# Relatório Técnico Parcial 5 do Projeto KnEDLe / NIDO

Alice Borges, Carolina A. Okimoto, Luís P. F. Garcia
Thiago P. Faleiros, Marcelo Mandelli, Vinícius R. P. Borges
Ricardo Marcacini, Andrei Lima Queiroz

**Universidade de Brasília**
**Departamento de Ciência da Computação**
http://nido.unb.br

28 de dezembro de 2022

## 1 Introdução

O presente relatório tem como objetivo elencar os resultados produzidos no Projeto de Pesquisa *KnEDLe - Extração de Informações de Publicações Oficiais usando Inteligência Artificial*. O projeto é fruto de uma parceria entre a Universidade de Brasília (UnB), a Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) e a Fundação de Empreendimentos Científicos e Tecnológicos (FINATEC)[1]. Este relatório trata das atividades e resultados produzidos na quarta fase do projeto (*release* 5) no período de 03/01/2022 a 20/07/2022. Cabe ressaltar que houve uma prorrogação da *release* por causa da finalização da parte de contratos e ajustes do *Dash* para integração da timeline. Essa intregração e os resultados foram apresentados no *Workshop* Nido que ocorreu em 29 de julho de 2022.

## 2 Viabilidade Técnica

Na *release 5* o projeto continuou focado no desenvolvimento do produto, porém, algumas pesquisas também foram realizadas. O foco dessa release foi a parte de contratos e licitações, onde tivemos o time *Contratos* trabalhando na identificação e segmentação dos atos. O enfoque também foi dividido para o time *Anotação* que ficaram responsáveis pela criação do corpus padrão ouro de contratos e licitação. Além disso, outro foco foi a indução para segmentação e reconhecimento de entidades para atos de pessoal. Para o produto acompanhar os avanços, a equipe também trabalhou bastante na ferramenta, tanto no DODFMiner para

[1]Este projeto possui estes registros nas respectivas instituições envolvidas: FAPDF convênio 07/2019; UnB SEI:23106.058975/2019-62; Finatec 6429 - FAPDF/CIC.

a adição de modelos como na interface com o usuário (o *frontend*), onde seria possível o usuário final utilizar facilmente novas funcionalidades.

O time do projeto KnEDLe, na *Release 5*, passou por algumas mudanças. Dois professroes/supervisores saíram do projeto, e também alguns alunos que já estávam a algum tempo se formaram e tiveram que sair. Portanto, foi necessário contratar novos alunos e professores e reestruturar alguns times.

A equipe de viabilidade técnica, nessa release, focou bastante em acompanhar os novos times e garantir que eles se adaptassem o melhor possível e assim evitar desalinhamento na realização das tarefas. Como nas releases anteriores, para que pudéssemos acompanhar o andamento dos times e suas realizações utilizamos das ferramentas já mencionadas anteriormente, OKR e para nosso kanban o Zenhub.

## 2.1 Zenhub

Após a adoção da ferramenta Zenhub na *Release 4*, o acompanhamento das atividades realizadas ficou bem mais simples e rápido pois todas as tarefas se encontram centralizadas em um só local. Dessa maneira, a equipe de viabilidade técnica conseguiu verificar o andamneto das tarefas de cada time semanalmente e garantir que estivessesm trabalhando dentro das metas estabelecidas pelo OKR. O pipeline das taferas continuou igual ao da *Release 4*. A estrutura do Kanbam é a seguinte: a) *Backlog* quando as tarefas são criadas; b) *To Do* para quando as tarefas estão planejadas para aquela *sprint*; c) *In Progress* para tarefas em andamento; d) *Done* para tarefas finalizadas e; e) *Closed* para tarefas finalizadas e revisadas.

A Figura 1 apresenta o progresso das tarefas realizadas pelo projeto a cada mês, portanto, é possível observar que o número de tarefas criadas foi maior do que o número de tarefas finalizadas, isso se deve ao faro de que algumas tarefas são repriorizadas de acordo com as necessidades de cada time e assim são deixadas para serem realizadas em outro momento ou até deixam de fazer sentido para o projeto. Apesar disso, essa diferença não é expressiva de fato. Outro fator interessante de analisar, são as tarefas propostas a serem feitas nas sprints (*"To Do"*), e aqui é notório que a maioria das tarefas propostas foram inicializadas e finalizadas, possuindo uma diferença mínima nesse número.

A Figura 2 também mostra um relatório das tarefas realizadas na *Release 5*, analisando o relatório, ele apresenta que 87% do escopo estimado foi concluído, onde to total de 329 issues e pull requests abertos, 298 foram concluídos.

## 2.2 OKR

O acompanhamento dos riscos foram realizados mensalmente por meio do *Objective and Key Results* (OKR)[17]. Coube ao time de Viabilidade Técnica conectar os objetivos do OKR com os da Estrutura Analítica do Projeto (EAP), os objetivos chave da *Release* 5 e as atividades registradas no ZenHub.

No início da *Release* 5, os objetivos chaves foram definidos para cada time e adicionados à planilha de acompanhamento. O acompanhamento do OKR foi realizado a cada 15 dias, onde os objetivos eram atualizados na reunião Geral com todos os alunos presente para garantir a evolução do projeto e identificação de lacunas a serem preenchidas. Além disso, todos do projeto conseguiriam ter ampla visão dessa evolução. Nessa *Release* a definição dos objetivos

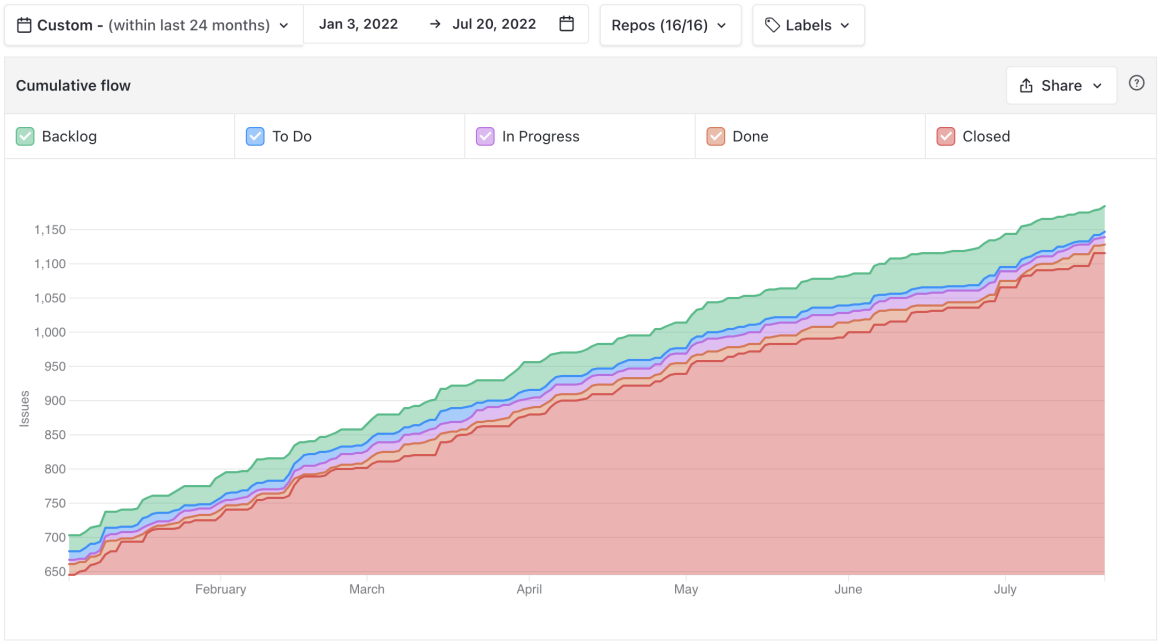



Figura 1: Acompanhamento das tarefas realizadas durante a *Release* 5

demorou um pouco mais para serem definidas, pois na *Release* anterior os objetivos ficaram extensos, então foi necessário uma definição mais acertiva.

