Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Inteligência Artificial no MPF: Uma Solução Baseada em IA para Pseudonimização de Dados Pessoais

Marcelo Anselmo de Souza Filho

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Bruno César Ribas

Brasília
2025

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Ai

Anselmo de Souza Filho, Marcelo
Inteligência Artificial no MPF: Uma Solução Baseada em IA para Pseudonimização de Dados Pessoais / Marcelo Anselmo de Souza Filho; orientador Bruno César Ribas. Brasília, 2025.
73 p.

Dissertação(Mestrado Profissional em Computação Aplicada) Universidade de Brasília, 2025.

1. Pseudonimização de Dado. 2. Reconhecimento de Entidade Nomeada. 3. Privacidade de Informação. 4. Textos Jurídicos. 5. LGPD. I. César Ribas, Bruno, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Inteligência Artificial no MPF: Uma Solução Baseada em IA para Pseudonimização de Dados Pessoais

Marcelo Anselmo de Souza Filho

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Bruno César Ribas (Orientador)
CIC/UnB

Prof. Dr. Thiago Paulo Faleiros
CIC/UnB

Prof. Dr. Willian Zalewski
UNILA

Prof.a Dr.a Edna Canedo
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 27 de junho de 2025

Dedicatória

À minha amada esposa, companheira incansável, que esteve ao meu lado em cada instante difícil, compartilhando minhas angústias com coragem e me incentivando, com amor e firmeza, a nunca desistir dos meus sonhos. Aos meus queridos filhos, que, mesmo tão jovens, demonstraram uma maturidade admirável ao compreenderem minha ausência em momentos preciosos da vida familiar. A todos os meus familiares e amigos, cuja presença, apoio e palavras de encorajamento foram fundamentais em minha caminhada.

Agradecimentos

Ao meu orientador, pela orientação valiosa, paciência e compreensão ao longo de toda essa jornada desafiadora. Sua condução me mostrou caminhos antes desconhecidos, que contribuíram profundamente para o meu crescimento pessoal e profissional.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

A evolução tecnológica tem transformado a sociedade, impactando o modo de vida das pessoas e o funcionamento das organizações. Desde a Revolução Industrial até a era da informação, essas mudanças moldaram atividades cotidianas e a estrutura institucional. O Ministério Público, como defensor dos direitos constitucionais, também tem sido influenciado por essas inovações. Diariamente no Ministério Público Federal (MPF) são inseridos milhares de registros por mais de 10 mil usuários em todo o país. Uma etapa relevante é manter os dados dos cidadãos seguros e protegidos. A Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil, em vigor desde 2020, estabelece diretrizes para a coleta, armazenamento e tratamento de dados pessoais, visando proteger a privacidade e a segurança dos cidadãos.

Atualmente, a pseudonimização manual no MPF é complexa e sujeita a erros. Técnicas automatizadas utilizando IA são fundamentais para eficiência e segurança. Portanto este trabalho visa apresentar o *LEGAL-BERT-LGPD*, um modelo baseado em BERT especializado em tarefas de pseudonimização de dados pessoais em conformidade com a LGPD. Partindo da arquitetura BERT, a abordagem proposta concentra-se na identificação e substituição de informações pessoais brasileiras em textos jurídicos por meio de tarefas de Reconhecimento de Entidades Nomeadas (Named Entity Recognition - NER). O estudo compara a performance do BERT proposto (GPU e CPU) com quatro grandes modelos de linguagem (LLMs): executados localmente, *DeepSeek-R1 8B* e *DeepSeek-R1 32B*, e em nuvem, *GPT-4o-mini* e *GPT-4.1*. Os experimentos mostraram que o *LEGAL-BERT-LGPD* alcançou uma posição equilibrada entre os modelos avaliados, ficando apenas à frente do *DeepSeek-R1 8B*. Percebemos que nosso modelo, mesmo com poucos parâmetros, consegue competir com grandes modelos. Portanto, a escolha do modelo deve refletir a criticidade dos dados: LLMs oferecem ganhos de qualidade em contextos menos restritivos, enquanto o *LEGAL-BERT-LGPD* se destaca em cenários de alta sensibilidade à privacidade.

