levantamento e análise de hipóteses de estratégias de mineração.

A correta especificação de dimensões e atributos necessários às respostas das perguntas de dados é determinante para a seleção apropriada de bases e amostras de dados, tarefa essa que pode ser dificultada tanto pelo tamanho e diversidade como pelo expressivo crescimento das bases disponíveis, sejam internas ou externas.

Desse modo, a definição da Área de Domínio do projeto, com suas respectivas dimensões e atributos, apresentada esquematicamente na Figura 6, é um dos marcos importantes, resultantes da Fase de Entendimento de Dados.

Definição de Área de Domínio:

- Definição de dimensões;
- Definição de atributos de cada dimensão;
- Avaliação de eventual necessidade de redução de dimensionalidade.

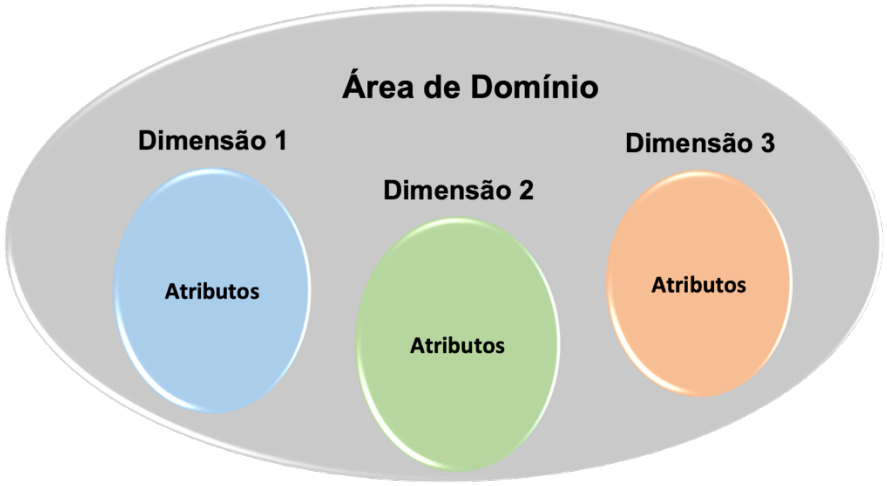

Figura 3 Área de Domínio. Fonte: iNuTech

- **Preparação dos Dados**

Essa é uma etapa realizada predominantemente por Engenheiros de Dados e por Analistas de Dados com o acompanhamento de Cientistas de Dados. O principal objetivo dessa fase é realizar todos os filtros e transformações necessárias para que os dados

tenham a melhor qualidade possível para serem usados. Várias técnicas podem ser empregadas de acordo com o contexto levantado nas fases anteriores e com a natureza dos dados coletados. Essa fase faz parte de um microciclo iterativo com as fases seguintes de Análise Exploratória e Modelagem, sendo constituída pelas seguintes macro-atividades:

- Eliminação de redundâncias e contradições;
- Correções de ambiguidades;
- Categorização;
- Sumarização;
- Tratamentos estatísticos.

- **Análise exploratória de amostra dos dados selecionada**

  Após a definição das perguntas de dados, da seleção, aquisição, tratamento e disponibilização de amostras de dados, inicia-se a o trabalho de análise exploratória, tem por objetivo a descoberta de padrões e correlações entre os diversos tipos de dados, e de preditores de maior relevância para os objetivos do projeto.

  Os resultados da análise exploratória de dados subsidiam escolhas e decisões sobre as possíveis alternativas para o projeto. Por meio da aplicação de técnicas e métodos diversos, separadamente ou em conjunto, como por exemplo abordagens estatísticas, aprendizado de máquina ou análise de redes complexas, obtêm-se como resultado análises de viabilidade técnica para as possíveis alternativas de soluções para o projeto.

- **Modelagem**

  Nessa fase os Cientistas de Dados atuam predominantemente com o acompanhamento de Engenheiros de Dados. É nesse momento que modelos analíticos descritivos, prescritivos, preditivos ou cognitivos são selecionados, criados, alterados, testados, calibrados e aplicados e/ou integrados, com a utilização de técnicas variadas, para a geração do "valor da informação", representado por um conjunto de novas de informações, inexistentes nas fontes originais, que revelam padrões, comportamentos e relações que podem responder às perguntas de negócio e resolver o problema proposto nas fases iniciais.

  Para cada modelo aplicado, gera-se um índice de eficácia usado como prova matemática e também para estabelecer um score de aferição de efetividade do modelo ao longo do tempo, na medida em que os dados das fontes evoluírem. Essa fase faz parte de um

microciclo iterativo com a fase anterior de Preparação dos Dados, sendo também esperada a repetição isolada da fase. As seguintes macro-atividades constituem essa fase:

- Estudo, seleção e testes de algoritmos;
- Calibragem dos algoritmos selecionados;
- Possíveis alterações no código dos algoritmos selecionados;
- Construção dos modelos analíticos.

- **Avaliação:**

  Nessa fase, o principal profissional envolvido é o Analista de Negócios, apoiado pelo Cientista de Dados, tendo por tarefa avaliar os modelos analíticos, os valores resultantes, os scores dos modelos e a qualidade do processo como um todo. Essa fase pode, eventualmente, determinar a necessidade de novos ciclos iterativos, englobando quaisquer das fases anteriores, para atendimento dos objetivos e critérios de qualidade estabelecidos nas fases iniciais do projeto. Essa fase é constituída pelas seguintes macro-atividades:

  - Validação dos Modelos Analíticos;
  - Verificação do processo de modelagem.

- **Implementação:**

  Finalmente, o Analista de Negócios e o Cientista de Dados definem os requisitos necessários para a construção de uma aplicação de software, seja um protótipo ou um produto final. Essas definições devem englobar as estratégias de apresentação e consumo dos dados para os diversos perfis de usuários atendidos pela aplicação, principalmente para os tomadores de decisão da organização. Gráficos e painéis são funcionalidades previstas na camada de apresentação da aplicação. Essa fase é constituída pelas seguintes macro-atividades:

  - Construção de uma aplicação de software;
  - Construção de camada de apresentação com representações gráficas, estáticas ou interativas, voltadas para os diferentes perfis de usuários.

4.2 Rotulação

Abordagens supervisionadas de aprendizado de máquina e processamento de linguagem natural demandam o uso de conjuntos de dados textuais rotulados, que definem pelo menos um atributo associado à categoria (ou classe) para cada instância. Esses modelos matemáticos possuem parâmetros que, com o processo de treinamento, aprendem os padrões e propriedades das instâncias de cada uma das classes presentes em um conjunto. Dados são naturalmente coletados ou obtidos sem quaisquer informações de categoria, demandando dos especialistas a tarefa de rotular cada instância para que esses dados sejam empregados em modelos baseados em aprendizado supervisionado.

A categorização ou rotulação de dados textuais é uma tarefa difícil e complexa, sendo comumente identificadas três abordagens: automática, semi-automática e manual. Na abordagem automática, pode-se citar o aprendizado não-supervisionado como uma estratégia clássica para obtenção dos rótulos a partir das subestruturas (grupos) de instâncias similares entre si (Asano et al., 2019). Como exemplos dessa estratégia, pode-se citar técnicas baseadas em aprendizado fraco (Ratner et al., 2017) e modelos de tópicos (Blei et al., 2003). Já nas abordagens semi-automáticas, os rótulos das instâncias de texto são parcialmente obtidos utilizando algoritmos semi-supervisionados de aprendizado de máquina, envolvendo a presença de especialistas do domínio no fornecimento de rótulos. Nesse cenário, uma das abordagens mais conhecidas é o aprendizado ativo (Settles et al., 2008), que será detalhado na Seção 4.3. Entretanto, as abordagens automáticas as e semi-automáticas não garantem a construção de corpos de textos de alta qualidade uma vez que os rótulos obtidos dependem da qualidade e da representatividade dos dados disponíveis, como também dos algoritmos de aprendizado de máquina em consideração.

Como alternativa, pode-se considerar a anotação manual para a construção de corpos de texto com o emprego de anotadores humanos. Assim, o processo de anotação pode levar a criação de corpos de texto de alta qualidade, uma vez que os anotadores utilizam seus conhecimentos para realizar as marcações no texto, como também no processo de revisão por pares que visa resolver problemas de discordância. Os corpos de texto construídos por meio de anotação manual e que apresentam alto grau de confiabilidade nos rótulos são conhecidos como padrão-ouro (Gold Standard Corpus) (Wissler et al., 2014) (Juckett D., 2012). Todavia, o processo de anotação para a obtenção de corpos de texto padrão ouro é complexo, custoso e caro pela definição das diretivas de anotação e do treinamento dos anotadores, além do minucioso processo de revisão por pares.

Vale ressaltar que o tipo de corpo de texto a ser construído depende da tarefa alvo e do domínio do conhecimento associado aos textos. Por exemplo, o processo de anotação para a classificação de documentos de texto consiste em atribuir rótulos para documentos, seções ou blocos de texto. Já para tarefas de reconhecimento de entidades nomeadas, a anotação demanda a marcação de palavras ou segmentos de texto que representam entidades específicas, sendo considerada um processo mais minucioso e detalhista do que a anotação de textos para classificação. Assim, a anotação deve ser customizada considerando diversos aspectos como a quantidade de anotadores disponíveis, a quantidade estimada de documentos ou entidades a serem anotados, a ferramenta de anotação e o tempo estimado para a conclusão da anotação.

Como mencionado anteriormente, corpo de texto padrão-ouro devem possuir alto grau de concordância entre anotadores em relação aos rótulos anotados. Por isso, é importante reservar uma parte dos dados textuais destinados à rotulação visando conduzir experimentos de avaliação da qualidade do corpo de texto em construção (Bowman et al., 2015), (Chakravarthi et al., 2020). Os experimentos de avaliação consistem em atribuir textos a serem rotulados para diferentes anotadores e assim calcular medidas estatísticas para avaliar a concordância de anotação (Bobicev et al., 2017) e a similaridade entre os segmentos de textos anotados por meio dos coeficientes Dice e de Jaccard (Jimenez et al., 2020).

Tradicionalmente, existem três critérios estatísticos para avaliar a concordância entre anotadores para objetos ou entidades categóricos: os coeficientes Cohen Kappa (Cohen J., 1960), Fleiss Kappa (Fleiss J., 1970) e Krippendorf's alpha (Krippendorff K., 2011). O coeficiente Cohen Kappa mede a concordância de rótulos considerando dois anotadores trabalhando simultaneamente nos mesmos documentos. O Cohen Kappa assume que os rótulos são aleatoriamente atribuídos às anotações sendo descritos por uma distribuição a priori para cada anotador. Já o coeficiente Fleiss Kappa generaliza essa ideia para o cenário de múltiplos anotadores, em que a concordância esperada é calculada com base no fato de que a atribuição aleatória de rótulos para as anotações, para qualquer anotador, é caracterizada pela distribuição das anotações entre os possíveis rótulos. Por fim, o coeficiente de Krippendorf's calcula a concordância com base na distribuição global das anotações, sem se preocupar com correspondências de anotação e anotador. Nesse sentido, esse critério é empregado em processos de anotação que possuem múltiplos anotadores e também é robusto à presença de valores ausentes.

4.3 Recuperação da Informação

Técnicas de Mineração de Texto podem ser utilizadas para identificar documentos de textos semelhantes. A identificação de processos (documentos de textos) é considerada uma forma de recuperação de informação, na qual objetiva-se extrair automaticamente informações associadas aos dados que se apresentam de natureza não estruturada.

Na literatura científica da área de Recuperação de informação, encontram-se várias técnicas que apresentam bons resultados, mas que também podem ter desvantagens para diferentes contextos, condições de estruturação dos dados e desempenho. Por isso, para facilitar a análise, essas técnicas são agrupadas em abordagens relacionadas às suas características comuns. Inicialmente, as abordagens vetoriais, ou "Saco de Palavras", representam documentos por vetores cujas dimensões correspondem às frequências de cada termo do vocabulário que será considerado. Essas técnicas apresentam bons resultados e estão presentes na maioria dos sistemas de indexação de documentos. Técnicas como LSI (Latent Semantic Indexing) (Hofmann, T., 1999), NMF (Non-negative Matrix Factorization) (Berry et al., 2007) e Okapi BM25 são adequadas para agrupar e encontrar a relação de similaridade entre os textos. Essas técnicas possuem a desvantagem de necessitar a alocação de toda a base para memória principal, o que a torna inviável para o contexto de grandes quantidades de dados.

Uma outra abordagem que se apresenta adequada para dados massivos, são aquelas baseadas em modelos probabilísticos. Modelos probabilísticos de tópicos, como o modelo base LDA (Latent Dirichlet Allocation) (Blei et al., 2003), possuem processo estocástico de inferência, além de possibilitar adaptações e adição de novos atributos de dados, além de simplesmente palavras. Por fim, as abordagens que exploram a representação baseada em imersão de palavras (word embeddings). Essas técnicas advêm de modelos de linguagem, área que apresenta opções robustas no estado-da-arte em várias tarefas de processamento de linguagem natural. Em especial, destacam-se técnicas que utilizam embeddings de contexto como arquiteturas neurais baseadas em Transformers (Lin et al., 2020).

Em função das características peculiares do contexto desta proposta, alguns autores propõem uma área específica dentro da área de recuperação de informação. Van Opijnem (Van Opijnen et al., 2017) aborda essas características e propõe abordar a relevância de documentos legais em seis dimensões, que são as seguintes. (i) Relevância algorítmica: Trata-se da semelhança entre o texto de pesquisa e o texto dos documentos atribuída por algoritmos, como ocorre na maioria dos sistemas de recuperação de informações. (ii) Relevância tópica: Trata-se da correspondência entre assunto ou tópico pesquisados com e os documentos recuperados. (iii)

Relevância Bibliográfica: Trata-se do alinhamento e coerência bibliográfica dos documentos recuperados. Esta dimensão da relevância mede a pertinência dos documentos no que se refere às legislações e normas referenciadas. (iv) Relevância Cognitiva: Trata-se da opinião ou preferência do pesquisador pelo modo como a questão legal é abordada nos documentos recuperados. (v) Relevância Situacional: Trata-se das características da atividade desempenhada pelo especialista para as quais busca suporte nos documentos. Em uma dada situação, o especialista busca por documentos que contenham argumentos de acusação, e em outras por argumentos de defesa. Por fim, (vi) Relevância de domínio: Refere-se à importância ou destaque atribuído ao documento pela comunidade jurídica.

Na literatura há diversas abordagens para estimar a relevância dos documentos legais. Entre as abordagens mais comuns está o uso de ontologias para mapear os conceitos presentes nos escritos jurídicos. Aqui, nesta proposta, descarta-se o uso de ontologias. Contudo, como é esperado no caso jurídico dos tribunais brasileiros, existe um grande espectro de temas jurídicos, o que inviabiliza a criação de taxonomias para um contexto tão amplo. Além disso, as frequentes alterações legislativas e volume jurisprudencial exigem constantes atualizações das ontologias.