A Figura 3 revela a evolução do OKR da equipe pelas *sprints* da *Release* 5. O eixo-$x$ destaca as *sprints* enquanto o eixo-$y$ destaca a porcentagem de conclusão dos objetivos da *release*. Os objetivos do time de Viabilidade Técnica, a evolução dos produtos gerados e as atividades de pesquisa são ilustrados pelas linhas verde, azul e amarela, respectivamente. Ao analisar o gráfico de evolução do OKR, tem-se que 100% dos objetivos de Viabilidade Técnica foram atingidos. Já os objetivos dos itens *Produto* e *Pesquisa* atingiram 71% e 74% respectivamente. Levando em consideração a metodologia OKR, as metas devem ser arrojadas de forma que o alcance de 70% signifique que o objetivo está praticamente atingido, já agregou valor e será concluído em breve. Com issso, podemos concluir que o projeto obteve sucesso em sua evolução nessa *Release*.

# 3 Knedash

Os DODFs são uma fonte de informação valiosa sobre o Distrito Federal, pois tem grande valor informativo, longevidade e validação. Contudo, ele é estruturado com linguagem natural, complicando o aproveitamento dessas informações em sistemas e algoritmos. Para sobrepor essa barreira, foi desenvolvida uma ferramenta entitulada *Knedash*[2] capaz de realizar a extração, a pesquisa e o armazenamento dessas informações por meio de técnicas de Mineração de Texto, Aprendizado de Máquina e Processamento de Linguagem Natural. A Figura 4 apresenta as principais telas da ferramenta *Knedash* e resultados da interação com o usuário.