Palavras-chave: Pseudonimização de Dados; Reconhecimento de Entidade Nomeada; Privacidade de Informação; Textos Jurídicos; LGPD; Transformer;

Abstract

Technological evolution has been transforming society, impacting both people's lifestyles and the functioning of organizations. From the Industrial Revolution to the Information Age, these changes have shaped daily activities and institutional structures. The Federal Prosecution Service (Ministério Público Federal - MPF), as a defender of constitutional rights, has also been influenced by these innovations. Every day at the Federal Prosecution Service (MPF), thousands of records are entered by more than 10,000 users across the country. A key step is ensuring that citizens' data is kept secure and protected. Brazil's General Data Protection Law (LGPD), in force since 2020, establishes guidelines for the collection, storage, and processing of Personally Identifiable Information (PII), aiming to protect citizens' privacy and security.

Currently, manual pseudonymization at the MPF is complex and prone to errors. Automated techniques using AI are essential for ensuring both efficiency and security. Therefore, this work aims to present *LEGAL-BERT-LGPD*, a BERT-based model specialized in pseudonymization tasks for personal data in compliance with the LGPD. Based on the BERT architecture, the proposed approach focuses on identifying and replacing Brazilian personal information in legal texts through Named Entity Recognition (NER) tasks. The study compares the performance of the proposed BERT model (on GPU and CPU) with four large language models (LLMs): two running locally, *DeepSeek-R1 8B* and *DeepSeek-R1 32B*, and two cloud-based, *GPT-4o-mini* and *GPT-4.1*. Experiments showed that *LEGAL-BERT-LGPD* achieved a balanced position among the evaluated models, ranking just ahead of *DeepSeek-R1 8B*. We observed that our model, even with fewer parameters, is capable of competing with larger models. Therefore, model selection should reflect the criticality of the data: LLMs deliver quality gains in less restrictive contexts, while *LEGAL-BERT-LGPD* excels in highly privacy-sensitive scenarios.

Keywords: Data Pseudonymization; Named Entity Recognition; Information Privacy; Legal Texts; LGPD; Transformer;

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

No contexto atual, em que a tecnologia e o poder computacional avançam rapidamente, os dados pessoais se tornam um recurso extremamente valioso [5]. Empresas coletam essas informações através de interações na internet, como em sites e redes sociais, para criar perfis detalhados dos usuários. Esses perfis podem incluir desde hábitos de consumo até informações sensíveis, como condições de saúde e preferências pessoais. Embora essa prática permita a personalização da experiência na internet, também expõe os usuários a riscos de privacidade, como a venda ilegal desses dados em mercados clandestinos, possivelmente para fins ilícitos [6].

Diante desse cenário, a proteção de dados pessoais se torna fundamental, não apenas para preservar a privacidade individual, mas também para a segurança da sociedade como um todo. A resposta a esses desafios veio por meio de regulamentações e leis específicas de proteção de dados, como o Regulamento Geral de Proteção de Dados (GDPR) na União Europeia e a Lei Geral de Proteção de Dados - LGPD [1] (Lei nº 13.079, de 14 de agosto de 2018) no Brasil [7].

O Ministério Público Federal (MPF), segue atento às orientações da LGPD. A principal missão do MPF é assegurar o respeito aos direitos dos cidadãos, atuando na fiscalização e na exigência da aplicação das leis [8]. Em cumprimento à LGPD e em respeito ao dever de transparência, o MPF fornece a todos os cidadãos, cujos dados pessoais são processados pela instituição, informações sobre as situações em que realiza o tratamento de dados pessoais.

Atualmente, a tarefa de pseudonimizar informações em documentos no MPF é realizada de forma manual, um processo lento, oneroso e suscetível a falhas humanas. A utilização de ferramentas genéricas, como editores de PDF, não apenas consome um tempo considerável dos usuários, mas também não assegura que os dados sensíveis sejam eficientemente ocultados. A necessidade de uma solução tecnológica se torna evidente. O

[1] `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`

próprio MPF, em seu planejamento estratégico, incentiva o uso de Inteligência Artificial (IA) para otimizar processos internos. Além disso, a Resolução nº 281 do Conselho Nacional do Ministério Público (CNMP) endossa a pseudonimização como prática recomendada para a proteção de dados.