Com isso, com o objetivo de melhorar a recuperação de informação, e também a similaridade entre os documentos, pretende-se usar não apenas modelos baseados em sacos de palavras, mas também utilizar características do contexto. Pode-se, por exemplo, explorar o uso de citações ou referências legais encontradas nos documentos. As citações informam o relacionamento do documento com as normas legais, e são utilizadas para enriquecer as soluções de recuperação de informação (RAGHAV; 2016).

4.4 Aprendizado Ativo

Modelos de aprendizado de máquina alcançam alto desempenho em várias tarefas, para isso, é necessário um grande conjunto de dados rotulados para treinamento supervisionado. Todavia, dois grandes problemas podem ocorrer na obtenção de grandes quantidades de dados rotulados: (1) dificuldade de adquirir novos exemplos devido a pouca ocorrência dos dados, e (2) o custo de rotular todos os conjuntos de dados excede o orçamento do projeto. A técnica de aprendizado ativo vem como solução para o último problema. Essa técnica assume que um grande conjunto de dados pode ser bem representado por uma pequena quantidade de exemplos, e idealmente, um modelo treinado com um pequeno subconjunto de dados tem desempenho próximo a um outro modelo treinado por uma grande quantidade de dados.

Nesse sentido, modelos baseados em aprendizado ativo são projetados para encontrar um pequeno subconjunto de amostras que podem ser anotadas por um oráculo (um humano). E com esse subconjunto de amostras, é possível treinar modelos com bom desempenho e que sejam capazes de identificar, com alta confiança, mais dados rotulados.

Em especial, para essa proposta, acredita-se que modelos de aprendizado ativo podem ser aplicados no conjunto de documentos não rotulados. Espera-se que seja possível recuperar, por meio de diversas funções de amostragem, um ranque de documentos mais informativos e que possam ser selecionados para construção da base rotulada.

De modo objetivo, pode-se descrever o aprendizado ativo por três partes principais: (1) uma estratégia para criar um conjunto inicial de dados rotulados, (2) uma função de amostragem para escolher os documentos mais informativos, e, por fim, (3) um critério de parada do algoritmo de aprendizado ativo.

Para a estratégia de criação do conjunto inicial, considerando o cenário onde não existem dados rotulados, pode-se aplicar medidas de informação mútua entre os documentos rotulados e agrupá-los. O agrupamento pode ser realizado por algoritmos como o K-Means ou técnicas de relação de similaridade semântica (como o Latent Dirichlet Allocation). Com a análise de documentos amostrados de cada grupo, o oráculo pode selecionar amostras informativas e diversidade.

Uma segunda parte importante, considerado o componente principal da técnica de aprendizado ativo, é a função de amostragem, ou estratégia de consulta. Essa estratégia é responsável pela seleção dos exemplos mais informativos do conjunto de documentos não rotulados e, consequentemente, para a apresentação desses documentos para a anotação do oráculo. Para essa proposta, acredita-se que funções baseadas em incerteza e diversidade, ou a combinação das duas, possam ser adequadas. Funções baseadas em incerteza assumem que as amostras não rotuladas que podem trazer mais informação ao modelo são aquelas que o modelo tem mais incerteza na predição. Os exemplos com menor confiança deve ser anotados pelo oráculo. Já funções baseadas em diversidade assumem que exemplos que trazem maior diversidade a base rotulada podem ser mais informativos ao conjunto a ser anotado.

O aprendizado ativo, de forma interativa, recupera exemplos não rotulados para a rotulação do oráculo. Com isso, objetiva-se maximizar o desempenho da classificação com a menor quantidade de dados anotados. O momento de finalização da anotação é definido pelo critério de parada, conforme: (i) o desempenho obtido na classificação de um conjunto de teste pré-definido; (ii) o valor do gradiente, que durante o treinamento pode ficar razoavelmente baixo,

indicando que pouca informação está sendo obtido dos dados rotulados; (iii) o critério de confiança, que pode alcançar um alto valor para os exemplos não rotulados; ou mesmo (iv) por especificidades e características do modelo. De qualquer forma, todos esses critérios de parada podem ser calculados para obter resultados confiáveis durante o processo de rotulação.

Por fim, é importante destacar que existem outras estratégias que podem ser utilizadas para tratar o problema da escassez de dados rotulados, como Transferência de conhecimento, Aprendizado semi-supervisionado, auto-aprendizado. A estratégia baseada em aprendizado ativo, juntamente com a rotulação realizada por um oráculo, pode ser apropriada para as tarefas de classificação previstas nesse projeto.

4.5 Supervisão Fraca

Nesta última década, o aprendizado profundo tem alcançado sucesso considerável em tarefas de classificação de dados textuais, evitando grande parte do processo manual de engenharia de características. Em dados de domínio específico, como o jurídico, a elaboração de características pode ser difícil e custosa. Para aproveitar características latentes nos dados, o processamento em redes neurais profundas é composto por arquiteturas complexas que requerem uma grande quantidade de dados. Com isso, chega-se novamente à necessidade de dados rotulados. Mesmo as técnicas de aprendizado ativo podem ser insuficientes para suprir a necessidade de dados rotulados que esses modelos exigem.

Para reduzir os encargos da anotação de dados de treinamento, uma estratégia interessante, e promissora para o contexto deste projeto, é recorrer a fontes mais baratas de dados rotulados. Acredita-se na possibilidade de usar técnicas baseadas em distância para encontrar documentos similares aos já anotados, aproveitar modelos já treinados e que também foram úteis para a estratégia de aprendizado ativo, ou mesmo utilizar heurísticas e regras previamente conhecidas por especialistas do domínio.

Essas estratégias de automatização da rotulação são propostas recentes, mas já apresentam frameworks programáticos para essas atividades. Especificamente, os usuários codificam fontes de supervisão fraca, por exemplo, heurística, bases de conhecimento e modelos pré-treinados, na forma de funções de rotulagem. As funções de rotulagem são rotinas definidas pelo usuário e que são capazes de fornecer rótulos para algum subconjunto dos dados. Essas funções devem ser variadas de modo que, coletivamente, geram um grande conjunto de rótulos de treinamento.

Os frameworks para o aprendizado fraco possibilitam a incorporação das várias funções de rotulação. É esperado que os rótulos gerados por essas funções de rotulação sejam ruidosos e conflitantes. Para tratar esse problema, os frameworks possuem agregadores de funções de rotulagem que produzem dados rotulados baseados no voto (ou rotulação) fornecido por essas funções.

As funções de rotulação são codificações variadas. Em aplicações práticas, um especialista do domínio pode criar expressões regulares que indicam a rotulação de alguns documentos. Algumas estratégias utilizam bases externas para encontrar alguma relação de similaridade entre os dados não rotulados e seus supostos rótulos. Além disso, pode-se utilizar modelos pré-treinados, que não apenas indiquem os rótulos, mas que possam realizar tarefas, como por exemplo, detecção de entidades nomeadas, e essas entidades podem auxiliar na indicação do rótulo.

Portanto, acredita-se que a supervisão fraca ofereça uma direção promissora para aumentar o volume de dados rotulados e que também diminua o esforço humano.

4.6 Representação Computacional de Termos

Embedding de palavras criam representações semelhantes para palavras que não possuem características comuns. Isso representa palavras e frases digitalmente, como uma lista de números. *Word Embedding* pode ser definido como um conjunto de vetores de números reais que representam palavras em um espaço *n*-dimensional. Esses vetores são gerados a partir da aplicação de um algoritmo sobre uma base textual, de forma que os vetores numéricos sejam capazes de representar os aspectos morfológico, sintático e semântico desse conjunto de dados textuais (Hartmann, N. *et al*., 2017). Esses vetores são relevantes e considerados como dados de entrada para soluções de aprendizagem de máquina que não consideram representação textual, mas sim representações numéricas. Existem diferentes técnicas para criar essa representação, cada uma com vantagens e desvantagens. Diferentes algoritmos têm sido propostos para gerar *Word Embeddings*. Por exemplo: Global Vectors (Glo-ve), Word2Vec, Wang2Vec, FastText and ELMo (Peters *et al*., 2018). Neste projeto, o uso de *Word Embeddings* será relevante para representar textos jurídicos de entrada para o treinamento de modelos. Uma tarefa relevante será a aplicação e avaliação das diferentes técnicas existentes no contexto do estudo.

4.7 Modelos de linguagem

A modelagem da linguagem é o processo de prever a chance de uma sequência específica de palavras aparecer em uma determinada frase. Os modelos de linguagem que geram texto como saída executam a modelagem de linguagem como parte de sua fase de treinamento. Modelos pré-treinados são considerados a espinha dorsal de vários sistemas modernos de processamento de linguagem natural (Qiu et al., 2020). Howard e Ruder (2018) propuseram um modelo e o ajustaram a uma infinidade de tarefas com resultados bem-sucedidos. Desde então, a abordagem padrão para muitas tarefas em processamento de linguagem natural tem sido ajustar um enorme modelo de linguagem pré-treinado na tarefa de destino específica para arquivar bons resultados. Em particular, neste projeto visamos explorar o conceito de treinamento currículo (Bengio et al., 2009) em que aplicamos uma sequência ordenada de refinamento de modelos de linguagem. A partir dos modelos mais genéricos de linguagem, o processo de treinamento sequencial segue em direção ao treinamento de modelos mais específicos para tarefas mais específicas. Nesse processo, o reuso de modelos existentes é uma prática. Como ponto de partida, visamos explorar modelos já treinados na Língua Portuguesa, como exemplo: BERTimbau (Souza *et al.*, 2020). Alguns modelos de última geração foram treinados usando técnicas de modelagem de linguagem, como BERT (Devlin et al., 2020) e GPT-2 (Radford et al., 2019). BERT foi alimentado com uma sequência de palavras, com 15% das palavras mascaradas, e deve produzir a sequência correta sem palavras mascaradas. GPT-2 foi alimentado com uma sequência de palavras e visou prever a próxima palavra dessa sequência.
Esses modelos exploram principalmente os conceitos de codificadores e decodificadores sendo uma família de modelos que aprendem a mapear dados em um domínio de entrada para um domínio de saída por meio de uma rede com dois estágios. A Arquitetura *Transformer (*Xia *et al.,* 2020) refere-se a uma arquitetura de rede neural baseada inteiramente em mecanismos de atenção (Bahdanau et al., 2015) em vez de Redes Convolucionais e Recorrentes. Esta proposta é composta de um codificador e um decodificador. O codificador processa uma sequência de informações como o texto para construir um vetor n-dimensional (representação da frase de entrada - mapeada para uma sequência contínua de representações). Então, este vetor é processado pelo decodificador para gerar outra sequência de saída. Nesse sentido, o *Transformer* é um modelo de sequência a sequência.
O *Transformer*, apresentado por Vaswani et al. (2017), é um modelo de transdução sequencial baseado inteiramente em modelos de atenção, substituindo camadas de recorrência por atenção própria multifacetada. Bahdanau, Cho e Bengio (2016) propuseram uma extensão ao

modelo de Máquinas Neurais de Tradução, tornando-o capaz de buscar por partes da sentença de entrada relevantes para a predição de uma palavra de saída. A busca por partes da sentença de entrada foi chamada, intuitivamente, de atenção. Desse modo, o decodificador decide quais partes da sentença de entrada ele deve 'prestar atenção'. O *Transformer* utiliza um empilhamento de atenção próprio e camadas completamente conectadas tanto para o codificador quanto para o decodificador.

Em geração de linguagem natural, dado um contexto predefinido, a tarefa de geração de linguagem consiste em sintetizar uma sequência de palavras com significados contextuais e construção gramatical correta. Modelos de linguagem que efetuam auto-regressão sensível ao contexto (Radford et al., 2019) possuem características para inferir sequências de textos com uma resposta. Usando a enorme capacidade de generalização que os modelos de linguagem pré-treinados tiveram em outras tarefas de processamento de linguagem natural, Radford *et al.* (2019) propuseram o GPT-2, que utiliza o decodificador de uma arquitetura *Transformer* para o treinamento do modelo. Essa arquitetura aprendeu a previsão da próxima palavra em um vasto conjunto de dados (~40 GB de texto) e muitas tarefas relacionadas à geração de linguagem. Tarefas relacionadas com o resumo de texto e geração de resposta foram aprimoradas pelo ajuste fino desse modelo pré-treinado. Mais recentemente, houve o lançamento do GPT-3 (Brown *et al.,* 2020) como uma nova versão do modelo com uma quantidade maior de parâmetros.

O BERT (Vaswani et al., 2017), ao contrário do GPT, utiliza um modelo de linguagem mascarada aplicado à parte do codificador do *Transformer;* e utiliza o token de sequência inicial (<CLS>) para realizar uma tarefa de classificação. Portanto, seu pré-treinamento consiste em um objetivo multitarefa. O BERT pré-treinado obteve o estado da arte em inferência de linguagem natural, reconhecimento de entidade nomeada e conjuntos de dados de resposta a perguntas.

5. OBJETIVOS

A partir da utilização de tecnologias digitais aplicadas ao setor jurídico e o sistema judiciário, pretende-se alcançar os seguintes objetivos:

a) Detecção de significado nas peças processuais: inclui identificação das alegações, exame de admissibilidade, cálculo da probabilidade de concessão de medidas cautelares;

b) Painel de jurimetria: inclui priorização de processos e comparação com causas anteriores.

c) Redação de peças: inclui geração de comunicações aos interessados e de instruções contendo sumarização de teses e predição da análise técnica e das propostas de encaminhamento.

d) Desenvolvimento e avaliação funcional da plataforma computacional visando a integração das tecnologias desenvolvidas

6. METODOLOGIA DE EXECUÇÃO DAS ETAPAS E CRONOGRAMA

A sistemática adotada no desenvolvimento de software terá como base uma metodologia ágil de ciência de dados, intitulada *CRISP-DM*. Esse método adota uma série de boas práticas utilizadas para entregar soluções inteligentes de maneira eficiente, de modo a também promover melhorias nas entregas parciais e definir as responsabilidades dos membros do time. O ciclo de vida do *CRISP-DM* contém 4 estágios principais:

- Entendimento do Negócio
- Aquisição e entendimento dos dados
- Modelagem
- Implantação

Sobre o ponto de vista da pesquisa e inovação, a metodologia proposta neste projeto possui características específicas e exige uma abordagem particular. A Figura 6.1 apresenta metodologia que guiará a pesquisa desenvolvida nesta proposta. É importante salientar que nossa estratégia orientadora é independente das tecnologias, dos volumes de dados ou das abordagens envolvidas. Nota-se que o ciclo iterado de desenvolvimento do CRISP-DM está incluso como um componente formador da metodologia.