[2] `http://knedash.unb.br/`

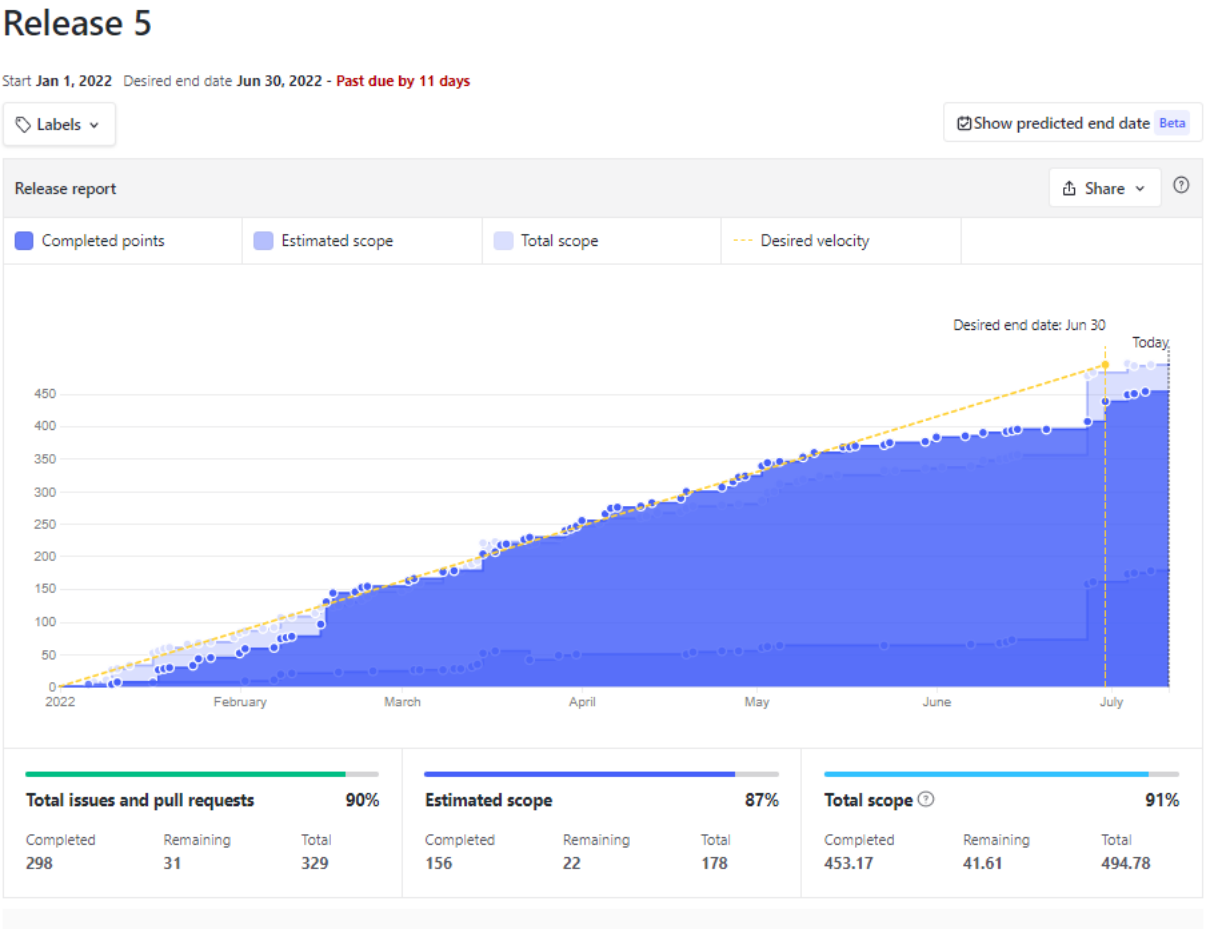



Figura 2: Relatório da *Release* 5

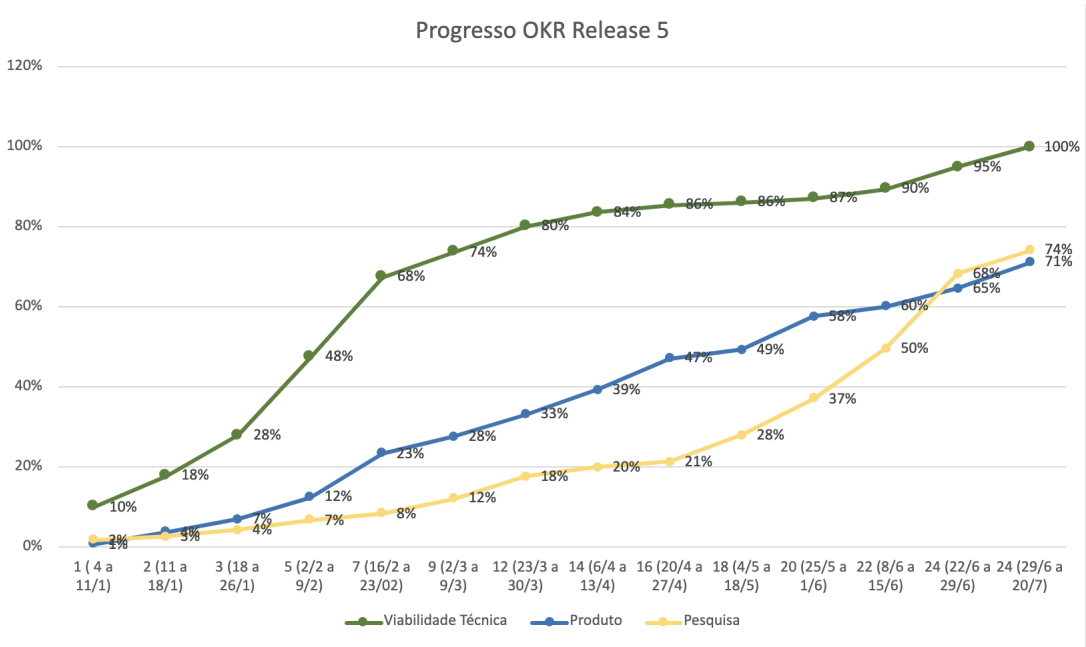



Figura 3: Evolução das atividades do OKR a cada duas *sprints*

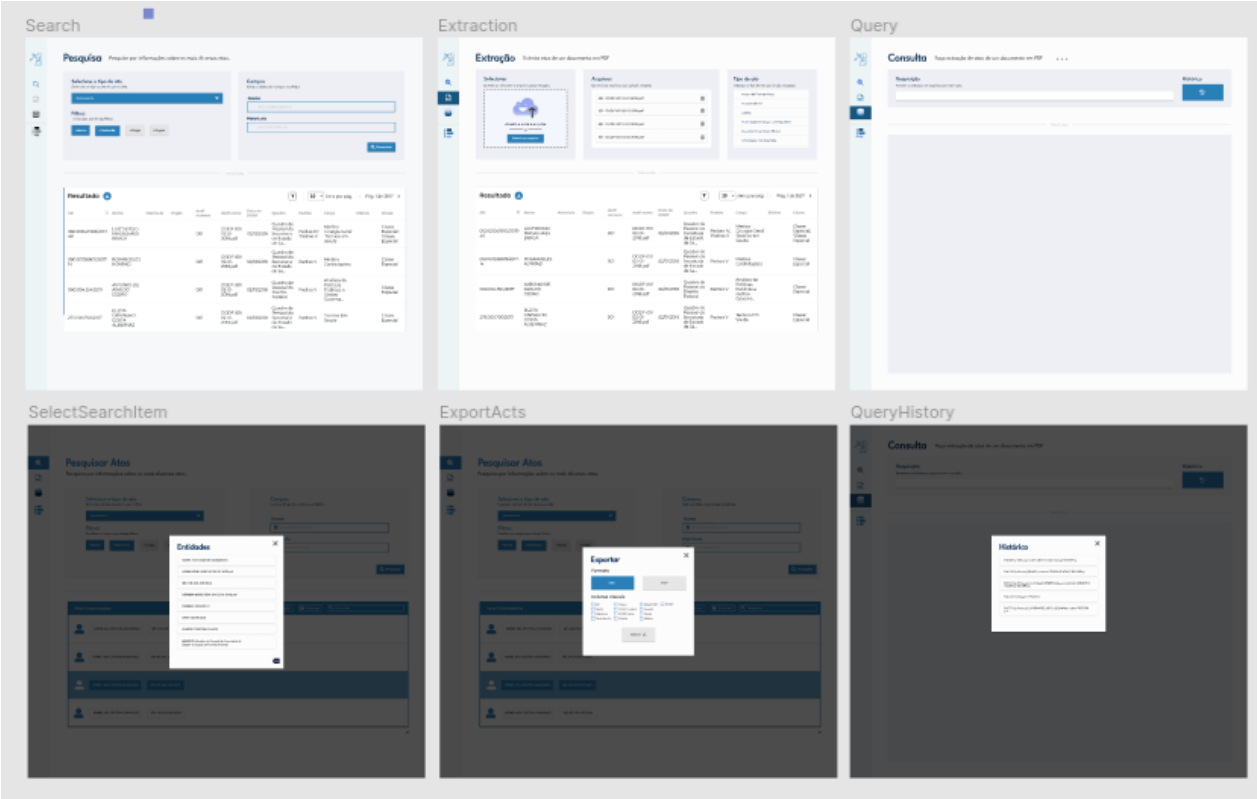



Figura 4: Telas do protótipo

É importante ressaltar que o *Knedash* faz uso das seguintes bibliotecas desenvolvidas nesse projeto: (*i*) DODFMiner[3] para a extração de atos e entidades de novos documentos PDFs; e (*ii*) DODFMinerAPI[4] para geração do Banco de Dados com as entidades estruturadas de Atos de Pessoal e Licitações e Contratos dos últimos 10 anos de DODFs.

## 3.1 Prototipação

Para o desenvolvimento da ferramenta, um ciclo de prototipação de alta fidelidade foi realizado com o objetivo de reimplementar a identidade visual do projeto e a estética da página, padronizar estruturas visuais e estudar as interações do usuário. Um protótipo de alta fidelidade deve se aproximar ao máximo dos aspectos visuais e funcionais do produto final, incluindo o conteúdo, fluxo de navegação e interações. São muito utilizados para testes e validação com usuário. Com o auxílio da ferramenta Figma[5] pode-se elaborar o protótipo [24].

## 3.2 Componentização

Após a prototipação e componentização das estruturas do *Knedash*, bem como o desenho de uma estrutura e interface, decidiu-se utilizar bibliotecas de componentes que fossem robustas e de fácil manutenção. Foi selecionado o ChakraUI[6], tendo em vista sua versatilidade e ampla utilização no mercado, para implementar todos os componentes prototipados. A ferramenta também foi dividida em páginas, utilizando o roteamento do Next.js[7], um framework focado em aplicar renderização *server-side* no React[8], uma bliboteca focada em aplicar renderização *client-side*[2]. Com isso, as responsabilidades dos contextos foram separadas, tornando a página mais responsiva.

## 3.3 React e Next.js

Enquanto o *React* é uma biblioteca *JavaScript* para desenvolvimento *client-side* que tem como um de seus objetivos facilitar a conexão entre diferentes partes de uma página por meio de componentes, o*Next.js* é um framework que permite a renderização *server-side* para aplicações web baseadas em *React* que permite evitar problemas de segurança e diminuir o tempo de carregamento.

A combinação desses dois componentes permite a renderização de parte das informações de forma local ou remota, melhorando o reaproveitamento de código e padronização de interface. Essas bibliotecas são flexíveis para a solução de problemas e para a construção de interfaces reutilizáveis, uma vez que cada um destes componentes pode ser manipulado de maneira distinta.

[3] `https://github.com/UnB-KnEDLe/DODFMiner`
[4] `https://github.com/UnB-KnEDLe/DODFMinerAPI`
[5] `www.figma.com`
[6] `https://chakra-ui.com/`
[7] `https://nextjs.org/`
[8] `https://reactjs.org/`

### 3.4 Substituição do Dash

Após alguns ciclos de ajuste, a antiga ferramenta entitulada *Dash* foi descontinuada e substituída pelo *Knedash*. Assim, a ferramenta foi otimizada para acelerar respostas.

## 4 Anotação de Textos

Na *Release 5*, foi iniciado um segundo processo de anotação no projeto KnEDLe para a construção de um *corpus* para os atos de contratos e licitações do Diário Oficial do Distrito Federal. A equipe **visualtm** ficou responsável por essa atividade, composta pelos discentes Artur Hugo, Alice de Lima e Matheus Stauffer. O objetivo desse *corpus* é ser utilizado para o treinamento de modelos de processamento de linguagem natural e de aprendizado para reconhecimento de entidades nomeadas e segmentação de textos a partir das publicações associadas a atos como avisos de abertura, revogação, suspensão e anulação de licitação, além dos extratos de contrato e de convênio. A construção desse *corpus* de contratos e licitações, denominado DODF Corpus II, foi realizada de maneira manual visando a obtenção de um *corpus* padrão ouro (*Gold Standard*) [27], uma vez que é importante a utilização de rótulos com alto grau de confiabilidade.

Como se sabe, um processo de anotação manual é desafiador e demorado, o que demanda um planejamento detalhado de suas etapas constituintes. O planejamento do processo de anotação foi iniciado em janeiro de 2022, em que foram selecionadas algumas edições do DODF para análise dos padrões das publicações associadas aos atos de contratos e licitações. Diversas reuniões ocorreram entre as diferentes equipes que estavam pesquisando sobre o treinamento de modelos para orientar a equipe **visualtm** na definição das entidades. Os especialistas do Tribunal de Contas do Distrito Federal (TCDF) foram consultados acerca da relevância das entidades escolhidas para serem rotuladas nas publicações. Como resultado, um tutorial de anotação [9] foi redigido para formalizar as caracteristicas dos atos considerados e suas entidades, como mostra a Tabela 1:

Tabela 1: Tipos de atos e o quantitativo de entidades por ato.

| Nome do Batch | Quantidade de entidades |
|---|---|
| Aviso de Abertura de Licitação | 16 |
| Aviso de Revogação/Anulação de Licitação | 7 |
| Aviso de Suspensão de Licitação | 6 |
| Extrato de Contrato | 15 |
| Extrato de Aditamento de Contrato | 12 |
| Extrato de Convênio | 18 |

Em seguida, estudantes dos cursos relacionados com a computação foram convidados para trabalhar como anotadores de forma voluntária. No total, 13 anotadores distintos participaram do processo de anotação, que foi viabilizado pela ferramenta de anotação NidoTat [10],

[9] Acessível em `https://github.com/UnB-KnEDLe/tutorial_anotacao_contratos_licitacoes`

[10] Acessível em `http://164.41.76.30/teamtat/`

adaptada da ferramenta TeamTat [7], para o processo de anotação, pela sua excelente capacidade de gerenciamento do processo de anotação e interface intuitiva para os anotadores. A Tabela 2 descreve como ficou dividida a quantidade total de documentos alocadas para anotação e criação do DODF Corpus II.