Contudo, o desenvolvimento de ferramentas de IA para o domínio jurídico brasileiro enfrenta um obstáculo significativo: a carência de conjuntos de dados anotados com informações pessoais formatadas segundo os padrões nacionais, como CPF e endereços. Essa falta de dados anotados impacta no treinamento de modelos NER que possam contribuir com as tarefas de pseudonimização. Já sobre os próprios modelos NER, temos o BERT, que trouxe um equilíbrio entre performance e custo computacional. Esse modelo captura o contexto de palavras em sentenças, resultando em uma exatidão muito maior para a tarefa de NER, apesar de ser mais pesado que métodos tradicionais, como o Regex. Técnicas mais modernas, como a Engenharia de Prompt, utilizam grandes modelos pré-treinados, como o GPT, para alcançar maior exatidão em NER. Apesar de mais eficientes, esses métodos costumam exigir maior capacidade computacional e custos adicionais devido à cobrança por uso de APIs, o que deve ser ponderado.

Portanto, tendo em vista o trabalho atual feito de forma manual no MPF e a quantidade massiva de processos com dados pessoais, este trabalho tem como finalidade elaborar uma solução que otimize a tarefa de pseudoanonimizar documentos jurídicos no MPF, acelerando o processo e reduzindo a possibilidade de erros humanos. Para isso, criaremos um modelo de IA que tenha um baixo custo computacional, conforme um conjunto de dados anotado, assegurando a proteção dos dados pessoais e a conformidade com a LGPD.

1.1 Objetivo

O novo modelo proposto será chamado de **LEGAL-BERT-LGPD** e fará parte de uma ferramenta desenvolvida internamente no MPF para pseudonimização. Esta ferramenta receberá um documento no formato PDF (Portable Document Format), realizará o reconhecimento de entidades nomeadas e habilitará a alteração do arquivo para a validação humana e, por fim, gerará um novo documento PDF pesquisável. Abaixo na Figura 1.1 temos um protótipo da ferramenta que será desenvolvida:

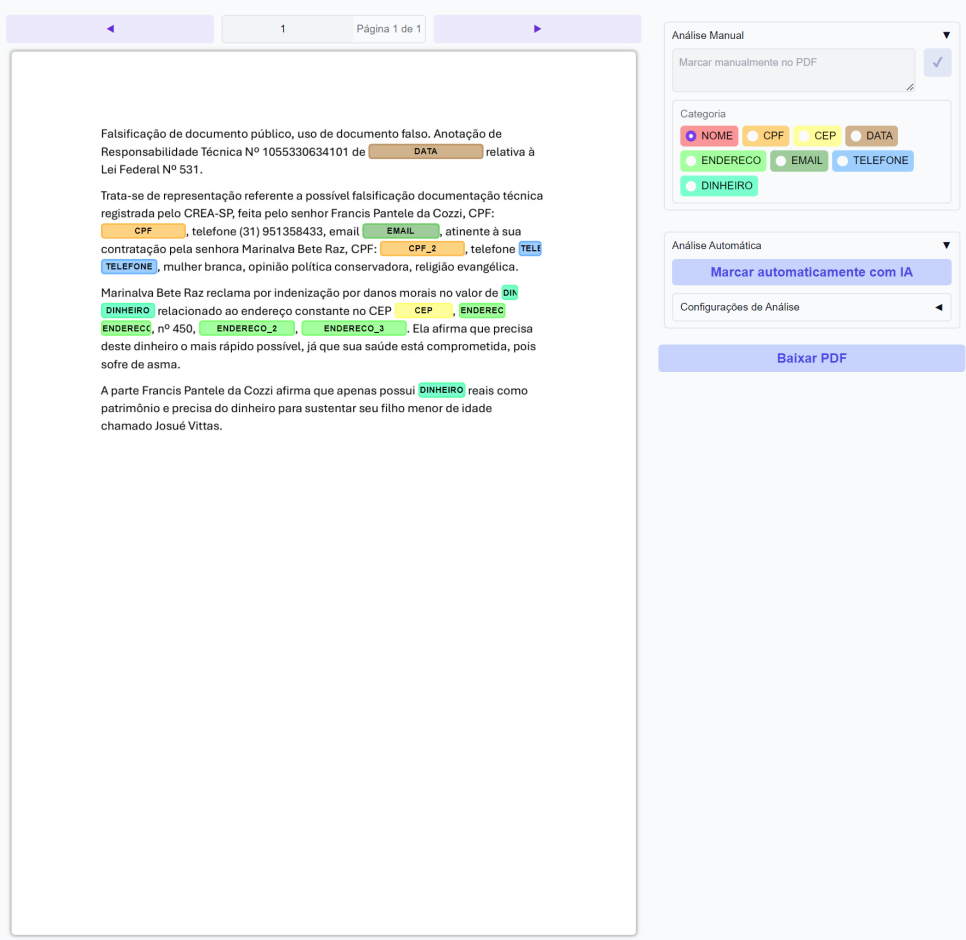

Figura 1.1: Protótipo.

1.1.1 Objetivos Específicos

Para atender ao objetivo, este estudo buscará dois pontos: treinar um novo modelo de NER, o qual terá o nome de **LEGAL-BERT-LGPD**, com dados específicos do domínio jurídico brasileiro, visando melhorar a exatidão na extração de informações relevantes. E outro, validar este novo modelo, comparando seu desempenho com outras LLMs (Large Language Models) populares em tarefas de classificação de tokens, especialmente no contexto jurídico. A Figura 1.2 ilustra o método deste estudo.

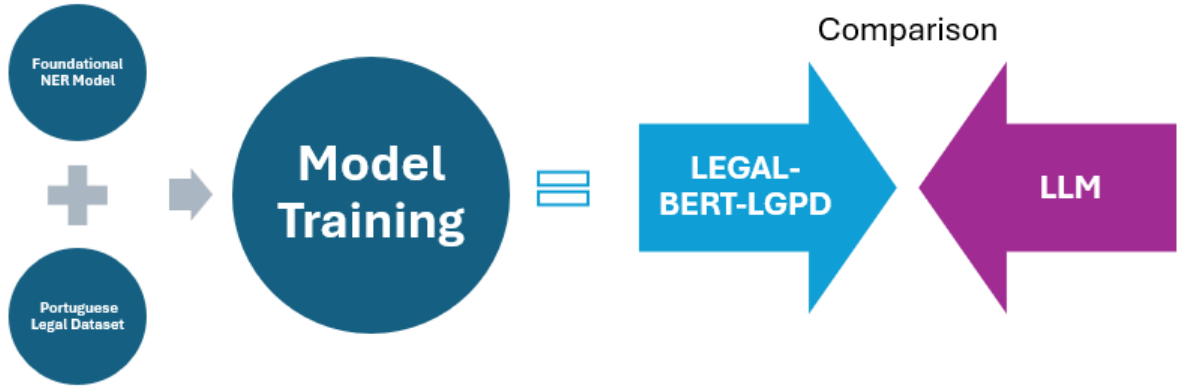

Figura 1.2: Método: treinamento do modelo LEGAL-BERT-LGPD, o qual consiste em utilizar um Modelo NER Fundacional mais um Conjunto de Dados Jurídico em Português, e em seguida, a comparação entre este modelo e os LLMs em cenários práticos de utilização.