Processo Offline – Treinamento

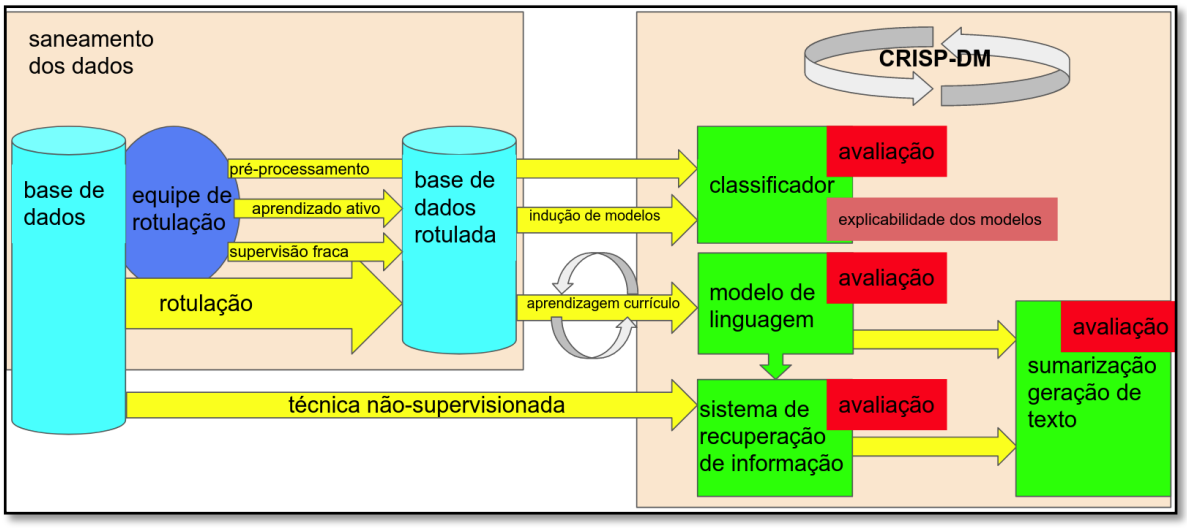

Figura 6.1 – Metodologia geral para o desenvolvimento da pesquisa e inovação. Os quadrados verdes representam modelos treinados para os diferentes objetivos desta proposta.

Na estratégia geral, um grande esforço será orientado para os grupos de tarefas relacionadas ao saneamento de dados e desenvolvimento dos modelos. E, seguida, descrevemos as atividades e os processos especificados na metodologia:

- Com a **base de dados** fornecida pelo tribunal, pretende-se realizar a coleta e **saneamento dos dados** para a integração dos dados, correspondentes a extração, transformação e carregamento.
- **Rotulação** da base de dados pela **equipe de rotulação**. A rotulação deve incluir a identificação de segmentos, entidades e classe de peça processual. Nessa etapa considerando a utilização de estratégias como **Aprendizado Ativo** e **Supervisão Fraca** para auxiliar a rotulação. Adicionalmente, investigaremos como a rotulação deve ser efetuada para atender as diferentes tarefas no projeto. Estamos assumindo que anotações de elementos como entidades nomeadas serão úteis para diversas tarefas. Contudo, tarefas específicas como sumarização e geração de texto podem demandar anotações específicas que deverão ser estudadas. Fato relevante é que a rotulação deve ser base para todas as tarefas de indução na proposta.
- **Pré-processamento** dos dados para a correção de possíveis problemas estruturais como desbalanceamento, valores ausentes, ruídos e duplicidades de dados.
- **Indução de modelos** de Aprendizado de Máquina para **classificação** e cálculo de probabilidade das peças processuais. Essa etapa deve incluir a análise e **avaliação** de diversos algoritmos do estado-da-arte.
- Definição da estratégia de **avaliação** dos resultados considerando as medidas mais adequadas para problemas multiclasse. O desbalanceamento das classes é um fator importante e deve ser considerado na análise das medidas mais adequadas.
- A **explicabilidade de modelos** é uma característica desejável para que possamos utilizar os modelos com maior garantia e qualidade. Investigaremos técnicas para melhorar a interpretabilidade dos modelos, principalmente daqueles que atuam como caixa preta.
- **Técnicas não supervisionadas** de **Recuperação de Informação** e Análise Semântica de documentos para calcular o grau de similaridade entre documentos de texto. Ser capaz de pesquisar informações é uma grande necessidade do trabalho moderno, sendo que a capacidade de localizar informações em ambiente virtual com rapidez e eficiência é fator diferencial para aumentar a qualidade e produtividade.

- A aplicação de técnicas de Análise de Similaridades semântica entre sentenças para auxiliar na qualidade da informação recuperada. Isso se dá pela capacidade de **modelos de linguagens** de correlacionar semanticamente textos curtos. Apesar de que os documentos sejam textos longos, artefatos produzidos em modelos de linguagens podem auxiliar outras tarefas.
- Estratégia de **avaliação** de resultados, sendo que as **técnicas de recuperação de informação** são **não-supervisionadas**. Com isso, é necessário determinar uma base padrão ouro, ou estratégias para avaliar a qualidade da similaridade com o auxílio de partes envolvidas e especialistas do domínio.
- Definição de uma metodologia baseada em **aprendizagem currículo** para refinamento de **modelos de linguagem** visando **geração de texto** no contexto jurídico.
    - Seleção de **modelos de linguagem** adequados à tarefa. Isso envolve a avaliação e teste preliminar conceitual e experimental de modelos existentes. O trabalho resultará em um subconjunto chave de modelos candidatos mais adequados ao contexto do projeto para as tarefas de interesse.
    - Método para refinamento de **modelos de linguagem** visando o aprendizado para a **geração automática de textos** não estruturados. Investigação de técnicas para considerar como modelos existentes são re-treinados com dados do domínio para a tarefa alvo.
- Concepção de técnicas para a **sumarização** automática de textos com base em **modelos de linguagem** refinados para o contexto de aplicação. Neste estágio o estudo visa investigar modelos treinados que contribuam para a **sumarização** de textos.
- Definição de um método para a **geração automática de textos**. Como requisito chave para a **geração de textos**, é necessário que a redação resultante da máquina expresse linguagem compreensível.
- **Avaliação** experimental dos métodos.
    - Implementação computacional de uma prova de conceito
    - Descrição do protocolo experimental incluindo métricas de **avaliação** e de sucesso nas análises experimentais.
    - Condução dos experimentos explorando a solução computacional desenvolvida.

No processo de desenvolvimento, seguiremos as seguintes etapas. A Figura 6.2 detalha a metodologia de desenvolvimento e o processo de aplicação dos modelos treinados.

- Criar sistemas computacionais que possibilitem a ingestão de dados a serem fornecidos pelos parceiros do TCU de forma contínua visando a melhoria dos modelos a serem implementados.
- Criação de interfaces para integração da redação de peças para integração com sistemas computacionais existentes, considerando acessos via web, mobile e demais plataformas a serem definidas.
- Criação de sistemas para medição de métricas dos resultados obtidos possibilitando a qualificação e refinamento dos modelos de aprendizado de máquina utilizados.

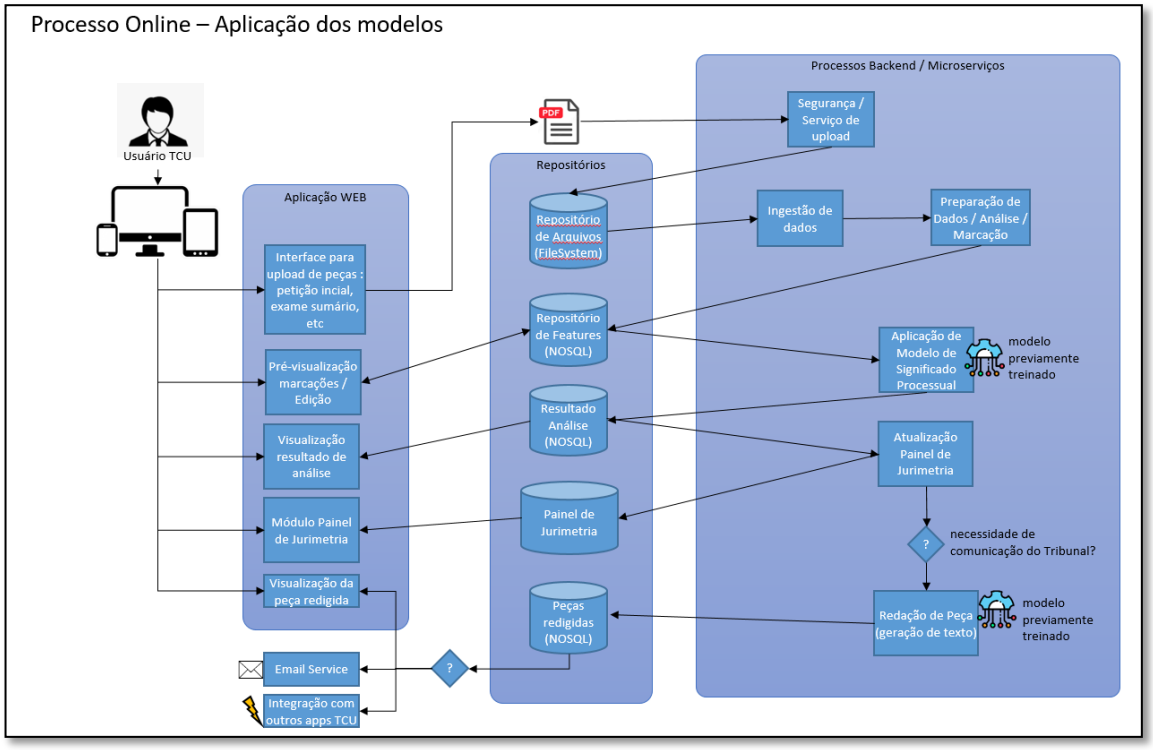

Figura 6.2 – Metodologia para o desenvolvimento e aplicação dos modelos.

A arquitetura de soluções de Inteligência Artificial se divide em Processos Online e Offline, conforme ilustra a Figura 6.2, os quais são Pipelines de aprendizagem de máquina que possuem propósitos diferentes. Os Processos Online referem-se às predições efetuadas durante a operacionalização do modelo treinado aprendizagem de máquina, ao passo que os Processos Offline estão relacionados ao treinamento dos modelos.

Dessa forma, organizamos o projeto em duas etapas bem definidas:

1. Processo Offline: Processo e ferramentas para treinamento dos modelos de ML;
2. Processo Online: Plataforma final, destinada aos usuários do TCU, onde os modelos de ML treinados na etapa 1 serão aplicados a novos documentos (do mesmo tipo de processos previamente treinados), à medida que forem requisitados.

A plataforma final (etapa 2) consiste em um aplicativo web, serviços de backend e repositórios destinados a manter o sistema em funcionamento.

Entregáveis / produtos da plataforma:

1. Módulo de cadastramento, validação e autenticação de usuários; por tratar de uma ferramenta que acessa dados processuais sujeitos a restrições de exibição ou sigilo, a ferramenta requer um controle de acesso, para garantir que somente pessoas autorizadas tenham acesso à plataforma.
2. Módulo para análise de processos e documentos: a plataforma oferecerá um conjunto de telas que permite ao usuário TCU submeter novos documentos (de tipos previamente treinados) ao sistema, para que seja feita a análise automatizada, fazendo uso dos modelos treinados.

   O processo de análise tem como início o upload de um PDF. Ao fazer o upload de um novo arquivo PDF, o sistema terá o serviço de Upload / Segurança para validá-lo, fazer as verificações de segurança necessárias e armazenar o arquivo em um repositório digital (File System), dando início ao processo.

   Uma vez que o arquivo teve o upload concluído, os Serviços de Ingestão de dados de e Preparação de Dados / Análise / marcação devem lidar com workloads de diferentes naturezas (streaming ou em lotes) e são responsáveis por efetivamente processar o arquivo PDF. Os resultados dessa etapa são a marcação de entidades e variáveis pré-determinadas, que serão armazenadas em um repositório com essa finalidade.

   O usuário então pode conferir os resultados das marcações de entidades e variáveis na interface de Pre-visualização das marcações / Edição, e proceder com alguma alteração, caso necessário.

   Uma vez que as marcações estão satisfatórias, o sistema faz a Aplicação de Modelo de Significado Processual, fazendo uso do modelo ML previamente treinado, com os objetivos:

   i. identificação das alegações;

ii. exame de admissibilidade;

iii. cálculo da probabilidade de concessão de medidas cautelares.

O resultado dessa análise é persistido em um repositório e pode ser visualizado pelo usuário do TCU por meio da interface de visualização Resultado de Análise, na forma de um relatório, com possibilidade de exportação.

Ao fim desse processo, o sistema automaticamente já inicia a atividade de atualização dos dados que alimentam o Painel de Jurimetria.

Dependendo das regras do painel de jurimetria, um conjunto de decisões de fluxo automatizadas podem detectar a necessidade de emitir comunicação aos interessados, que resulta na redação automática da peça de comunicação correspondente do Tribunal.

3. Módulo Painel de Jurimetria – tem por objetivo fornecer uma visualização com informações para auxiliar o trabalho de instrução, permitindo ao usuário realizar seleção de processos pelo número ou de parâmetros para a realização de queries na base de conhecimento; ademais, pode-se ainda realizar pesquisas por vários parâmetros, a serem definidos em conjunto com o TCU;

4. Módulo de Redação e Geração de Peças: De forma automatizada, ou se requerido pelo usuário do TCU, o sistema será capaz de redigir o texto de Comunicação aos interessados para um processo previamente analisado pela ferramenta. Essa etapa faz uso de modelos de linguagem previamente treinado e seus resultados são armazenados em repositório, para visualização e consulta. Adicionalmente, o sistema pode exibir o resultado da redação para o usuário, realizar o encaminhamento da informação por meio de e-mail ou interfacear com sistemas já existentes do TCU para dar o encaminhamento necessário.

Para o desenvolvimento da solução proposta, as instituições executoras proverão uma equipe multidisciplinar contratada formalmente com os recursos financeiros relacionados no Item 14. A execução do projeto se dará ao longo de 36 meses em 26 etapas.

6.2 Cronograma

Anexo II – Planilhas

6.3 Descrição das etapas

<u>Etapa 01 –</u> Workshop de Kickoff do projeto

Descrição: Nesta etapa será realizado workshop de kickoff, por videoconferência ou na sede do TCU, com participação da equipe técnica das executoras. O kickoff é uma cerimônia prevista para projetos, além de ser uma boa prática de gerenciamento de projetos.

Produtos esperados: Registro da reunião de Kickoff do projeto apresentando os objetivos, participantes e suas funções, cronograma e riscos. Apresentação de slides a ser anexada aos artefatos do projeto.

Responsável: UnB, Unicamp, Eldorado e INutech

Mês do Projeto: 01

<u>Etapa 02 –</u> Entendimento do Negócio

Descrição: Nesta fase serão realizadas reuniões de trabalho que objetivam o entendimento detalhado e abrangente dos problemas que serão resolvidos, especificando os problemas em termos das respectivas perguntas de negócios e de dados. Será utilizada a metodologia *CRISP-DM*, descrito no item 4.1 desta proposta técnica comercial.