Tabela 2: Descrição dos batches de DODFs concebidos para o processo de anotação.

| Nome do Batch | Quantidade de documentos | Período das edições |
|---|---|---|
| Batch Experimental | 33 | 2018 e 2019 |
| Batch 1 | $X$ | Outubro/2021 à Março/2022 |
| Batch de Validação | $Y$ | Janeiro/2020 à Setembro/2021 |

O processo de anotação envolvendo os anotadores voluntários durou de fevereiro de 2022 até maio de 2022. Devido à disponibilidade dos anotadores ser menor do que o tempo necessário para a rotulação e a revisão das anotações, a parte de revisão ficou encarregada da equipe a partir de junho de 2022. Nesse momento, a equipe **visualtm** contava com apenas dois membros, fazendo com que o processo de revisão se prolongasse até o final da Release 5. Para fazer as revisões, os membros da equipe corrigiam as anotações feitas pelos voluntários com feedback dos especialistas do TCDF, que tiravam as dúvidas relacionadas aos casos mais difíceis de rotular, como as entidades "tipo de objeto" e "órgão".

# 5 Licitações e Contratos

Dando continuação aos trabalhos realizados na Release 4, dividimos os objetivos dessa entrega em duas atividades principais, sendo a Atividade A - Criação dos modelos de Extração de Entidades Nomeadas (NER), e a Atividade B - Criação da base de dados em Grafos.

Como visto na Figura 5, a Atividade A contém as tarefas de criação de modelos de Extração de Entidades Nomeadas (NER) para cada ato (i.e., Aviso de Licitação, Extrato de Contrato/Convênio, Extrato de Aditamento Contratual, Aviso de Suspensão de Licitação e Aviso de Revogação/Anulação de Licitação).

Já para atividade B, temos a tarefa de criação de uma estrutura de dados do tipo grafos para representar as entidades informações de cada um dos atos citados acima, pois estes serão incluídos no banco de dados Neo4j. Além disso, outras tarefas dessa atividade são: a inclusão dos atos publicados nos DODFs nos períodos entre julho de 2021 à julho de 2022, e a codificação de serviços REST para possibilitar o acesso dos atos dos DODFs por outras aplicações, no nosso caso, possibilitar que a ferramenta knedash (http://knedash.unb.br/) visualize os atos inseridos na base de dados.

Por fim, a conclusão das atividades A e B estão relacionados aos seguintes objetivos: Para a Atividade A, média de 80% de F1-score para cada modelo criado e para a Atividade B, uma API REST para acesso aos dados e documentada via Swagger.

## 5.1 Criação de modelos NER

Foram criados modelos NER para os atos de Aviso de Licitação, Extratos de Contrato/-Convênio e Aditamento Contratual. No entanto, os modelos NER para os atos de aviso de

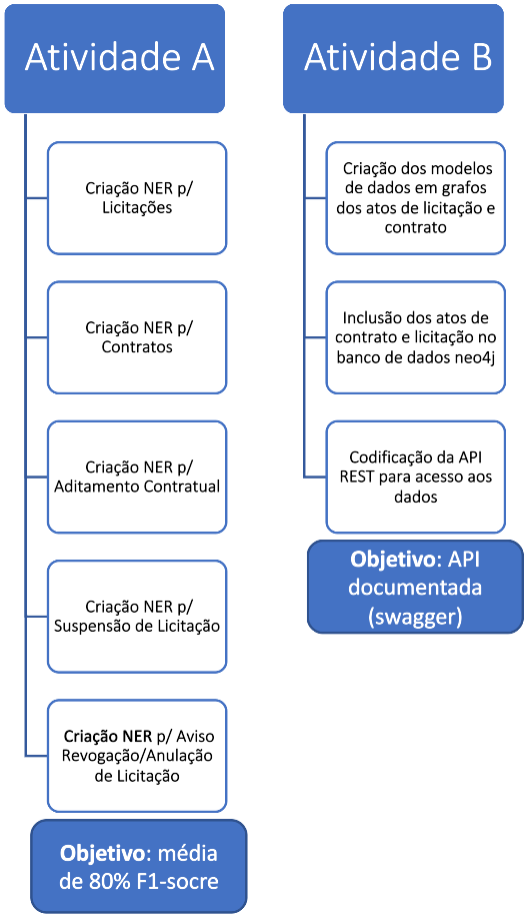



Figura 5: Atividades e objetivos da equipe de Contratos e Licitações para a *Release 5*

Suspensão e Revogação/Anulação de Licitação não foram criados, pois possuímos poucos dados anotados para realizar o treino do modelo.

Com relação aos algoritmos e técnicas utilizadas para a criação dos modelos, utilizamos o tradicional algoritmo Conditional Random Field (CRF) [9] para treinar o modelo NER dos atos de Abertura de Licitação. Para os atos de Extratos de Contrato/Convênio e Aditamento Contratual, utilizamos modelos baseados em redes neurais. No caso de Extratos de contrato e Convênio, utilizamos uma combinação de algoritmos CNN-CNN-LSTM [23], e para Aditamento Contratual, utilizamos CNN-biLSTM-CRF [12]. A Tabela 3 abaixo mostra os resultados obtidos com os modelos criados. Como pode ser visto, os valores médios de F1-score atingiram o objeto da atividade A (80% mínimo de F1-score), sendo 80%, 85% e 83% para os modelos NER de Licitação, Contratos/Convênios e Aditamento Contratual respectivamente. Vale salientar que o modelo não performou bem para algumas poucas entidades, por exemplo, a entidade "Órgão Licitante"(ato Licitação), obtivemos 28%, em "Vigência"(ato Contra./Convênios), 50%, e em "Objeto de Aditam. Contratual"(ato Aditam. Contratual), 51% de F1-score.

| **Abert. Licitação** | | **Contrat./Convênios** | | **Aditam. Contratual** | |
|---|---|---|---|---|---|
| **Entidade** | **F1-score** | **Entidade** | **F1-score** | **Entidade** | **F1-score** |
| Modalidade Lic. | 91% | Número de Contrato/Convênio | 98% | Num. Termo Aditivo | 96% |
| Num. Licitação | 86% | Objeto de Contrato/Convênio | 68% | Num. Contrato | 96% |
| Orgão Licitante | 28% | Ent. Contratante/Convenente | 81% | Orgão Contratante | 91% |
| Sistema de Compras | 78% | Ent. Contratada | 83% | Obj. Aditam. Contratual | 51% |
| Objeto da Licitação | 84% | Processo do GDF | 99% | | |
| Valor Estimado | 82% | Número de Licitação | 73% | | |
| Data de Abertura | 77% | Unidade Orçamentária | 95% | | |
| Processo | 76% | Programa de Trabalho | 95% | | |
| Nome do Responsável | 48% | Nota de Empenho | 88% | | |
| Código do Sis. Compras | 76% | Natureza de Despesa | 94% | | |
| | | Valor Total | 92% | | |
| | | Vigência | 50% | | |
| | | Data da Assinatura | 98% | | |
| Média F1-score (em porcentagem) | 80% | | 85% | | 83% |