Produtos esperados: Relatório com as especificações das perguntas de negócios e de dados do problema a ser resolvido.

Responsável: UnB, Unicamp. iNuTech, Eldorado

Mês do Projeto: De 1 a 3

<u>Etapa 03 –</u> Saneamento dos dados

Descrição: Nessa etapa será realizada a integração dos dados, correspondentes a extração, transformação e carregamento. Nesse processo, os dados são retirados do sistema fonte do fornecedor e convertidos para um formato que possa ser utilizado nas próximas etapas. Nessa etapa, também serão realizados: a anonimização, o pré-processamento de texto, que inclui atomização, correções ortográficas, redução lexical, remoção de stopwords, normalização, dentre outros.

Produtos esperados: Relatório Técnico e Base semiestruturada

Responsável: UnB

Mês do Projeto: De 1 a 6

Etapa 04 – Planejamento da etapa de rotulação

Descrição: Organização, definição de protocolos de rotulagem para cada tipo de documento e treinamento da equipe para as atividades de rotulação das peças processuais. Definição do tipo de base de dados, seleção, implantação e customização de ferramenta de rotulação. As bases podem ser: (1) ouro que necessita de revisão por pares ou; (2) prata ou bronze que fazem uso de técnicas como Aprendizado Ativo, Supervisão Fraca ou suporte externo em conjunto com os anotadores. Uma forma de avaliar a qualidade da base é utilizar medidas estatísticas como os coeficientes Kappa e Krippendorff's Alpha para avaliar a concordância entre as anotações. A ferramenta de rotulação também será definida nessa etapa. Alternativas como LabelBox, LabelStudio, Inception ou TeamTat.

Produtos esperados: Metodologia e protocolos de rotulação, ferramenta para as rotulações escolhida, implantada e customizada.

Responsável: UnB, INutech

Mês do Projeto: De 1 a 4

Etapa 05– Rotulação de peças processuais

Descrição: Essa etapa fornecerá as bases rotuladas que serão utilizadas em todas as etapas do projeto. No caso da escolha da base ouro, a equipe será organizada em pares e um revisor ficará responsável por validar as rotulações. No caso da geração da base prata ou bronze, técnicas como Aprendizado Ativo e Supervisão Fraca poderão ser utilizadas em conjunto para selecionar as peças com diferentes vieses e que geram maior ganho de informação para compor a base rotulada. Faz parte da etapa de rotulação (1) a identificação de segmentos em cada processo e (2) a rotulação de entidades nomeadas. Para esse processo, ferramentas de rotulação mencionadas na Etapa 3 serão utilizadas. Os processos a serem rotulados serão divididos em quatro lotes, para que as próximas etapas possam ser iniciadas, com uma base rotulada parcial. O Anexo 1 da presente proposta apresenta o planejamento detalhado dessa etapa.

Produtos esperados: Base rotulada

Responsável: UnB, INutech

Mês do Projeto: De 3 a 15

Etapa 06 – Pré-processamento

Descrição: A etapa de pré-processamento tem como objetivo preparar e ajustar os dados para a indução dos modelos. Problemas como desbalanceamento, valores ausentes, ruídos e

duplicidades de dados podem ser tratados nessa etapa por meio de técnicas de Mineração de Texto e Aprendizado de Máquina.

Produtos esperados: Base rotulada filtrada

Responsável: UnB

Mês do Projeto: De 7 a 19

Etapa 07 – Indução de modelos para classificação

Descrição: A indução de modelos de classificação pode ser feita inicialmente por Aprendizado Ativo e Supervisão Fraca. Nessa etapa podemos estabelecer um baseline de desempenho. Em uma etapa posterior, novos modelos estado-da-arte como modelos baseados em Transformers podem ser utilizados.

Produtos esperados: Benchmark de desempenho, Modelos de linguagens, Modelos de classificação.

Responsável: UnB

Mês do Projeto: De 10 a 34

Etapa 08 – Estratégia para avaliação dos resultados

Descrição: A avaliação de modelos de Aprendizado de Máquina deve ser feita considerando as características como o desbalanceamento e o número de classes. Portanto, medidas que consideram essas características como AUC, GMean e Kappa podem ser índices interessantes para avaliar o desempenho dos modelos propostos. Além disso, é importante fornecer garantias estatísticas sobre os resultados por meio de técnicas de validação cruzada e testes estatísticos de significância.

Produtos esperados: Relatório Técnico, Benchmark de desempenho

Responsável: UnB, INutech

Mês do Projeto: De 10 a 34

Etapa 09 – Explicabilidade de modelos

Descrição: A explicabilidade de modelos de Aprendizado de Máquina tem por objetivo detectar vieses, preconceito, falhas e vulnerabilidades. Além disso, ela pode mensurar a importância dos atributos, entender o comportamento dos modelos e das predições. Portanto, tem por objetivo atender os domínios que necessitam de explicações das decisões. Técnicas como Local Interpretable Model-Agnostic Explanation (LIME), SHapley Additive exPlanations (SHAP), dentre

outras podem ser utilizadas para identificar essas vulnerabilidades nos modelos treinados e permitir que maior robustez aos modelos gerados.

Produtos esperados: Relatório Técnico

Responsável: UnB

Mês do Projeto: De 10 a 34

Etapa 10 – Recuperação de Informação e Análise Semântica

Descrição: Técnicas de Mineração de Texto podem ser utilizadas para identificar documentos de textos semelhantes. A identificação de processos semelhantes é considerada uma forma de recuperação de informação, o que justifica a avaliação de técnicas tradicionais, como aquelas baseadas em decomposição matricial (LSI – Latent Semantic Indexing, NMF – Non-negative Matrix Factorization) e modelos probabilísticos de tópicos (como o LDA – Latent Dirichlet Allocation). Concomitantemente a outras etapas, pode-se gerar a imersão de palavras (Word Embedding) e imersão de palavras dado contexto (Context Word Embedding) para auxiliar na tarefa de análise semântica de texto. Pretende-se aproveitar Modelos de Linguagens e modelos supervisionados para melhorar a recuperação de informação.

Produtos esperados: Sistema de recuperação de informação. Modelos de Análise Semântica treinados. Embeddings de contexto, Modelos Sequenciais de Texto

Responsável: UnB

Mês do Projeto: De 4a 28

Etapa 11 – Análise de Similaridades entre sentenças

Descrição: Similaridade entre sentenças ou similaridade semântica entre textos curtos mede a similaridade entre dois segmentos de textos. Sabe-se que as técnicas modernas de modelos de linguagens são capazes de trazer bons resultados no grau de significado em textos curso, diferente do caso de textos longos. Com isso, pode-se aproveitar os recursos produzidos em outras etapas para tratar o problema de similaridade entre sentenças. Os artefatos produzidos nesta etapa podem auxiliar na geração de textos, classificação e recuperação de informação.

Produtos esperados: Relatório Técnico com estatísticas e avaliações de similaridades

Responsável: UnB, Unicamp

Mês do Projeto: De 6 a 30

Etapa 12– Avaliação de resultados das Técnicas de Similaridade

Descrição: Há várias métricas para avaliação de sistemas de Recuperação de informação e análise de similaridade. As mais básicas e frequentes são precisão e revocação, combinadas de diferentes maneiras. Aqui, pretende-se compor um conjunto de dados controlado, onde se saiba os resultados esperados. Cada documento de texto será submetido a base para a recuperação dos mais similares, e os resultados serão comparados com o ranque estabelecido previamente. Pretende-se calcular a média das precisões para cada nível de revocação, para ter uma avaliação do ponto de vista do usuário.

Produtos esperados: Métricas indicando a acurácia dos modelos

Responsável: UnB, Unicamp, INuTech

Mês do Projeto: De 7 a 36

Etapa 13 – Metodologia para refinamento de modelos de linguagem

Descrição: Desenvolvimento de uma metodologia explorando métodos do estado da arte para o refinamento de modelos de linguagem. Desenvolveremos uma sucessão de procedimentos de treinamentos de modelos em uma abordagem de aprendizagem currículo. Nessa abordagem, a cada passo os modelos serão treinados com dados mais específicos visando obter conhecimentos mais especializados no modelo.

Produtos esperados: Metodologia

Responsável: Unicamp

Mês do Projeto: De 2 a 18

Etapa 14 – Sumarização automática de textos

Descrição: Desenvolvimento de técnicas para a sumarização de textos a partir de um repertório de documentos textuais. Investigaremos modelos responsáveis em gerar textos sintéticos como resumo de um documento fonte de maneira automatizada como tarefa alvo. Exploraremos técnicas de manipulação de dados e refinamento de granularidade de modelos para esse fim.

Produtos esperados: Técnicas e Implementação de provas de conceito

Responsável: Unicamp e Eldorado

Mês do Projeto: De 06 a 35

Etapa 15 – Geração automática de textos

Descrição: Estudo de métodos para a geração de textos sintéticos com base no treinamento de modelos a partir de modelos pré-treinados existentes. Para esta tarefa, exploraremos uma abordagem arquitetural de aprendizado utilizando codificador e decodificador com modelo. Discussão e disseminação de resultados.

Produtos esperados: Técnicas e Implementação de provas de conceito.

Responsável: Unicamp e Eldorado

Mês do Projeto: De 10 a 36

Etapa 16 – Avaliação experimental da técnica

Descrição: Serão realizadas simulações e experimentos controlados executados considerando conjunto de dados gerados. Iremos validar as soluções propostas a serem empregadas, comparando-as com as soluções existentes na literatura por meio de metodologias e métricas quantitativas e qualitativas. Métricas como acurácia, precisão e medida-f1 serão exploradas nesta etapa.

Produtos esperados: Relatórios técnicos

Responsável: Unicamp e Eldorado

Mês do Projeto: De 08 a 36

Etapa 17 – Disseminação de resultados

Descrição: Elaborar artigos para submissão em eventos científicos e periódicos especializados divulgando os resultados encontrados. Resultados derivados do projeto serão submetidos na forma de artigos científicos a periódicos e conferências seletivas qualificadas. Isso será relevante para a divulgação da pesquisa e na formação dos alunos. A versão final da solução será integralmente documentada em um relatório técnico.

Produtos esperados: Artigos científicos e relatórios técnicos

Responsável: Unicamp, UnB e Eldorado

Mês do Projeto: De 14 a 17; 24 a 27; 32 a 36

Etapa 18 – Definição de Requisitos para implementação da Plataforma Computacional, incluindo o desenho das interfaces para os usuários.

Descrição: Nesta etapa serão levantados os requisitos funcionais para a implementação do software do sistema por meio de entrevistas com a equipe técnica do TCU, e documentados

pela equipe de desenvolvimento de software. Será também desenhada a interface do usuário em sistema de mock-up. Para esta fase os profissionais de desenho de interface assim como analistas de software da equipe de desenvolvimento serão envolvidos.

Equipe técnica envolvida: PO e time de UX design.

Produtos esperados: Conjunto de requisitos documentados em ferramenta apropriada, como por exemplo estórias de usuário no Jira ou similar, assim como documentação gráfica da interface do usuário.

Responsável: Unicamp, Eldorado, UnB e INutech

Mês do Projeto: De 11; 17; 30.

Etapa 19 – Implementação inicial dos protótipos funcionais

Descrição: Nesta etapa, tendo como base a documentação de requisitos criada na etapa anterior, serão desenvolvidos protótipos funcionais do software utilizando técnicas de codificação. Esta etapa poderá ser iniciada antes da finalização da etapa anterior de levantamento de requisitos.

Equipe técnica envolvida: Product Owner, Scrum Master, time de desenvolvimento, time de UX design.

Produtos esperados: Código fonte dos protótipos funcionais.

Responsável: Eldorado, Unicamp, UnB e INutech

Mês do Projeto: De 12 a 13; 18.

Etapa 20 – Disponibilização para feedback e alterações iniciais

Descrição: Os protótipos implementados serão disponibilizados para o TCU para testes e feedback. O Product Owner será responsável por coletar os feedbacks do TCU e informar as equipes de desenvolvimento e teste das necessidades de alterações, utilizando ferramentas computacionais apropriadas.

Equipe técnica envolvida: Product Owner.

Produtos esperados: Relatórios e arquivos binários para execução das funcionalidades previstas.

Responsável: Unicamp, Eldorado, UnB

Mês do Projeto: De 14; 19 e 31

Etapa 21 – Implementação dos softwares da Plataforma Computacional

Descrição: Tendo como base os requisitos levantados para o sistema a ser implementado, o time de desenvolvimento de software irá implementar o código de acordo com o desenho de interface e dos requisitos funcionais. O software será implementado por uma equipe de desenvolvedores de aplicativos web front-end e back-end.

As atividades envolvidas nesta fase incluirão o projeto e especificação de banco de dados para a aquisição e manutenção de dados, implementação de software back-end para acesso ao banco de dados e a implementação do software de front-end para interface dos usuários.

O processo de desenvolvimento do software irá seguir padrões de indústria utilizando metodologias ágeis, com o planejamento de Sprints para um conjunto de requisitos a serem entregues periodicamente.

Equipe técnica envolvida: Product Owner, Scrum Master, time de desenvolvimento, time de UX design.

Produtos esperados: Código fonte, relatórios e arquivos binários para execução das funcionalidades previstas

Responsável: Unicamp, Eldorado, UnB

Mês do Projeto: De 15 a 36

Etapa 22 – Testes e Verificação dos Softwares da Plataforma Computacional

Descrição: Criação de test cases para verificação do sistema implementado e execução dos testes nas versões disponibilizadas pela equipe de desenvolvimento.

Equipe técnica envolvida: Analistas de teste, analistas de desenvolvimento, time de UX design

Produtos esperados: Relatórios de execução de testes, documentação periódica listando as alterações a serem realizadas pela equipe de desenvolvimento do software do sistema.

Responsável: Unicamp, UnB e Eldorado

Mês do Projeto: De 16 a 36

Etapa 23 – Rollout das soluções / Treinamento das equipes envolvidas

Descrição: Esta etapa consiste em efetuar a transferência das soluções desenvolvidas no projeto para o TCU. Isso envolve comunicação entre equipes técnicas, compartilhamento de configuração de ambientes e procedimentos de *deployment* do software. Adicionalmente, esta etapa envolve o treinamento de pessoas no TCU para a utilização do software gerado.

Produtos esperados: Manuais de utilização das soluções

Responsável: Unicamp, UnB, Eldorado e INutech

Mês do Projeto: De 30 a 36

Etapa 24 – Validação dos resultados obtidos

Descrição: Esta etapa visa conduzir uma análise de resultados obtidos em diferentes momentos do projeto a partir de resultados de avaliação experimental de técnicas produzidas no projeto. A análise também envolve inspeções de uso preliminar do produto por partes interessadas.