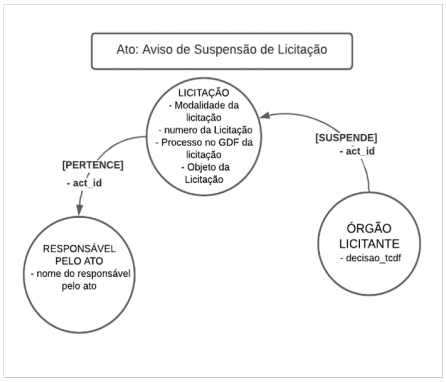
Tabela 3: Resultado dos modelos para cada ato e respectivas entidades

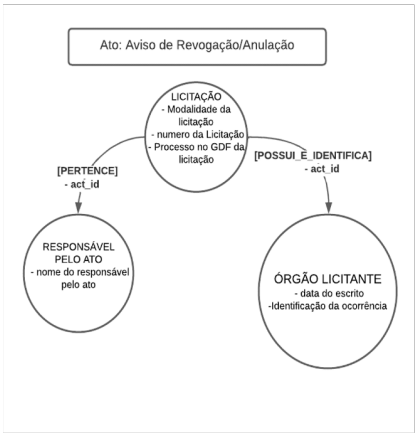
## 5.2 Criação de modelo de dados em grafos dos atos

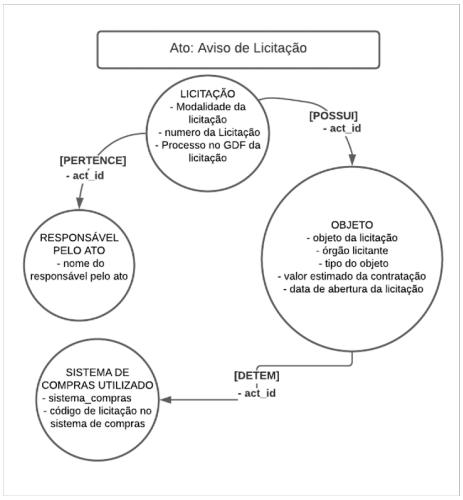
A modelagem de dados em grafos é utilizada para descrever informações e enfatizar relações entre entidades de um domínio. Devido a essa característica, escolhemos a estrutura de dados do tipo grafo para representar o domínio de conhecimento de atos de Abertura de Licitação e Extratos de Contratos/Convênios. Cada ato possui um rol de entidades, sendo que no momento foram destacadas 10 entidades para os atos licitação, 13 para os atos de Contrat./Convênios e 4 para Aditamento Contratual, como visto na Tabela 4.

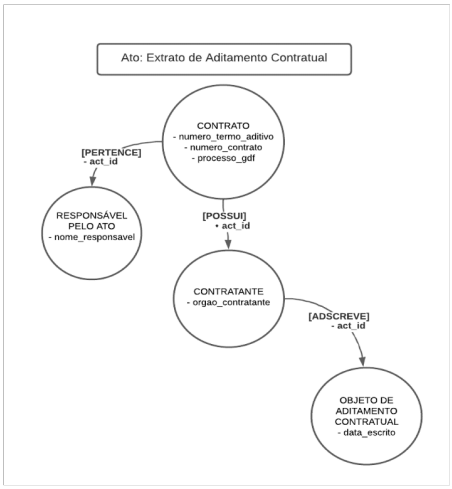
Algumas dessas entidades podem estar presentes e mais de um ato, como por exemplo, #Modalidade de Licitação, que está presente nos atos de Aviso de Licitação, Aviso de Revogação/Anulação e Aviso de Suspensão, e também que a entidade #Num. Licitação, que está presente em todos os atos, menos no ato de Extratos de Contrat./Convênio.

Dada as entidades e o entendimento de suas relações, foram criados modelos em grafos para cada um dos atos como mostra a Figura 6.



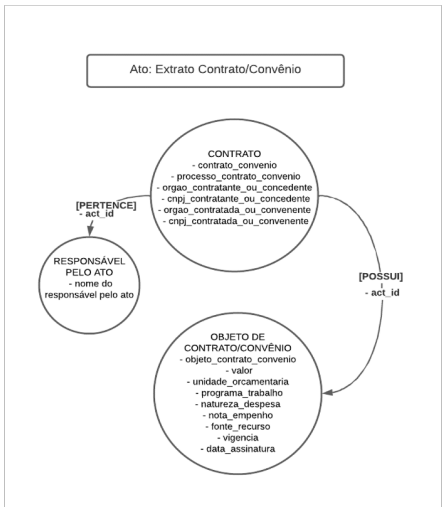




Figura 6: Modelagem em grafos dos atos de Aviso de Licitação, Extratos de Contrato/-Convênio, Aditamento Contratual, Aviso de Revogação/Anulação, Aviso de Suspensão de Licitação

| Tipo de Ato | Entidade |
|---|---|
| Aviso de Licitação | #Modalidade de Licitação, #Num. Licitação, #Orgão Licitante, #Sistema de Compras, #Objeto da Licitação, #Valor Estimado, #Data de Abertura, #Processo, #Nome do Responsável, #Código do Sistema de Compras |
| Extratos de Contrato e Convênio | #Número de Contratos/Convênio, #Objeto de Contrato/Convênio, #Ent. Contratante/Convenente, #Ent. Contratada, #Processo do GDF, #Número de Licitação, #Unidade Orçamentária, #Programa de Trabalho, #Nota de Empenho, #Natureza de Despesa, #Valor Total, #Data da Assinatura |
| Aditamento Contratual | #Objeto do Aditivo, #Num. do Contrato, #Num. do Aditivo, #Contratante, #Processo, #Nome do Responsável, #Data Escrito |
| Aviso de Revogação/Anulação | #Modalidade da Licitação, #Número da Licitação, #Processo no GDF da Licitação, #Nome do responsável, #Data do escrito, #Identificação da ocorrência |
| Aviso de Suspensão de Licitação | #Modalidade da Licitação, #Número da Licitação, #Processo no GDF da Licitação, #Nome do responsável, #Objeto da Licitação, #decisão TCDF |

Tabela 4: Entidades para cada tipo de ato

## 5.3 Inserção dos atos dos DODFs de julho/2021 à julho/2022 no banco de dados neo4j

As etapas para a realização dessa tarefa foram:

1. Coleta dos aquivos .json referentes aos DODFs do período jul/2021 à jul/2022.
2. Filtagem dos atos (selecionar somente Abertura de Licitação, Extrato de Contrato e Convênio e Aditamento Contratual).
3. Extração das entidades dos atos utilizando os modelos NER criados (ver seção 5.1)
4. Inserção dos atos conforme modelo apresentado na seção 5.2

No momento, os dados foram inseridos no banco de dados Neo4j no servidor na UnB. Essas dados são acessados pela aplicação knedash (`http://knedash.unb.br/knedash`)

## 5.4 Criação de API REST para acesso aos atos

Para realizar o acesso as informações de atos inseridas no banco de dados Neo4j, foi necessária a criação de uma API serviços do tipo REST. Essa API faz parte do projeto DODFminerAPI que encontra-se no repositório: `https://github.com/UnB-KnEDLe/DODFMinerAPI`.

Foram criados os *endpoints* para os seguintes atos: Abertura de Licitação, Extrato de Contrato/Convênio e Aditamento Contratual. Os *endpoints* podem ser acessados através do link `http://knedash.unb.br/dodfminner/api/swagger-ui/`.

## 5.5 Outras Atividades Realizadas

- **Criação de funcionalidade de download para DODFminer:** Os modelos recém-treinados que utilizam técnicas de redes neurais e aprendizado profundo são mais complexos, pois usam algoritmos mais avançados e até mesmo *embeddings* pré-treinados. Essa maior complexidade reflete no tamanho dos modelos, que passou de 30 MB para mais de 700 MB. O projeto DODFMiner estava lidando com o armazenamento de modelos através do projeto no GitHub, mas com modelos pesados essa solução não era mais razoável. Portanto, foi necessária a criação de uma funcionalidade de que fizesse o download de modelos criados e pré-treinados para dentro do projeto DODFminer.

- **Criação de um modelo de Autômato Determinístico Finito (DFA) para segmentação de atos:** Foi feito o uso de Expressões Regulares (REGEX) para a criação dos DFAs para a detecção de atos de Abertura, Suspensão, Revogação e Resultado de Licitação. Essa tarefa foi necessária para que pudéssemos capturar os atos de dentro de um arquivo .txt que foi gerado através de um processamento do tipo OCR realizado em um arquivo .pdf do um DODF.

# 6 Aprendizado Ativo e *Human-in-the-Loop* para NER

Métodos estado-da-arte para reconhecimento de entidades nomeadas (NER) são baseados em modelos de linguagem neural profunda [15], como BERT [4]. Embora tais modelos obtenham resultados promissores, eles dependem de um grande conjunto de treinamento com dados rotulados, muitas vezes limitados na prática, principalmente em aplicações específicas de domínio. Assim, há uma necessidade crescente de coleta contínua de dados e abordagens de rotulagem para atualizar e acelerar o treinamento do modelo NER [29].

O Human-in-the-Loop (HITL) é um conjunto de estratégias para incorporar o conhecimento e a experiência humanos para aumentar a precisão de um modelo de aprendizado de máquina e, ao mesmo tempo, atingir uma certa precisão de destino para um modelo mais rápido [28, 14]. Recentemente, o HITL foi aplicado com sucesso para lidar com seleção de dados, rotulagem e aceleração do treinamento de modelos para tarefas NER [14]. Em particular, um passo crucial no HITL é a técnica de amostragem de dados para interação humana. Técnicas de Aprendizagem Ativa (AL - Active Learning) são exploradas nesta etapa, com o objetivo de selecionar o melhor subconjunto de dados [20]. Em particular, AL são métodos de aprendizado de máquina que consultam interativamente humanos ou bases de conhecimento para rotular dados, geralmente minimizando o número de consultas de acordo com alguns critérios. Muitos critérios foram propostos nas últimas décadas [1, 20], como amostragem de incerteza [10], amostragem de densidade [21] e consulta por amostragem de comitê [22]. Este último critério é especialmente útil, uma vez que humanos podem interagir com diferentes modelos de comitê (interação homem-modelo), e os modelos de um comitê podem interagir entre si (interação modelo-modelo).

Neste projeto, o objetivo é investigar e estender uma técnica denominada Query by Committee (QBC), proposta originalmente por Seung, Opper e Sompolinksy [22] para aprendizado ativo de tarefas de classificação. Inicialmente, diferentes modelos são treinados a partir do mesmo conjunto de dados rotulados. Em seguida, os membros do comitê podem votar em

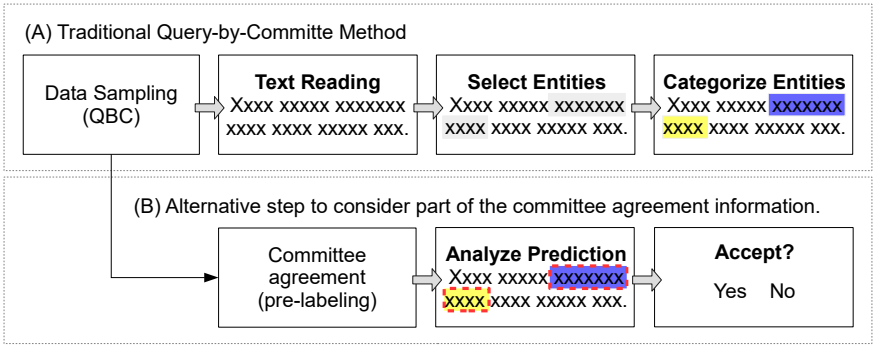



Figura 7: Ilustração do processo de anotação de dados para tarefas de Reconhecimento de Entidades Nomeadas.

cada instância (dados não rotulados) para definir a categoria da entidade. A ideia básica do QBC é que as instâncias mais informativas são aquelas em que a maioria dos modelos discorda. A consulta retorna tais instâncias para que os humanos possam incorporar sua experiência e conhecimento. Estudos anteriores demonstraram formalmente que QBC generaliza o aprendizado incorporando diversidade no conjunto de treinamento de instâncias que causam discordância máxima entre o comitê [22, 13], ou seja, consultando instâncias não rotuladas de regiões controversas do espaço de entrada [8] .

Um dos principais resultados deste projeto é uma extensão do método QBC para cenários envolvendo NER. Enquanto o QBC tradicional se concentra apenas no desacordo, também foram exploradas as informações de concordância do comitê, na premissa de que são úteis para apoiar a rotulagem entidades nomeadas quando assistidas por humanos. A Figura 7 ilustra o relaxamento QBC proposto. Na Figura 7(A), apresentamos o processo de rotulagem tradicional, no qual QBC seleciona uma instância textual informativa para rotulação. Nesse sentido, os humanos realizam um custoso processo de análise de texto, envolvendo a leitura do documento, bem como a identificação e categorização das entidades nomeadas. Esses dados rotulados são inseridos no conjunto de treinamento para a próxima iteração de atualização dos modelos NER. Observe que, nesse caso, as informações com algum nível de concordância são descartadas e não são usadas na interface de rotulagem. Por outro lado, na Figura 7(B), optamos por não descartar instâncias com algum nível de concordância e utilizá-las para pré-rotulagem do documento textual. Os usuários podem aprovar ou desaprovar a sugestão de instância rotulada, ou seja, uma interação binária que é muito menos onerosa no processo de anotação. Alguns estudos afirmam que as respostas binárias são a melhor estratégia para controle de qualidade e devem ser escolhidas sempre que possível no processo (ver [14] para uma discussão de interfaces para aprendizagem ativa.

A seguir é apresentado um resumo das atividades desenvolvidas no período de Abril a Setembro de 2022:

**Extensão do método QBC para AL-NER** : o método QTC tradicional explora principalmente desacordos entre modelos para seleção de instâncias para anotação. No entanto, introduzimos uma heurística para incluir algumas instâncias que possuem um nível de concordância para acelerar o processo de rotulagem sem reduzir a precisão dos modelos NER treinados. Ressaltamos que incorporar sugestões de anotações a partir

de informações de concordância visa pré-anotar dados e reduzir o custo de interação humana, principalmente com apoio de interface. Uma visão geral da proposta está ilustrada na Figura 8.

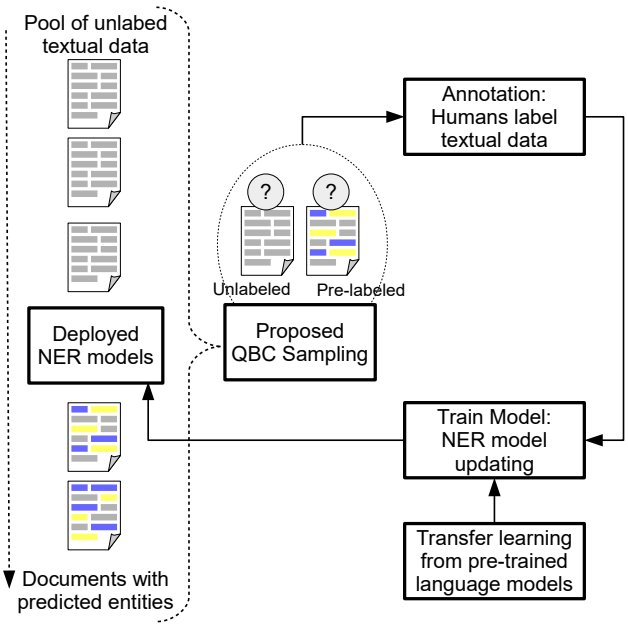



Figura 8: Um modelo abstrato do processo HITL e Aprendizado Ativo para tarefas de Reconhecimento de Entidades Nomeadas. Adaptado de [14].

**Redução Antecipada da Propagação de Erros** : modelos podem concordar e estar incorretos sobre a categoria da entidade nomeada. Esta situação não é tratada diretamente pelo QBC original. Tais instâncias também são informativas, e nosso método permite que humanos corrijam o erro antecipadamente (por exemplo, reprovação de uma previsão), evitando assim a propagação desse erro para as próximas iterações.