Produtos esperados: Relatório técnico de validação da ferramenta; Entrega da versão final da plataforma computacional

Responsável: Unicamp, Eldorado, UnB e INuTech

Mês do Projeto: 15; 20;24; 29 a 35

Etapa 25 – Workshop de encerramento

Descrição: Realização do workshop de encerramento, sendo este uma cerimônia prevista para projetos, além de ser uma boa prática de gerenciamento de projetos. Nessa etapa será conduzido pela equipe técnica das entidades executoras o workshop de encerramento, por vídeo conferência ou presencial nas instalações do TCU.

Produtos esperados: Workshop

Responsável: Unicamp, Eldorado, UnB e INutech

Mês do Projeto: 36

Etapa 26 – Elaboração do relatório final

Descrição Consolidação dos relatórios das etapas objetivando a elaboração do relatório final do projeto.

Produtos esperados: Relatório final

Responsável: Unicamp, Eldorado, UnB e iNuTech

Mês do Projeto: 34 a 36

7. PRODUTO PRINCIPAL

O produto principal do desenvolvimento é uma Plataforma Computacional de aprendizado de máquina e processamento de linguagem natural para textos jurídicos

O produto principal embarca:

a) Criação de um conjunto de protótipos funcionais para integração com sistemas internos do TCU, a serem definidos de acordo com os requisitos de sistemas internos (web/mobile/dashboards).
b) Produtização dos protótipos funcionais criados na fase anterior do projeto
c) Criação de ferramentas para a ingestão de dados, geração de métricas de desempenho e ajustes de parâmetros dos protótipos funcionais.
d) Modelos de Aprendizado de Máquina para classificação e cálculo de probabilidade das peças processuais.
e) Técnicas de Recuperação de Informação e Análise Semântica de documentos jurídicos.
f) Metodologia para refinamento de modelos de linguagem visando geração de texto no contexto jurídico.
g) Concepção, desenvolvimento e avaliação de técnicas para sumarização e geração de textos de maneira automática no domínio jurídico.
h) Base de dados pré-processada, semi-estruturada e rotulada para diversas tarefas do problema (classificação, segmentação de texto e extração de entidades nomeadas).

7.1 Produtos secundários

Os seguintes produtos secundários serão disponibilizados:

- Conjunto de dados devidamente anotados para apoio ao treinamento de modelos de aprendizagem de máquina.
- Análise experimental comparativa de algoritmos do estado da arte para indução sobre peças processuais.
- Análise comparativa de técnicas de Análise de Similaridades entre sentenças para recuperação de informação
- Estudo sobre a interpretabilidade de modelos de aprendizagem de máquina visando garantir qualidade e racional de decisões da máquina.
- Protótipos funcionais que implementam as técnicas estudadas e desenvolvidas.

- Estudo piloto da solução por meio de avaliação experimental dos protótipos desenvolvidos.
- Disseminação dos resultados acadêmicos alcançados.

De modo complementar deverão ser entregues também os seguintes produtos:

UnB

- 2 artigos científicos em periódico internacional/nacional;
- 3 artigos científicos em congresso internacional/nacional;
- 3 dissertações de mestrado em temas correlatos ao projeto;
- 5 trabalhos de conclusão de curso em temas correlatos ao projeto;

Unicamp

- 2 artigos científicos em periódico internacional/nacional;
- 3 artigos científicos em congressos internacional/nacional;
- 2 dissertações de mestrado em temas correlatos ao projeto;
- 1 trabalhos de conclusão de curso em temas correlatos ao projeto;

iNuTech

- 1 artigo científico em periódico internacional/nacional;
- 1 artigo científico em congressos internacional/nacional;
- 1 dissertação de mestrado em temas correlatos ao projeto;
- 1 Workshop sobre o estado da arte em NLP, com a utilização de modelos explicáveis.

Eldorado

- 1 artigo científico publicado em periódicos nacionais ou internacionais.
- Documentações da Plataforma Computacional: Arquitetura, Especificação Técnica dentre outros;
- Estruturação de Oficinas de aprendizado no contexto do Instituto para expansão do conhecimento.
- Ciclo de Palestras públicas nos temas de desenvolvimento do projeto e/ou meetups de desenvolvimento.
- Workshops de design para treinamentos a cada seis meses para todos os envolvidos no projeto.
- Avaliação de oportunidades de registro de software computacional junto ao INPI.

8. PREMISSAS

- O Tribunal de Contas da União será responsável pela contratação de serviços e aquisição de ferramentas, bibliotecas e/ou licenças de SW e demais infraestruturas tecnológicas necessárias à implantação, utilização da solução contratada.
- O Tribunal de Contas da União deverá fornecer e ou permitir os detalhamentos de fluxos e processos pela Contratada, a serem implementados, assim como validar a implementação de acordo com as regras de negócios e regulamentações. Deverá ainda, disponibilizar dados estruturados, essencialmente, metadados ligados ao protocolo de cada processo.
- O Tribunal de Contas da União deverá disponibilizar todos os documentos dos 14 mil processos de representações de denúncias, a partir e 2010, e fornecer os dados referentes aos 14 mil processos que constem no Sistema e-TCU ou franquear o acesso ao mesmo.
- Após a fase de saneamento de dados a ETEC será reavaliada quanto à sua viabilidade em termos financeiros e técnicos, destacando as métricas de desempenho acordadas entre as partes do contrato, podendo ser encerrada.
- O Tribunal de Contas da União afirmou em documento, enviado por e-mail no dia 17/02/22, às 18h50, referente às dúvidas do Edital de Chamamento Público para a Encomenda Tecnológica de Instrução Assisitida por Inteligência Artificial (anexo a esta proposta), que os documentos não estruturados são os PDFs/DOCXs das peças processuais citadas no Termo de Referência. As petições iniciais são digitalizadas, e têm problemas de reconhecimento de caracteres. As demais peças são nativamente criadas em meio digital.
- O Tribunal de Contas da União afirmou em documento, enviado por e-mail no dia 17/02/22, às 18h50, referente às dúvidas do Edital de Chamamento Público para a Encomenda Tecnológica de Instrução Assisitida por Inteligência Artificial (anexo a esta proposta), que não estão claros os critérios para aceite dos produtos, em termos de métricas de acurácia, precisão, recall e etc. De fato, não está clara nem a possibilidade de existência do produto. Para o Projeto de P&D, pede-se que a Contratada apresente uma proposta de quais métricas utilizar para cada marco do projeto. Porém, ao final da fase de saneamento de cada ciclo temático, as partes negociarão as métricas, as faixas de bônus e seus valores. Esta negociação poderá ser revista ao longo do

desenvolvimento do produto, sempre que ambas as partes concordarem que houve impacto devido ao risco tecnológico

- O Tribunal de Contas da União deverá indicar e alocar potenciais usuários da solução para a fase inicial de pesquisas e validação de testes de usabilidade.
- O Tribunal de Contas da União deverá fornecer uma definição básica do grupo de usuários que a solução pretende impactar, definindo nichos que representem a escala quantitativa.
- O Tribunal de Contas da União deverá definir e indicar os níveis de acessibilidade a serem cobertos pelo projeto.
- O sistema será desenvolvido em Português-BR.
- A solução será desenvolvida considerando uma interface web responsiva.
- O Tribunal de Contas da União deverá viabilizar os testes sistêmicos da solução em condições compatíveis com as que o sistema será utilizado e validado.
- O Tribunal de Contas da União deverá ser responsável por garantir a segurança da informação contra possíveis ataques de intrusão, captura de dados e paralização dos serviços.
- A data de início do projeto será alinhada de comum acordo entre as partes.
- O Tribunal de Contas da União afirmou em documento, enviado por e-mail no dia 17/02/22, às 18h50, referente às dúvidas do Edital de Chamamento Público para a Encomenda Tecnológica de Instrução Assistida por Inteligência Artificial (anexo a esta proposta), que os pagamentos serão mensais mediante apresentação de fatura dos serviços prestados, no valor do custo fixo mensal acertado entre as partes para o respectivo marco do projeto, na fase de negociação, com base nas planilhas de custo fornecidas pela Contratada.

9. RISCOS

Todo projeto de pesquisa e de desenvolvimento de forma inerente envolve riscos para a obtenção dos seus resultados objetivados. A seguir são descritos alguns riscos vislumbrados pelas executoras, probabilidade de ocorrência, impactos no projeto, plano de contingência para a redução e/ou eliminação dos mesmos caso ocorram e os responsáveis pelo plano de contingência.

Risco	Probabilidade de Ocorrência	Impacto	Plano de Contingência	Responsável pelo plano de contingência
Variação nos preços dos equipamentos, e materiais, assim como atraso na entrega de equipamentos	Baixa	Médio	Concluir a aquisição nos primeiros meses do projeto	Comitê Estratégico e GP TCU
Atrasos por parte do TCU em disponibilizar o acesso aos dados e outros sistemas legados	Média	Alto	Realizar acompanhamento periódico com o TCU e dar alta prioridade para coleta de dados. Alocar recurso da TI de forma a garantir acesso aos dados nos prazos previstos em cronograma	Comitê Estratégico e GP TCU
Manutenção da equipe executora durante o período do projeto	Médio	Alto	Trabalhar com o banco de talentos com perfis similares ao exigido para a entrega do produto	Comitê Estratégico
Atraso no pagamento das entregas do projeto	Médio	Alto	Empréstimo para possíveis atrasos no pagamento	Comitê Estratégico
Custo de infraestrutura de TI exceder o planejado	Médio/Alto	Alto	Rever definição dos objetivos do projeto e (talvez) orçamento. Buscar alternativas nacionais, regionais ou em outros projetos.	Comitê Estratégico
Segurança da Informação	Alto	Alto	TCU deverá ser responsável por garantir a segurança da informação contra possíveis ataques de intrusão, captura de dados e paralização dos serviços	GP TCU

10. ORIGINALIDADE

10.1 Propriedade intelectual

Estão previstos a produção e os seguintes depósitos de Propriedade Intelectual junto ao INPI – Instituto Nacional da Propriedade Industrial.

Tema/Linha de Pesquisa	Patente ou Registro de Software	Período previsto para depósito
Software com Algoritmos de IA para a Classificação de textos jurídicos	Programa de Computador	Mês 35
Software com Algoritmos de IA para geração automática de textos jurídicos	Programa de Computador	Mês 35

10.2 Fatores de Originalidade

O projeto na fase de inovação apresenta os seguintes quesitos de originalidade:

- Ineditismo: as “Buscas de Anterioridades do Projeto” realizadas no item anterior não encontrou qualquer projeto, artigo, dissertação ou tese, processo de patente e produto comercializado com similaridade a plataforma a ser obtida, sendo por isso constatada a sua originalidade no campo técnico de seu conhecimento e de sua inovação.

- Propriedade Intelectual: estão previstos a produção e depósito de dois: (i) Software com Algoritmos de IA para a Classificação de textos jurídicos e (ii) Software com Algoritmos de IA para geração automática de textos jurídicos.

- Produção acadêmica: estão previstas 5 (cinco) dissertações de mestrado, 6 (seis) trabalhos de fim de curso em temas relacionados ao projeto, e publicação de 11 (onze) artigos classificados na lista Qualis Periódicos como A1, A2 ou B1.

11. APLICABILIDADE

A Plataforma Computacional de aprendizado de máquina a ser desenvolvida permitirá o processamento de linguagem natural para textos jurídicos, bem como o fornecimento de estimativas diversas com o auxílio de uma aplicação de jurimetria preditiva.

11.1 Motivações para a construção da solução proposta

O Sistema Judicial Brasileiro, operando como o instrumento máximo de poder para a garantia dos direitos previstos na Constituição, promove uma supervalorização dos mecanismos oficiais

para a resolução de conflitos e aplicação de justiça, assumindo no inconsciente coletivo o papel de um superpoder a quem se deve delegar a tarefa de resolução de todas as disputas existentes entre os indivíduos, empresas e órgãos governamentais. Nesse contexto, diante da incapacidade de atendimento de tais anseios da população, das instituições civis e do próprio governo, surge o cenário atual de hiperjudicialização e crise do Poder Judiciário.

A hiperjudicialização, consequência de uma cultura de litigância, caracterizada pela prática do cidadão brasileiro de promover a delegação da resolução dos conflitos ao Sistema Judiciário, abarrota os Tribunais com milhões de novos processos anuais, acarretando, entre outros fatores, a demora de anos para a tramitação dos processos judiciais, e uma distorção funcional do Sistema Judiciário.

A cultura da litigância reflete uma ideia fortemente arraigada na sociedade brasileira, de que qualquer conflito para ser definitivamente resolvido deve ser judicializado e concluído sob a forma de uma sentença adjudicada, dotada de força coercitiva. As raízes da elevada litigiosidade da sociedade brasileira são de origens e motivações distintas, possuindo como elementos principais a história social e política dos últimos séculos, o desenvolvimento atual sócio-econômico, educacional e cultural, além dos níveis de credibilidade e de eficiência que os Tribunais conquistaram no imaginário popular.

A cultura da litigância em conjunto com os rituais processualísticos de cunho protelatório estabelecidos no Sistema Judicial Brasileiro, promovem a crise do Poder Judiciário, tornando-o ineficiente e incapaz de atender de modo satisfatório as demandas que recebe. A frustração das expectativas de resolução satisfatória das demandas encaminhadas ao Sistema Judiciário, por outro lado, realimenta o espírito litigante das partes em conflito, que terminam por utilizar todos os recursos legais disponíveis e o envio das ações judiciais às instâncias superiores do Sistema Judiciário, culminando em um círculo vicioso.

Adicionalmente, a formação acadêmica histórica dos cursos de Direito, que formam juízes, advogados, analistas legislativos ou judiciários, não contribui para que esses operadores do Direito valorizem e priorizem soluções consensuais para os conflitos, resultando ainda, quando combinada com a cultura de litigância, nos mais baixos índices globais de acordos por mediação ou conciliação.