**Usando Previsões Parcialmente Corretas** : pequenas mudanças nas palavras de uma entidade podem ser consideradas um erro pelos modelos NER durante a avaliação do modelo, impactando negativamente a estratégia tradicional de desacordo QBC. Nosso método explora previsões parcialmente corretas no relaxamento QBC, para que os humanos possam acelerar a rotulagem de dados com ajustes simples na entidade pré-rotulada (por exemplo, simples adição ou remoção de um token de uma entidade pré-rotulada).

Uma avaliação experimental preliminar foi utilizado um conjunto de dados extraído de Diário Oficial, fornecido pelo projeto KnEDLe, em que as entidades nomeadas envolvem nomes de pessoas, cargos e respectivos cargos para agências governamentais, conforme descrito na Tabela 5.

Foram utilizados cinco modelos NER com os seguintes modelos pré-treinados.

- BERTimbau (base-cased) [25]: Modelo BERT pré-treinado com corpus textual português.

Tabela 5: Overview of the three datasets used in the experimental evaluation

| Dataset | #Train Docs | #Test Docs | #Named Entities | Train | Test |
|---|---|---|---|---|---|
| Government gazettes | 2113 | 211 | cargo comissionado | 1837 | 209 |
| | | | cargo efetivo | 580 | 70 |
| | | | hierarquia lotacao | 1805 | 209 |
| | | | matricula | 745 | 94 |
| | | | matricula SIAPE | 82 | 10 |
| | | | nome | 1836 | 211 |
| | | | orgao | 1824 | 210 |
| | | | simbolo | 1786 | 204 |

- DistilBERT-PT (base)[11]: Uma versão mais leve (destilada) do BERT, pré-treinado com corpus textual português.
- NER-News-Portuguese (base-case)[12]: Um modelo BERTimbau ajustado para tarefas NER em notícias portuguesas.
- Lener-BR (base-cased)[13]: Um modelo BERTimbau ajustado para textos legislativos portugueses.
- Wikineural-Multilingual [26]: Um modelo BERT multilíngue pré-treinado com textos de diferentes idiomas da Wikipedia.

Foi simulado um processo de anotação e aprendizado ativo de acordo com as seguintes etapas:

1. Os modelos são inicialmente treinados considerando uma pequena porcentagem $p = 0.2$ (20%) do conjunto de treinamento.
2. Para cada iteração, o método QBC deve retornar um número de documentos para anotação humana. O tamanho da consulta usado foi de 250. Usamos cinco iterações para avaliar o impacto do processo de AL.

Na Tabela 6 é apresentada uma visão geral dos resultados experimentais. São comparados os desempenhos (F1-Score) de reconhecimento de entidades nomeadas do modelo inicial, do QBC tradicional por desacordo e do QBC ajustado para este projeto. É possível observar que o QBC proposto é competitivo, mantendo o desempenho do QBC original, mas reduzindo o esforço humano de anotação em aproximadamente 8%. Para calcular a redução do esforço, foram adaptados parâmetros propostos em [3]. No caso desta presente proposta, cada instância pré-rotulada corretamente reduz o esforço de anotar uma instância específica em 75%. Já instâncias pré-rotuladas parcialmente corretas reduzem o esforço do usuário em 25% na tarefa de anotação. Essas medidas são aplicadas a todas as consultas QBC propostas durante as 5 iterações.

A próxima etapa do projeto é integrar o QBC proposto uma ferramenta computacional para apoiar de ponta-a-ponta humanos na tarefa de anotação de dados para NER. Essa

[11] `https://huggingface.co/adalbertojunior/distilbert-portuguese-cased`
[12] `https://huggingface.co/monilouise/ner_news_portuguese`
[13] `https://huggingface.co/pierreguillou/bert-base-cased-pt-lenerbr`

Tabela 6: F1-Score de cada modelo inicial e dos respectivos modelos após cinco iterações de aprendizado ativo via QBC.

| | Government gazettes | | |
|---|---|---|---|
| Model | **Initial Model** | **QBC (Disagreement)** | **QBC (Ours)** |
| NER-BERTimbau | 0.80 | 0.90 | 0.87 |
| NER-DistilBERT-PT | 0.73 | 0.87 | 0.87 |
| NER-LenerBR | 0.80 | 0.86 | 0.87 |
| NER-News | 0.74 | 0.86 | 0.87 |
| NER-Wikineural | 0.73 | 0.86 | 0.85 |
| Labeling Effort Reduction | — | — | 8% |

ferramenta é baseada no LabelStudio[14], que conta com uma interface web de anotação de entidades nomeadas em textos. A integração permite que sugestões de anotações possam ser apresentadas para os usuários, conforme respostas obtidas pelo QBC proposto e avaliado neste projeto. Assim, como resultado esperado, será obtida uma base de dados anotados para tarefas NER com menor esforço.

Uma versão preliminar da ferramenta foi desenvolvida e está em fase de testes, disponível em `https://github.com/UnB-KnEDLe/active_learning_tool`.

# 7 Pesquisa

Nesta seção serão relatados as metas de pesquisa alcançadas na *Release 5* do projeto KnEDLe. Para isso, essa seção será organizada entre os relatos das pesquisas dos atuais alunos de mestrado.

No projeto KnEDLe estão atuando dois alunos de mestrado. Uma aluna, Lucélia Vieira Mota, já apresentou o texto exigido no exame de qualificação. A pesquisa da aluna Lucélia está relacionada com a criação de bases de dados rotuladas em português a partir da abordagem de supervisão fraca. O outro aluno, Micael Filipe Ribeiro de Lima, agendou a qualificação para novembro de 2023. O trabalho do aluno Micael está relacionado com o problema de encontrar segmentos de textos dentro do Diário Oficial do Distrito Federal. A seguir, serão descritos os resumos das pesquisas desses dois alunos.

## 7.1 Criação de base de dados rotulados em português a partir da abordagem de supervisão fraca (aluna Lucélia Vieira Mota)

A rotulagem de dados de treinamento tornou-se um dos principais obstáculos ao uso do aprendizado de máquina. Entre vários paradigmas de rotulagem dos dados, a supervisão fraca tem mostrado como uma oportunidade para aliviar o gargalo da rotulagem manual, pois a partir da supervisão podemos sintetizar programaticamente o treinamento de rótulos de múltiplas fontes geradas por supervisão potencialmente ruidosa.

A dissertação apresenta pela aluna apresentou experimentos sobre uma das abordagens de aplicação da supervisão fraca. Em particular, foi realizada uma breve revisão bibliográfica sobre a base teórica que fundamenta o uso dessa abordagem e descreve de forma geral, um

[14] `https://labelstud.io/`

fluxo de trabalho de aprendizado e rotulação dentro problema de reconhecimento de entidade nomeada a partir da supervisão fraca. Por fim, realizou-se experimentos para avaliar os ganhos de se utilizar essa abordagem para auxliar na rotulação de bases dentro do contexto da Administração Pública no Brasil, e assim, inspirar futuras direções de pesquisa no campo.

Com isso, o objetivo da pesquisa de mestrado é aplicar uma abordagem alternativa a tarefa de rotulação de dados apoiada na técnica de supervisão fraca para gerar uma base dados rotulada com uma acurácia aceitável (aceitável no sentido de agregar valor ao processo de rotulação em detrimento ao manual). Trata-se de uma abordagem que utiliza funções de rotulagem de dados que automaticamente anotam documentos com rótulos de entidades nomeadas.