Tomando como base os relatórios disponibilizados anualmente pelo CNJ (Conselho Nacional de Justiça), podemos comprovar, assim como dimensionar as questões e problemas anteriormente tratados:

- A quantidade de ações judiciais em tramitação no Sistema Judicial Brasileiro, é desproporcional à população, quando comparada aos índices globais. Considerando apenas as ações judiciais efetivamente ajuizadas pela primeira vez em 2020 – 17,6 milhões ações originárias – temos um índice de cerca de 8.263 ações novas para cada grupo de 100 mil habitantes (projeção de 213,3 milhões de habitantes em agosto/2021);
- O Poder Judiciário finalizou o ano de 2020 com 62,4 milhões de ações judiciais em tramitação, aguardando alguma solução definitiva;
- Durante o ano de 2020, em todo o Poder Judiciário, ingressaram 25,8 milhões de processos e foram baixados 27,9 milhões;
- Se forem consideradas apenas as ações judiciais efetivamente ajuizadas pela primeira vez em 2020, sem computar os casos em grau de recurso e as execuções judiciais (que decorrem do término da fase de conhecimento ou do resultado do recurso), tem-se que ingressaram 17,6 milhões ações originárias em 2020, -12,5% do que no ano anterior;
- O Relatório Anual do ano de 2011 do Conselho Nacional de Justiça apontou que a taxa de congestionamento na fase de execução de primeiro grau da Justiça Estadual chega a 89,8%, ou seja, de cada 100 processos sentenciados apenas 10 foram garantidos ou quitados (CNJ, 2012, p.43), acusando a falta de efetividade na execução das decisões proferidas;
- A maior parcela dos contendedores do Sistema Judiciário são agentes públicos. O Conselho Nacional de Justiça, em março de 2011, divulgou a lista dos 100 maiores litigantes. Dos dez maiores litigantes, seis integram a Administração Pública, quais sejam: INSS - INSTITUTO NACIONAL DO SEGURO SOCIAL (22,33%), CEF - CAIXA ECONÔMICA FEDERAL (8,50%), FAZENDA NACIONAL (7,45%), UNIÃO (6,97%), BANCO DO BRASIL S/A. (4,24%), ESTADO DO RIO GRANDE DO SUL (4,24%). Em dados mais concretos, de todos os processos em curso no Judiciário nacional, a Administração Pública é parte em 53,73%, seja como autora ou ré.
- No setor privado, o setor Bancário e de Telefonia disparam com a quantidade de ações ajuizadas ou como réus: em 7º, 8º, 9º e 10º lugar da pesquisa estão, respectivamente, o BANCO BRADESCO S/A (3,84%), BANCO ITAÚ S/A (3,43%), BRASIL TELECOM CELULAR S/A (3,28%) e BANCO FINASA S/A (2,19%).
- Juntos, bancos e telefonia, mais o setor público (Federal, Estadual e Municipal) representam 95% do total de processos dos 100 maiores litigantes nacionais.

11.2 Âmbito de aplicação do produto principal do projeto

Reverter a cultura de litigância, com o objetivo de reduzir o número de novos processos judiciais e do número de recursos encaminhados anualmente ao Sistema Judiciário, assim como aumentar os percentuais de acordos por mediação ou conciliação, certamente será um processo educacional longo e custoso, que demandará décadas, na melhor das hipóteses. Desse modo, resta como alternativa melhorar a produtividade e a qualidade da tramitação de processos judiciais, sendo esse o objetivo principal da solução descrita na presente proposta.

O uso de assistentes virtuais especializados, capazes de eliminar a maior parte do tempo gastos em atividades repetitivas, apontando ainda as prioridades em que devem ser empregados os esforços das equipes profissionais, proporciona simultaneamente o aumento de produtividade e a melhoria da qualidade dos produtos e serviços nas mais diversas áreas.

Na área do Direito, a consulta de leis, decretos e jurisprudência, a análise de doutrinas mais adequadas para determinada ação, a elaboração de pareceres jurídicos e a construção de peças, assim como o cálculo de estimativas de êxito, de valores e de prazos de tramitação de processos judiciais, já são atividades realizadas com bem mais rapidez, eficiência e acurácia pelas aplicações de Inteligência Artificial.

A solução apresentada na presente proposta, concebida como uma Plataforma Computacional, composta por um conjunto de aplicações supervisionadas de Inteligência Artificial, tem por objetivo auxiliar os profissionais do TCU em atividades consideradas críticas e de elevada importância, como aquelas descritas acima.

A plataforma a ser desenvolvida, utilizando processamento de linguagem natural e sofisticados mecanismos de reconhecimento de padrões, combinando abordagens híbridas de aprendizado de máquina e análise de redes semânticas com métodos estatísticos, será capaz de realizar a análise, classificação e priorização de processos de Controle Externo – Representações e Denúncias – oferecerá um conjunto de ferramentas que poupam horas de trabalho à advogados, analistas, auditores e demais profissionais do TCU, na leitura e análise de peças processuais, na pesquisa de teses jurídicas, na seleção de processos similares para estudo dos resultados dos julgamentos anteriores, permitindo uma orientação ágil, objetiva e segura da melhor estratégia de tese jurídica a ser adotada para cada caso.

A plataforma também deverá disponibilizar consultas típicas de jurimetria, fornecendo estimativas robustas de êxito, de valores e de prazos de tramitação, para o temas de Aquisições Públicas (licitações, compras diretas ou contratos decorrentes desses instrumentos), fundadas

nas especificidades de cada caso a ser analisado, apontando os fatores relevantes e determinantes dos resultados do julgamento de casos similares, complementando as facilidades para a priorização de processos e para a construção de peças jurídicas.

13. RELEVÂNCIA

13.1 Contribuições e impactos tecnológicos e científicos

13.1.1 Apoio à infraestrutura laboratorial

Para a execução do projeto estão previstas as seguintes aquisições de materiais permanentes e de softwares para apoio a infraestrutura laboratorial:

- Para apoio as atividades dos pesquisadores da UnB:
  - Servidor de aplicação: Para desenvolvimento e teste de algoritmos, armazenamento de dados e processamento básico de dados.
  - Modem 3G/4G industrial: Para fazer a comunicação entre os computadores industriais e servidor por meio da Internet (3G/4G).
  - Estação de trabalho tipo 1 (incluindo CPU e monitor 27"): Para o desenvolvimento e teste de algoritmos e simulações computacionais.
  - Estação de trabalho tipo 2 (incluindo CPU e monitor 24"): Para o desenvolvimento e teste de algoritmos e simulações computacionais.
  - Notebook com IP53 (Ingress Protection 53) destinado para atividades em campo: Para atividades de campo em condições ambientais outdoor, faz-se necessário um notebook robusto com IP53 e tela apropriada para ambiente com muita luz.
- Para apoio as atividades dos pesquisadores da Unicamp:-
  - Estação de trabalho de alto desempenho (incluindo CPU e monitor): Para o desenvolvimento e teste de algoritmos de IA e simulações computacionais.
  - Notebooks de alto desempenho e SSD para atividades de pesquisa e desenvolvimento da equipe
- Para apoio as atividades dos pesquisadores da Eldorado:-
  - Ambiente Cloud
  - 5 kits de desenvolvimento de alto desempenho (notebook com GPU embarcada e outros acessórios de informática)
  - Outros acessórios de informática como nobreaks para apoio as atividades dos pesquisadores do Instituto Eldorado

13.1.2 Capacitação Profissional

Estão previstas as seguintes capacitações profissionais ao longo de realização do projeto.

Tipo	Membro da Equipe	Instituição de ensino	Tema	Início (mês/ano)	Término (mês/ano)
Treinamento Operacional	TCU	Eldorado	Plataforma Computacional (documentação, telas, fluxos, regras, entre outros)	Etapa 21	Etapa 21
Workshop	TCU	INuTech	Estado da arte em NLP	-	-

13.1.3 Produção Técnica-Científica

As seguintes publicações técnico-científicas estão previstas para divulgação de conhecimento ao longo do projeto, cujos os temas, eventos e periódicos serão selecionados adequadamente conforme o tema das contribuições científicas.

item	Executora	Tema/Linha de Pesquisa	Título do Evento Periódico	É Qualis A1, A2 ou B1? (S/N)	Período previsto para envio
1	UnB	a definir	a definir	S	2024/2025
2	UnB	a definir	a definir	S	2024/2025
3	UnB	a definir	a definir	S	2024/2025
4	UnB	a definir	a definir	S	2024/2025
5	UnB	a definir	a definir	S	2024/2025
6	Unicamp	Análise de similaridade semântica entre sentenças em corpora de textos jurídicos	IEEE International Conference on Semantic Computing	N	2024/2025
7	Unicamp	Metodologia para refinamento de modelos de linguagem visando aplicações no domínio jurídico	Annual Conference of the North American Chapter of the Association for Computational Linguistics	S	2024/2025
8	Unicamp	Investigação de técnicas para sumarização automática de textos na Língua	ACL - Annual Meeting of the Association for	S	2024/2025

		Portuguesa com base em aprendizagem de currículo	Computational Linguistics.		
9	Unicamp	Treinamento Multi-Sequencial para geração de texto com base em corpos textuais do domínio jurídico	International Natural Language Generation Conference	S	2024/2025
10	Unicamp	Geração de texto em linguagem natural em Português no domínio jurídico	Journal Computational Linguistics	S	2024/2025
11	Unicamp	Resultados em geração automática de linguagem no Português. Aplicações no contexto jurídico	Journal Expert systems with application	S	2024/2025
12	INutech	a definir	a definir	S	2024/2025
13	INutech	a definir	a definir	S	2024/2025
14	Eldorado	a definir	a definir	S	2024/2025
14	Eldorado	a definir	a definir	S	2024/2025
15	Eldorado	a definir	a definir	S	2024/2025
16	Eldorado	a definir	a definir	S	2024/2025

Outras conferências candidatas: Annual International Conference on Machine Learning; Conference on Empirical Methods in Natural Language Processing (EMNLP); ACM International Conference on Information and Knowledge Management.

15. BIBLIOGRAFIA

AB2L. “Radar de Lawtechs e Legaltechs Associadas”. Versão Janeiro de 2022. In: https://ab2l.org.br/ecossistema/radar-de-lawtechs-e-legaltechs/. Acessado em 26 de janeiro de 2022.

ASANO, Y. M., Rupprecht, C., Vedaldi, A., Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371, 2019.

BAHDANAU D., KYUNGHYUN C., BENGIO Y., Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2015.

BENGIO, Y., LOURADOUR J., COLLOBERT R., WESTON J., Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09), Association for Computing Machinery, New York, NY, USA, pp. 41–48., 2009.

BERRY, M. W., BROWNE, M., LANGVILLE, A. N., PAUCA, V. P., PLEMMONS, R. J., Algorithms and applications for approximate nonnegative matrix factorization, Computational Statistics & Data Analysis, vol. 52, edição 1, pp. 155-173, 2007.

BOBICEV, V., SOKOLOVA, M., Inter-annotator agreement in sentiment analysis: Machine Learning Perspective, 2017.

BOWMAN, S. R., ANGELI, G., POTTS, C., MANNING, C. D., A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

BROWN T. B., et al. Language Models are Few-Shot Learners. ArXiv:2005.14165, 2020.
CHAKRAVARTHI, B. R., JOSE, N., SURYAWANSHI, S., SHERLY, E., McCRAE, J. P., A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210, 2020.

COHEN, J., A coefficient of agreement for nominal scales, Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46, 1960.

CONSELHO NACIONAL DE JUSTIÇA. “Inteligência artificial no poder judiciário brasileiro”, Coordenação: José Antônio Dias Toffoli; Bráulio Gabriel Gusmão. Brasília/DF, 2019.

DUARTE, V. et al, Software livre: tendências, oportunidades e desafios, Observatório Softex, 2014.

DEVLIN, J., CHANG M. W, LEE, K., TOUTANOVA K., BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the North American Chapter of the Association for Computational Linguistics, 2019.

FERNANDES, G. L. Direito e Ciência de Dados: tendências e impactos da Quarta Revolução Industrial. In Inteligência Artificial Aplicada ao Processo de Tomada de Decisões, organizado por Henrique Alves Pinto, Jefferson Carús Guedes e Joaquim Portes de Cerqueira César. São Paulo: Editora D’Plácido. 2020.

FLEISS, J. L. Measuring nominal scale agreement among many raters, Psychological Bulletin, vol. 76, no. 5, pp. 378, 1971.

DZMITRY, B., CHO, K., BENGIO, Y., Neural Machine Translation by Jointly Learning to Align and Translate, arXiv: 1409.0473, 2016.

FULLER, S., 10 predictions: the legal function in 2025: Leaner, Faster, Smarter, KPMG, https://home.kpmg/xx/en/home/insights/2020/12/future-oflegal-article-series.html. Acessado em 12 de janeiro de 2022.

HARTMANN, N. et al., Portuguese word embeddings: Evaluating on word analogies and natural language tasks, arXiv preprint arXiv:1708.06025, 2017.

HOFMANN, T., Probabilistic latent semantic indexing, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50-57, 1999.

HOWARD, J., RUDER S., Universal language model fine-tuning for text classification, Proceedings of ACL - Annual Meeting of the Association for Computational Linguistics, 2018.
ILTA`s 2021 Technology Survey, Executive Summary, Net Documents, 2021.

JUCKETT, D., A method for determining the number of documents needed for a gold standard corpus. Journal of Biomedical Informatics, vol. 45, no. 3, 460-470, 2012.

JIMENEZ, S., GONZALEZ, F. A., GELBUKH, A., Mathematical properties of soft cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance, Information Sciences, vol. 367, pp. 373-389, 2016.

KRIPPENDORFF, K., Computing Krippendorff's alpha-reliability, 2011.

LIN, T., WANG, Y., LIU, X., QIU, X., A survey of transformers, arXiv preprint arXiv:2106.04554, 2021.

PETERS, M. E., NEUNMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K., ZETTLEMOYER, L., Deep contextualized word representations, Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACCL), 2018.

QIU, X., SUN, T., XU, Y., SHAO, Y., DAI, N., HUANG, X., Pre-trained models for natural language processing: A survey, Science China Technological Sciences, vol. 63, ed. 10, pp. 1872–1897, 2020.

RADFORD, A., WU J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER I., et al., Language models are unsupervised multitask learners, OpenAI Blog, vol. 1, no. 8, 9 páginas, 2019.

RAGHAV, K., REDDY, P. K., REDDY, V. B., Analyzing the extraction of relevant legal judgments using paragraph-level and citation information, AI4JCArtificial Intelligence for Justice, vol. 30, 2016.

RATNER, A., BACH, S. H., EHRENBERG, H., FRIES, J., WU, S., RÉ, C., Snorkel: Rapid training data creation with weak supervision, Proceedings of the VLDB Endowment: International Conference on Very Large Data Bases, vol. 11, no. 3, p. 269, NIH Public Access, 2017.

SCHMIDT, E., COHEN, J. The new digital age: transforming nations, businesses, and our lives. New York Times, 2013.

SETTLES, B., CRAVEN, M., An analysis of active learning strategies for sequence labeling tasks. Proceedings of the 2008 conference on empirical methods in natural language processing, pp. 1070-1079, 2008.

SOUZA F., NOGUEIRA R., LOTUFO R., Bertimbau: Pretrained bert models for Brazilian Portuguese, Intelligent Systems, Springer International Publishing, pp. 403 – 417, 2020.

VAN OPIJNEN, M., SANTOS, C., On the concept of relevance in legal information retrieval, Artificial Intelligence and Law, vol. 25, pp. 65-87, 2017.

VASWANI, A., et al., Attention Is All You Need, arXiv: 1706.03762, 2017.

WISSLER L., ALMASHRAEE, M., DÍAZ, D. M., PASCHKE, A., The Gold Standard in Corpus Annotation. IEEE German Student Conference, 21, 2014.

WOHNOUTKA, B. "ISPS6 - Inevitable Cloud Migration: Actionable Strategies for Healthcare Providers". Apresentação realizada no Gartner Symposium/ITxpo – Orlando, Fl. – 6 de outubro, 2013.

XIA, P., WU S., DURME, V. B., Which* BERT? a survey organizing contextualized encoders., Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 7516–7533, 2020.

YEUNG, L. et al. , Informatização Judicial e Efeitos sobre a Eficiência da Prestação Jurisdicional e o Acesso à Justiça, Insper, dezembro de 2020.

Artigos em sites:

Barreto, G. "O que você precisa saber sobre a nova realidade da Advocacia 5.0", 12 fev 2020. Atualizado em 14 dez 2021. In: https://www. aurum.com.br/blog/advocacia5-0/ Acessado em 14 de fevereiro de 2022.

Equipe TD. "Revolução no sistema jurídico: entenda o que são Lawtechs". TD. In: https://transformacaodigital.com/revolucao-no-sistema-juridico-entenda-o-que-saolawtechs/ Acessado em 14 defevereiro de 2022.

Equipe TD. "ROSS, o primeiro robô advogado do mundo". In: https://transformacaodigital. com/juridico/ross-o-primeiro-robo-advogado-do-mundo/ Acessado em 14 de fevereiro de 2022.

Grillo, B. "Excesso de plataformas de processo eletrônico atrapalha advogados". Atualizado em 4 de outubro de 2017. In: https://www.conjur.com.br/2017-out03/excesso-sistemas-processo-eletronico-atrapalham-advogados. Acessado em 14 de fevereiro de 2022.

Neoway. "Entenda o que é Inteligência Jurídica e quais suas vantagens". In: https://blog.neoway.com.br/inteligencia-juridica/ Acessado em 14 de fevereiro de 2022.

Neoway. Jurimetria: O que é, para que serve e seus pilares para o Direito. In: https://blog.neoway.com.br/o-que-e-jurimetria/ Acessado em 14 de fevereiro de 2022

Sites visitados:

https://www.tikal.tech. Acessado em 15 de fevereiro de 2022.

https://jusapi.com.br. Acessado em 15 de fevereiro de 2022.

https://datalawyer.com.br. Acessado em 15 de fevereiro de 2022.

https://neoway.com.br. Acessado em 15 de fevereiro de 2022.

http://dria.unb.br/teste-top. Acessado em 18 de fevereiro de 2022

http://nido.unb.br Acessado em 20 de fevereiro de 2022

http://dria.unb.br/teste-top Acessado em 20 de fevereiro de 2022

https://ailab.unb.br/projetos Acessado em 22 de fevereiro de 2022

Anexo I – Detalhamento da Etapa de Rotulagem

Tipos de Documentos

1. Por tipos de processos de Controle Externo:
    i. Representações de Aquisições Públicas;
    ii. Denúncias de Aquisições Públicas.
2. Por tipo de peça jurídica:
    i. Petição inicial;
    ii. exame sumário;
    iii. instruções preliminares;
    iv. comunicações do Tribunal;
    v. manifestações dos interessados;
    vi. instrução de mérito;
    vii. pronunciamento da subunidade;
    viii. pronunciamento da unidade técnica.

Quantidade de Documentos a serem rotulados

- Segundo o Termo de Referência, há um total aproximado de **2.000 Representações e Denúncias** de aquisições públicas encerradas e disponíveis como massa de treinamento para modelos computacionais.
- Estimando-se **uma média de três peças diferentes para cada processo encerrado**, teremos cerca de **6.000 documentos para serem rotulados**.
- Estimando-se que cada estagiário consiga rotular **dois documentos por hora**, e cada documento seja rotulado por **dois anotadores**, conforme metodologia a ser adotada de revisão por par, serão necessárias **6.000 horas de rotulagem**.
- Uma equipe de **10 estudantes do curso de Direito**, com dedicação de **80 horas mensais**, totalizando **800 horas mensais**, precisaria de **7,5 meses para rotular** todos os documentos. Somado esse prazo a mais duas semanas de treinamento no uso da ferramenta a ser escolhida e na aplicação dos protocolos que serão elaborados, termos o total de **8 meses para a etapa de rotulagem**.
- Após a atividade de rotulagem, deverá haver a atividade de **conciliação de divergências** das anotações, que em geral, pode ser estimada em metade do esforço da rotulagem,

ou seja, cerca de mais **4 meses**, totalizando **12 meses para a disponibilização da base de documentos rotulados**.

Planejamento da Execução da Etapa de Rotulagem

- O esforço total de rotulagem e conciliação de divergências deverá ser dividido em 4 lotes de 500 processos de Representações e Denúncias;
- De acordo com as estimativas anteriores, o prazo para a rotulagem de cada lote de 500 processos seria de cerca de 2 meses, somados a um mês adicional para o tratamento das divergências observadas, totalizando assim 3 meses para todo o procedimento de cada lote.

Anexo III - Métricas de performance aplicadas à modelos de aprendizado de máquina e aprendizado profundo para a classificação

A presente Nota Técnica tem como objetivo apresentar as principais métricas utilizadas em problemas de classificação, suas características e propósitos, tanto sob perspectivas técnicas como de negócio.

No desenvolvimento de projetos de Ciência de Dados utilizando técnicas de aprendizado de máquina ou de aprendizado profundo, deve ser escolhido um conjunto apropriado de métricas para a avaliação da qualidade dos modelos de classificação, de acordo com as especificidades de cada tipo de problema enfrentado e de aplicação.

As métricas de avaliação de modelos de classificação são funções matemáticas com o propósito de medir as taxas de erros e de acertos desses modelos, não sendo nenhuma dessas métricas genericamente melhor que as demais. Algumas métricas são matematicamente mais simples, enquanto outras um pouco mais complexas, ajustando-se às características dos datasets: algumas métricas funcionam melhor de acordo com o tamanho do dataset, enquanto outras de acordo com a homogeneidade ou não da proporção de dados no dataset pertencentes a cada classe.

Para ilustrar a aplicação dessas métricas, utilizaremos dois tipos de exemplos: um na área de saúde e outro na área jurídica, mas que poderão ser facilmente repensados para qualquer outra área de atividade.

As quatro métricas de performance comumente utilizadas para a avaliação de modelos de aprendizado de máquina, voltados à classificação e categorização de entidades e variáveis são as seguintes:

- Acurácia;
- Precisão;
- Revocação (recall, ou sensibilidade);
- F1-Score

A acurácia, ou taxa de acerto de um modelo de classificação, indica o percentual de predições corretas em relação ao total de elementos presentes em um dataset. Apesar de configurar um tipo de métrica importante, e que indica a qualidade geral de um modelo, para certas situações, nas quais as taxas de erro e de acerto de cada classe tenham importâncias diferentes para o problema de negócio, ou que haja desproporção entre a presença dos elementos de cada classe, também deve-se levar em consideração as métricas de precisão, revocação e F1-Score, para a avaliação da qualidade de tais modelos.

- **Valores de Classe:** no caso de variáveis binárias, as classes podem assumir os seguintes valores ou categorias:
    - Verdadeiro Positivo (TP): classes corretamente classificadas pelo modelo como positivas;
    - Falso Positivo (FP): classes erroneamente classificadas pelo modelo como positivas;

- Verdadeiro Negativo (TN): classes corretamente classificadas pelo modelo como negativas;
- Falso Negativo (FN): classes erroneamente classificadas pelo modelo como negativas.

Ex.: No caso dos exemplos anteriores, de pacientes com suspeita de câncer, ou de decisões em processos trabalhistas:

- Verdadeiros Positivos correspondem às predições corretamente classificadas de pacientes com tumores, ou de tumores corretamente classificados como malignos, ou de trechos de texto corretamente classificados como decisões, ou de decisões corretamente classificadas como favoráveis ao empregado;
- Falsos Positivos correspondem às predições erroneamente classificadas de pacientes com tumores, ou de tumores erroneamente classificados como malignos, ou de trechos de texto erroneamente classificados como decisões, ou de decisões erroneamente classificadas como favoráveis ao empregado;
- Verdadeiros Negativos correspondem às predições corretamente classificadas de pacientes sem tumores, ou de tumores corretamente classificados como não sendo malignos, de trechos de texto corretamente classificados como não sendo decisões, ou de decisões corretamente classificadas como não sendo favoráveis ao empregado;
- Falsos Negativos correspondem às predições erroneamente classificadas de pacientes sem tumores, ou de tumores erroneamente classificados como não sendo malignos, de trechos de texto erroneamente classificados como não sendo decisões, ou de decisões erroneamente classificadas como não sendo favoráveis ao empregado.

- **Matriz de Confusão:** matriz de valores que podem ser assumidos pelas classes de uma amostra.

  É utilizada para o cálculo das métricas de acurácia, precisão e revocação, a partir de uma tabela com os erros e acertos do modelo, quando comparados com os dados rotulados como referência no dataset de treinamento.

  Para a avaliação de classes binárias, a Matriz de Confusão apresenta apenas duas linhas e duas colunas, como ilustra a Figura 1, a seguir:

MATRIZ DE CONFUSÃO		Classe Real	
		Positiva	Negativa
Predição de Classe pelo Modelo	Positiva	Verdadeiro Positivo (TP)	Falso Positivo (FP)
	Negativa	Falso Negativo (FN)	Verdadeiro Negativo (TN)

Figura A: Matriz de confusão para classificação binária. Fonte: iNuTech.

Para o caso dos pacientes com suspeita de câncer, o dataset rotulado para a matriz de confusão do exemplo abaixo, é constituído por exames de imagem, previamente analisados, identificados pelos resultados de tumores presentes como malignos ou não, enquanto para o exemplo das decisões de processos trabalhistas, o dataset rotulado é constituído por sentenças e acórdãos previamente analisados, com as decisões identificadas e classificadas como favoráveis ou não ao empregado ou ao empregador.

Ex.: Considere um conjunto de 1.000 exames de imagem, ou de decisões de processos judiciais trabalhistas, dos quais 100 exames são de pacientes realmente com câncer, apresentando tumores malignos, ou que 100 decisões são favoráveis ao empregador, e que um modelo treinado em um dataset previamente rotulado, gere as seguintes predições hipotéticas para esse conjunto de dados, conforme a matriz de confusão, apresentada na Figura B.

MATRIZ DE CONFUSÃO		Classe Real	
		Pacientes com câncer / decisões favoráveis à empresa	Pacientes sem câncer / decisões favoráveis ao empregado
Predição de Classe pelo Modelo	Com câncer/ decisão favorável a empresa	Cenário #1 - TP 80	Cenário #2 - FP 80
	Sem câncer/ decisão favorável ao empregado	Cenário #3 - FN 20	Cenário #4 - TN 820

Figura B: Exemplo de Matriz de Confusão para classificação binária. Fonte: autoria própria.

- Cenário #1: O modelo corretamente efetuou predições de câncer para pacientes com câncer, ou de decisões favoráveis à empresa (Verdadeiro Positivo);
- Cenário #2: O modelo erroneamente efetuou 80 predições positivas de câncer para pacientes sem câncer, ou de decisões favoráveis à empresa, que foram na realidade favoráveis ao empregado (Falso Positivo);
- Cenário #3: O modelo erroneamente efetuou 20 predições de inexistência de câncer para pacientes com câncer, ou de decisões favoráveis ao empregado, que foram na realidade favoráveis à empresa (Falso Negativo);

- Cenário #4: O modelo corretamente efetuou predições de inexistência de câncer para pacientes sem câncer, ou de decisões favoráveis ao empregado (Verdadeiro Negativo).

Das quatro situações hipotéticas acima, os cenários #1 e #4 são os ideais, porém as predições errôneas dos cenários #2 e #3, são resultados indesejados, que podem gerar diferentes desdobramentos:

- O Cenário #2, de Falsos Positivos (FP), indica que dos 900 pacientes que em realidade possuem tumores benignos, o modelo fez a predição de que 80 desses pacientes tinham câncer. Ou, no caso das decisões trabalhistas, que das 900 decisões que foram de fato favoráveis ao empregado, o modelo fez a predição de que 80 dessas decisões seriam favoráveis à empresa;
- O cenário #3, de Falsos Negativos (FN), indica que dos 100 pacientes que em realidade tinham câncer, o modelo fez a predição de que 20 desses pacientes possuíam tumores benignos. Ou, no caso das decisões trabalhistas, que das 100 decisões que de fato foram favoráveis à empresa, o modelo fez a predição de que 20 dessas decisões seriam favoráveis ao empregado.

Para pacientes com tumores benignos, que erroneamente tenham falsos positivos, esse resultado certamente irá gerar stress, exames e gastos financeiros desnecessários. Porém, para pacientes com tumores malignos, mas que erroneamente tenham falsos negativos, esse resultado pode levar a risco de morte.

Em função dos diferentes cenários é possíveis resultados descritos anteriormente, e das características de cada tipo de aplicação, são utilizadas diferentes métricas de qualidade, descritas a seguir:

- **Acurácia:** proporção de predições corretas geradas por um modelo (Verdadeiros Positivos + Verdadeiros Negativos), sobre o total de elementos de um determinado dataset:
  - $\text{Acurácia} = \frac{\text{Verdadeiros Positivos (TP)+Verdadeiros Negativos (TN)}}{\text{Verdadeiros Positivos (TP)+ Verdadeiros Negativos (TN)+ Falsos Positivos (FP)+ Falsos Negativos (FN)}}$

Ex.: Para o caso da matriz de confusão da figura 2, temos a seguinte métrica de acurácia, indicando que o modelo prediz corretamente o resultado de 90% dos exames que apresentam tumores, ou do resultado das decisões de processos trabalhistas:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} = \frac{820 + 80}{1.000} = 90\%$$

- **Precisão:** indica o percentual das predições positivas corretas, em relação ao total das ocorrências que são realmente positivas. É uma métrica adequada aos casos em que Falsos Positivos (TP) são considerados mais críticos que Falsos Negativos (FN).

$$Precisão = \frac{VerdadeirosPositivos(TP)}{VerdadeirosPositivos(TP) + FalsosPositivos(FP)}$$

Ex.: Para o caso da matriz de confusão da figura 2, temos a seguinte métrica de precisão, indicando que o modelo predisse 160 ocorrências positivas (pacientes com câncer, ou decisões favoráveis à empresa), mas apenas 80 eram de fato positivas, configurando uma precisão de 50%:

$$Precisão = \frac{TP}{TP + FP} = \frac{80}{80 + 80} = 50\%$$

- **Revocação (recall):** indica o percentual amostras reconhecidas como sendo de uma classe, que o modelo categoriza corretamente como positivas, sobre o total das ocorrências dessa classe que são realmente positivas. É uma métrica adequada aos casos em que Falsos Negativos (FN) são considerados mais críticos que Falsos Positivos (TP).

$$Revocação = \frac{VerdadeirosPositivos(TP)}{VerdadeirosPositivos(TP) + FalsosNegativos(FN)}$$

Ex.: Para o caso da matriz de confusão da figura 2, temos a seguinte métrica de revocação, indicando que o modelo predisse corretamente 80 ocorrências positivas (pacientes com câncer, ou decisões favoráveis à empresa), mas de fato 100 eram ocorrências positivas, configurando uma revocação de 80%:

$$Revocação = \frac{TP}{TP + FN} = \frac{80}{80 + 20} = 80\%$$

Em resumo:

  - Quanto maior o percentual de Falsos Positivos (FPs), menor a precisão. Baixa precisão indica a ocorrência de percentual elevado de falsos positivos;
  - Quanto maior o percentual de Falsos Negativos (FNs), menor a revocação (ou recall). Baixa revocação indica a ocorrência de percentual elevado de falsos negativos;
  - Há sempre que se considerar o *trade-off* entre precisão e revocação. O esforço para aumentar a precisão, implica no custo de se reduzir a revocação, e vice-versa;
  - Um modelo com elevado percentual de acurácia, pode eventualmente produzir resultados com baixos índices de precisão ou de revocação. Quando falsos positivos ou falsos negativos tiverem uma importância relativamente elevada, torna-se aconselhável utilizar concomitantemente com a acurácia, as métricas de precisão e revocação;
  - A métrica de acurácia presta-se principalmente para os casos em que falsos positivos e falsos negativos possuem impactos semelhantes. Se os impactos de falsos positivos e falsos negativos forem muito diferentes, como no caso dos

exames para avaliar a suspeita de câncer, torna-se aconselhável também considerar as métricas de precisão e revocação dos modelos de predição.