A fim de nortear o alcance do objetivo desta pesquisa foram definidos os seguintes objetivos específicos:

1. Investigar algoritmos de geração de entidades nomeadas (NER) com uso da técnica de supervisão fraca.

2. Treinar os modelos em bases de dados já rotuladas e avaliar o reaproveitamento desses modelos na construção das funções de rótulos;

3. Aplicar outras funções de rótulos para abarcar as entidades não contempladas pelas bases de dados já rotuladas.

4. Diminuir o ruído dos rótulos por meio do uso de um modelo generativo.

5. Criar base de dados rotuladas para a tarefa de NER no contexto de contratação pública;

A partir dos experimentos inicias (os resultados estão descritos no documento da qualificação), foi visto que algoritmos de NER juntamente com as abordagens de supervisão fraca têm o potencial de resolver alguns dos problemas atualmente pouco abordados na literatura de rotulação de entidades. Nesse sentido, os próximos trabalhos se concentrarão no desenvolvimento dos novos experimentos e na compilação de seus resultados no documento final da dissertação.

## 7.2 Segmentação de texto do Diário Oficial do Distrito Federal

O processamento (PLN do inglês Natural Language Processing) envolve tarefas de rotulagem de sequências, como marcação de partes do discurso *(PoS tagging)*, reconhecimento de entidades nomeadas (NER do inglês Named Entity Recognition) e Extração de informação(IE). [16]. A etapa de rotulagem é uma das principais etapas do PLN e necessita de conjuntos de dados rotulados para treinar modelos preditivos. Entretanto, a rotulagem manual é uma tarefa árdua e custosa, em termos de tempo e de recursos humanos. Além de exigir dos anotadores, uma expertise sobre o assunto e conhecimento dos termos técnicos a serem rotulados.

Nesse sentido vários pesquisadores tem proposto abordagens a fim resolver o problema de NER e automatizar a geração desses rótulos, tais como HMMs (do inglês Hidden Markov Model) [18] LSTM (do inglês Long Short-Term Memory) [6].

O Reconhecimento de Entidades Nomeadas (NER) constitui um componente central em muitos pipelines de PNL e é empregado em uma ampla gama de aplicações, como por exemplo, extração de informações [19]. Ou seja, dado um documento, o objetivo do NER é identificar e classificar *tokens* referentes a uma entidade pertencente a categorias pré especificadas, como pessoas, organizações ou localizações geográficas. [11].

Os métodos de segmentação de texto são úteis para identificar o diferentes tópicos que aparecem em um documento. O objetivo da segmentação de texto é dividir um texto em segmentos homogêneos, de modo que cada segmento corresponda a um determinado assunto. [5]

A pesquisa de mestrado realizada pelo aluno busca combinar os benefícios da segmentação de texto e do problema de NER ao examinar a performance de algoritmos de segmentação para extrair várias classes existentes em um documento oficial.

### 7.2.1 Motivação

Atualmente, já existem vários modelos NER treinados, entretanto para sua completa utilização é preciso encontrar um modelo que atenda ao domínio e o idioma que se pretende rotular. Nesse sentido, quando se trabalha com domínios restritos, como domínios ligados às comunicações oficiais, por exemplo, há uma limitação para reuso desses modelos existentes, sendo necessário criar novos modelos com regras bem definidas que sejam capazes de extrair as sequências de maneira mais assertiva possível, o que exige conhecimento especializado para realizar a classificação.

No contexto em questão, o domínio a ser atendido está relacionado ao interesse do Governo do Distrito Federal (GDF) em trabalhar com dados em português. Os dados estão relacionados aos atos públicos praticados pelo GDF e que são publicados diariamente no diário oficial do DF [15] .

### 7.2.2 Descrição do Problema

Explicar o problema (ponto de vista computacional - desafio computacional) Comentar das lacunas e desafios desse problema Descrever as hipóteses

### 7.2.3 Objetivo

Nesta pesquisa de mestrado será abordado um dos desafios enfrentados pela tarefa de NER para rotulação massiva de dados. Dessa forma, a questão que deve ser respondida por esta dissertação é:

1. Investigar técnicas de tagueamento de sequencias para segmentação texto aplicadas à publicações oficiais; tais técnicas tragam resultados efetivos com alto valor de acurácia

Para responder esta questão, foi elaborada uma hipótese com uma potencial solução que será apresentada na seção **??**.

A fim de nortear o alcance do objetivo desta pesquisa foram definidos os seguintes objetivos específicos:

[15] Disponível em `https://www.dodf.df.gov.br`

1. Descrever o problema de segmentação de texto;
2. Identificar algoritmos correlatos ao problema de segmentação de texto;
3. Propor e implementar uma abordagem para segmentação de texto no contexto do Diário Oficial do Distrito Federal;
4. Avaliar experimentalmente a abordagem proposta.

Por fim, é importante salientar que esta pesquisa ainda está em desenvolvimento. O aluno está se preparando para o exame de qualificação.

# 8 Considerações Finais

No relatório foram elencados os resultados produzidos durante a quinta etapa do projeto. Nesta etapa foram identificadas novas necessidades e funcionalidades para as ferramentas desenvolvidas. Assim, o projeto mantém o cronograma previsto. Porém, deve-se destacar que, apesar da evolução do OKR (*Objective and Key Results*) indicar mais de 75 % de atividades completas, as atividades restantes podem ser acumuladas no próximo semestre, que é justamente o último semestre do projeto. Isso pode impactar na correta entrega das ferramentas. Por isso, a equipe do projeto elaborou um plano de prorrogação do projeto para mais quatro meses. Esse prorrogamento propiciar o prazo adequado para a elaboração da versão final da biblioteca, manuais para os usuários e instrução e treinamento para o uso do sistema desenvolvido.

# 9 Equipe

Além dos autores deste relatório, o trabalho realizado no referido período contou com as participações dos seguintes bolsistas: Alice Lima, Arthur Hugo Cunha, Daniel de Sousa Oliveira Melo Veras,Felipe Xavier Barbosa da Silva, Gabriel Mendes Ciriatico Guimaraes, Gabriel da Silva Corvino Nogueira, Ian Filipe Pontes Ferreira, Jonatas Gomes Barbosa da Silva, Larissa Santana de Freitas Andrade, Luis Fernando Ferreira, Maicon Rodrigues Queiroz, Rafael Amaral Soares, Thais Rebouças Araujo, Vitor Vasconcelos de Oliveira, Manuela Matos Correia de Souza, Matheus Stauffer Viana de Oliveira e Tatiana Franco Pereira. Também contamos com o trabalho dos alunos de pós-graduação Micael Filipe Ribeiro de Lima e Lucelia Vieira.

# Referências


[1] Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. Active learning: A survey. In *Data Classification*, pages 599–634. Chapman and Hall/CRC, 2014.

[2] Sanchit Aggarwal. Modern web-development using reactjs. *International Journal of Recent Research Aspects*, 5(1):133–137, 2018.

[3] Omar Alonso. Algorithms and techniques for quality control. In *The Practice of Crowdsourcing*, pages 53–63. Springer, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[5] Pavlina Fragkou. Text segmentation using named entity recognition and co-reference resolution in english and greek texts. *CoRR*, abs/1610.09226, 2016.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11, 2020.

[8] Punit Kumar and Atul Gupta. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35(4):913–945, 2020.

[9] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[10] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[11] Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. Named entity recognition without labelled data: A weak supervision approach, 2020.

[12] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[13] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.

[14] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[15] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39, 2021.

[16] An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[17] Paul R Niven and Ben Lamorte. *Objectives and key results: Driving focus, alignment, and engagement with OKRs*. John Wiley & Sons, 2016.

[18] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[19] J. Raiman and O. Raiman. O. deeptype: Multilingual entity linking by neural type system evolution., Apr 2018.

[20] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

[21] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

[22] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[23] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. arXiv, 2017.

[24] Bruno Santana da Silva Simone Diniz Junqueiro Barbosa. *Interação Humano-Computador*, volume 1. Editora Campus, 2010.

[25] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer, 2020.

[26] Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, 2021.

[27] Lars Wissler, Mohammed Almashraee, Dagmar Monett Díaz, and Adrian Paschke. The gold standard in corpus annotation. In *IEEE Germany Student Conference*, pages 1–4, 2014.

[28] Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

[29] Shanshan Zhang, Lihong He, Eduard Dragut, and Slobodan Vucetic. How to invest my time: Lessons from human-in-the-loop entity extraction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2305–2313, 2019.