- **F1-Score:** é uma média harmônica das métricas de precisão e de revocação, considerando simultaneamente falsos positivos e falsos negativos. Sendo comumente utilizada para avaliar a qualidade de modelos de classificação, a métrica F1-Score, combina precisão e revocação em uma única métrica, indicando a qualidade geral do modelo.

  O F1-Score é particularmente adotado para a avaliação de modelos de processamento de linguagem natural, por ser apropriado o seu emprego a datasets com amostras desproporcionais de classes presentes, situação comum em PLN.

  Um resultado baixo do F1-Score é um indicativo de que uma das métricas, ou as duas, de precisão e revocação, estão igualmente baixas.

$$F1Score = \frac{2 * (Revocação * Precisão)}{(Revocação + Precisão)}$$

  Ex.: Para o caso da matriz de confusão da figura 2, temos o seguinte F1-Score:

$$F1score = \frac{2 * (0{,}80 * 0{,}50)}{(0{,}80 + 0{,}50)} = \frac{2 * 0{,}40}{1{,}30} = 0{,}615$$

Considera-se um bom modelo de classificação, aquele que consegue um balanceamento equilibrado entre precisão e revocação, de acordo com as características da aplicação a que se destina o modelo. Percentuais elevados de precisão e revocação, implicam em percentual elevado de acurácia.

- **Quadro Resumo**

Métrica	Fórmula
Acurácia	$\frac{TP + TN}{TP + TN + FP + FN}$
Precisão	$\frac{TP}{TP + FP}$
Revocação	$\frac{TP}{TP + FN}$
F1-Score	$\frac{2 * (Precisão * Revocação)}{(Precisão + Revocação)}$

Fazendo TP + FP = P (soma dos positivos),

$$TemosquePrecisão = \frac{TP}{P}$$

TP = Precisão x P

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

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

$FN = \frac{Precisão x P}{Recall} - Precisão x P$ = Precisão ($\frac{P}{Recall} - P$

Fazendo TP + FP + TN + FN = 1 (soma de todos os elementos do conjunto).

Temos que

Acurácia = TP + TN e

TN = 1 – P – FN

Assim,

Acurácia = TP + 1 – P – FN = (Precisão x P) + 1 – P – $\frac{Precisão x P}{Recall} + Precisão x P$ = 2P x

Acurácia = 2P x Precisão – $\frac{Precisão x P}{Recall}$ + 1 – P = Precisão (2P - $\frac{P}{Recall}$) + 1 – P

Anexo IV – Respostas aos Questionamentos apresentados

RES: Dúvidas: Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assistida por Inteligência Artificial

Seafi <seafi@tcu.gov.br>
Wed 09/03/2022 17:57

To: Monika Gomes Heringer <monika.heringer@eldorado.org.br>; Seafi <seafi@tcu.gov.br>
Cc: Joao Marcos da Paixao <Joao.Paixao@eldorado.org.br>; Marcia Regina Guedes Brandao <marcia.brandao@eldorado.org.br>

Prezada sra. Monika Heringer,

Uma das caracterís cas do Chamado Público em questão é a possibilidade de revisão, pelo TCU, dos critérios con dos no Termo de Referência da ETEC. Inclusive do orçamento es mado. É preciso, para tal, que a Proposta de P&D apresente planilha de formação de preços detalhando os custos e prazos es mados para o alcance dos marcos do projeto.

Atenciosamente,

Comissão de Seleção da ETEC
SEAFI - Serviço de apoio à fiscalização de contratos de TI
https://tcu.gov.br/etec

Missão do TCU
Aprimorar a Administração Pública em bene cio da sociedade por meio do controle externo.

De: Monika Gomes Heringer <monika.heringer@eldorado.org.br>
Enviada em: sexta-feira, 4 de março de 2022 15:44
Para: Seafi <seafi@tcu.gov.br>
Cc: Joao Marcos da Paixao <Joao.Paixao@eldorado.org.br>; Marcia Regina Guedes Brandao <marcia.brandao@eldorado.org.br>
Assunto: Re: Dúvidas: Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assis da por Inteligência Ar ficial

Prezado Comitê,

O Ins tuto de Pesquisa Eldorado, a Universidade de Brasília - UnB -, a Universidade Estadual de Campinas e o Ins tuto NuTech de Pesquisa Aplicada em Ciência, Tecnologia e Inovação, estão bastante dedicados na elaboração de uma proposta robusta do Projeto de P&D,

referente ao Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assis da por Inteligência Ar ficial, a ser enviada até o dia 25/03/22.

Entretanto gostaríamos de saber há alguma expecta va de alteração doorçamento referente à Encomenda Tecnológica, entendendo a complexidade do projeto e visando uma entrega de al ssima qualidade?

Aguardamos o retorno.

Atenciosamente,

Monika Heringer
Consultora de Novos Negócios
61 98177-1380
monika.heringer@eld
orado.org.br Ins tuto
de Pesquisa Eldorado
www.eldorado.org.br

From: Seafi <seafi@tcu.gov.br>
Sent: 17 February 2022 18:50
To: Monika Gomes Heringer <monika.heringer@eldorado.org.br>; Seafi <seafi@tcu.gov.br>
Cc: Joao Marcos da Paixao <Joao.Paixao@eldorado.org.br>; Marcia Regina Guedes Brandao <marcia.brandao@eldorado.org.br>
Subject: RES: Dúvidas: Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assis da por
Inteligência Ar ficial

Prezados,

Seguem as respostas solicitadas. A Comissão de Seleção da ETEC continua à disposição dos senhores, caso desejem agendar uma videoconferência para esclarecimentos adicionais.

1. Especialistas do TCU nas áreas de domínio do projeto irão par cipar da equipe de desenvolvimento? (Na pág. 40 do Edital é mencionado que o TCU irá *"designar equipe da comissão de negócio com usuários finais do sistema de instrução de representações e denúncias"*.)

a. Com que dedicação?
b. Qual quan dade de profissionais do TCU?
c. Poderão par cipar das a vidades de orientação aos cien stas de dados?
d. Poderão par cipar das a vidades de orientação aos desenvolvedores de front-end?

Irão compor a própria equipe ágil de desenvolvimento: a) um gerente de projeto do TCU, encarregado de fazer a ponte entre a contratada com a equipe de TI que desenvolve a solução de instrução assistida; b) o dono do produto (product owner), que fará a ligação com a equipe de usuários finais da solução.
O gerente de projeto participará de todas as cerimônias ágeis, incluindo as reuniões diárias. O dono do produto participará ao menos das cerimônias ágeis de encerramento e início das sprints.
Os dois profissionais citados (e seus substitutos eventuais) representam cerca de 15 outros profissionais das equipes de desenvolvimento da solução e de instrução de processos. A função deles é facilitar a interação, sempre que necessário, da equipe da contratada com os profissionais das equipes internas do TCU. Portanto, os auditores do TCU encarregados do desenvolvimento da solução ou da instrução dos processos poderão sim participar das atividades de orientação aos profissionais da contratada, tanto cientistas de dados quanto desenvolvedores. Há como avaliar a documentação que servirá para subsidiar o treinamento dos modelos?

e. Qual o tamanho médio, em número de caracteres/páginas de cada po de documento?
f. Es ma va de custos das a vidades de rotulagem de documentos para treinamento de modelos (ver págs. 46 e 47 do Edital: *"A contratante não se envolverá neste serviço, exceto como consultora e como avaliadora, por amostragem, da qualidade dos dados anotados"* e *"Caso o fornecedor interessado entenda ser necessário rotular ou estruturar tais documentos, para poder usá-los no treinamento de modelos computacionais, este passo deve estar explicitamente considerado no Projeto de P&D apresentado. Incluindo a es ma va de custos da mão de obra humana necessária ao trabalho"*.

Há grande variação entre os documentos e não sabemos responder seu tamanho médio, em páginas ou caracteres. Por isso a primeira fase da contratação foi pensada como fase de saneamento de dados, na qual a contratada deverá gerar como produto para o TCU um banco com os dados rotulados e anotados. Após a primeira fase a ETEC será reavaliada quanto à sua viabilidade em termos técnicos e financeiros, podendo ser encerrada.
Na fase de saneamento do primeiro ciclo temático (Aquisições Públicas), há ~2000 processos (aproximadamente 8000 documentos, compreendendo petição inicial, instruções e comunicações) a serem trabalhados. Cabe aos interessados estimarem os custos de rotulagem.

2. Existem dados confiáveis, suficientes para responder cada pergunta de negócio? (Na pág. 40 do Edital é mencionado que o TCU irá *"disponibilizar as bases de dados as quais os colaboradores da empresa terão acesso"*, e nas págs. 43 e 44, há uma relação e quan dade de documentos a serem disponibilizados.)
   a. Que po de dados estruturados?
      i. Onde buscar?
      ii. Quais formatos?
      iii. Dados públicos/privados?

iv. Dados sensíveis?

b. Quais documentos com dados não estruturados?

i. Onde buscar?

ii. Quais formatos?

iii. Algum po de documento como imagem ou com problemas de digitalização? (ver páginas

iv. Dados públicos/privados?

v. Dados sensíveis?

Os poucos dados estruturados são, essencialmente, metadados ligados ao protocolo de cada processo. Exemplos:

- Data de autuação,
- Órgão/entidade jurisdicionada,
- Responsáveis,
- Interessados,
- Natureza,
- Assunto,
- Irregularidades
- alegadas Existe
- pedido

cautelar?

Existe pedido de oitiva?

Existe proposta de ações saneadoras?

Deliberações

No entanto, não se pode garantir a qualidade do preenchimento destes metadados.

Os documentos não estruturados são os PDFs/DOCXs das peças processuais citadas no Termo de Referência. As petições iniciais são digitalizadas, e têm problemas de reconhecimento de caracteres.

As demais peças são nativamente criadas em meio digital.

Todos os dados a serem disponibilizados são públicos, peças de processos já encerrados. No entanto, podem conter dados sensíveis (identificação de pessoas), que não devem ser utilizados para o treinamento dos modelos computacionais.

3. Qual a forma de prestação de contas?

a. Será necessária a emissão de notas fiscais?

b. Existe um Manual Orienta vo para a apresentação da Prestação de Contas?

c. A Prestação de Contas financeira se dará apenas ao final do contrato?

d. Deverá ser realizada por Regime Caixa?

e. Se sim, será preciso conta bancária especifica para o projeto?

f. A conta bancária poderá ser aberta em qualquer banco?

Os pagamentos serão mensais mediante apresentação de fatura dos serviços prestados, no valor do custo fixo mensal acertado entre as partes para o respectivo marco do

projeto, na fase de negociação, com base nas planilhas de custo fornecidas pela interessada. É mais compatível, então, com o Regime de Caixa, mas não é necessário ter conta específica para o projeto, a menos que tal exigência exista por motivos da própria contratada. Como não se trata de reembolso, não há necessidade de prestação detalhada de contas das despesas realizadas.

4. Estão claros os critérios para o aceite dos produtos de dados?
    a. Quais as métricas de acurácias/precisão/recall para cada modelo?
    b. As métricas serão acordadas a posteriori?
    c. Riscos a considerar (Ver pág. 48 do Edital): *"Como o formato do documento é livre, cada denunciante/representante tem liberdade para estruturar a peça como bem entender, o que confere grande variabilidade ao histórico de pe ções disponível".*

Não estão claros os critérios para aceite dos produtos, em termos de métricas de acurácia, precisão, recall etc. De fato, não está clara nem a possibilidade de existência do produto. Para o Projeto de P&D, pede-se que a interessada apresente uma proposta de quais métricas utilizar para cada marco do projeto. Porém, ao final da fase de saneamento de cada ciclo temático, as partes negociarão as métricas, as faixas de bônus e seus valores. Esta negociação poderá ser revista ao longo do desenvolvimento do produto, sempre que ambas as partes concordarem que houve impacto devido ao risco tecnológico.

5. Estão claros os critérios para a remuneração de incen vo?
    a. Quais as métricas de acurácias/precisão/recall para cada modelo?
    b. As métricas serão acordadas a posteriori? Vide questão anterior.
6. Como e quando serão definidas as reuniões com estes fiscais e apresentação de resultados, dados e documentos do projeto?

Estas definições fazem parte da proposta de Projeto de P&D da interessada. O edital exige que seja adotada metodologia ágil com cerimônias frequentes e artefatos pré-definidos.

Atenciosamente,

Comissão de Seleção da ETEC
SEAFI - Serviço de apoio à fiscalização de contratos de TI
https://tcu.gov.br/etec

De: Monika Gomes Heringer <monika.heringer@eldorado.org.br>
Enviada em: terça-feira, 15 de fevereiro de 2022 22:40
Para: Seafi <seafi@tcu.gov.br>
Cc: Joao Marcos da Paixao <Joao.Paixao@eldorado.org.br>; Marcia Regina Guedes Brandao <marcia.brandao@eldorado.org.br>

Assunto: Dúvidas: Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assis da por Inteligência Ar ficial

Prezado Comitê,

O Ins tuto de Pesquisa Eldorado vem por meio deste, solicitar esclarecimentos das dúvidas, documento anexo neste email, referentes ao Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assis da por Inteligência Ar ficial.

Aguardamos o retorno.

Atenciosamente,

Monika Heringer
Consultora de Novos Negócios
61 98177-1380
monika.heringer@eld
orado.org.br Ins tuto
de Pesquisa Eldorado
www.eldorado.org.br