1.2 Contribuições

Esperamos que este trabalho contribua para a proteção de dados pessoais no MPF, fornecendo uma solução eficaz e eficiente para a pseudonimização de documentos jurídicos. Para concluir, este trabalho apresentará suas conclusões e identificará áreas que ainda não foram exploradas para pesquisas futuras. Portanto, os resultados esperados incluem:

- **Criação de um conjunto de dados rotulado**: Como parte do processo, será desenvolvido um conjunto de dados anotado de forma a facilitar o treinamento e aprimoramento contínuo dos modelos de pseudonimização, contribuindo para o avanço do campo de textos jurídicos.
- **Redução de custos computacionais**: Utilizando um modelo que consuma menos recurso computacional, baseado em arquiteturas como Transformers, espera-se minimizar a necessidade de recursos computacionais elevados, tornando a solução acessível para ambientes com limitações tecnológicas.
- **Segurança**: Rodando localmente, garantir que os dados sensíveis não sejam expostos a terceiros, alinhando-se às diretrizes da LGPD e às recomendações do CNMP.

A estrutura restante do documento é a seguinte: no capítulo 2 serão discutidos os conceitos teóricos essenciais para compreender o trabalho; o capítulo 3 consiste em uma revisão da literatura, que teve como objetivo localizar estudos relacionados; o capítulo 4 descreve os resultados alcançados, enquanto o capítulo 5 resume as conclusões tiradas até agora e sugere possíveis trabalhos futuros nesse contexto.

Capítulo 2

Referencial teórico

Neste capítulo abordaremos uma contextualização do MPF, bem como os conceitos técnicos necessários para compreender a proposta de automação da pseudonimização.

2.1 Ministério Público Federal (MPF)

De acordo com a Constituição Federal de 1988, cabe ao Ministério Público brasileiro, como função essencial à Justiça, a defesa: [8]:

- Dos direitos sociais e individuais indisponíveis;
- Da ordem jurídica e;
- Do regime democrático.

O Ministério Público brasileiro é composto pelos Ministérios Públicos nos estados (atuam perante a Justiça estadual), e pelo Ministério Público da União (MPU), que, por sua vez, possui quatro ramos:

- Ministério Público Federal (MPF)
- Ministério Público do Trabalho (MPT)
- Ministério Público Militar (MPM)
- Ministério Público do Distrito Federal e Territórios (MPDFT)

O MPF atua como fiscal da lei, mas tem atuação também nas áreas cível, criminal e eleitoral. Na área eleitoral, o MPF pode intervir em todas as fases do processo e age em parceria com os ministérios públicos estaduais. O MPF atua na Justiça Federal, em causas nas quais a Constituição considera haver interesse federal. O MPF também age

Figura 2.1: Organização do Ministério Público Brasileiro.

preventivamente, extrajudicialmente, quando atua por meio de recomendações, audiências públicas e promove acordos por meio dos Termos de Ajuste de Conduta (TAC).

O MPU e o MPF são chefiados pelo(a) procurador(a)-geral da República, nomeado pelo presidente da República, com autorização da maioria absoluta do Senado Federal. A sede administrativa do MPF é a Procuradoria-Geral da República.

2.1.1 Estrutura do MPF

O MPF, assim como o MPB, não faz parte de nenhum dos três poderes (Executivo, Legislativo e Judiciário) e tem independência funcional assegurada pela Constituição Federal. O MPF atua em casos federais, regulamentados pela Constituição e pelas leis federais, sempre que a questão envolver interesse público. Além disso, o Ministério Público tem autonomia na estrutura do Estado: não pode ser extinto ou ter atribuições repassadas a outra instituição. Os membros (procuradores e promotores) possuem as chamadas autonomia institucional e independência funcional, ou seja, têm liberdade para atuar segundo suas convicções, com base na lei.

O Ministério Público Federal tem mais de 200 unidades espalhadas em todo o país [9]. A estrutura conta com:

- Procuradoria-Geral da República (PGR);
- Procuradorias Regionais da República (PRRs);
- Procuradorias da República nos estados e no Distrito Federal (PRs); e
- Procuradorias da República nos municípios (PRMs).

2.1.2 Recomendação do MPF para o Uso de IA

Conforme o Mapa Estratégico de 2022 a 2027 [10], na seção de Processos Internos há, entre outras, a seguinte descrição: "Incrementar o uso de inteligência artificial para auxiliar no processo de tomada de decisões e na automatização de procedimentos". A Figura abaixo mostra este ponto 2.2.

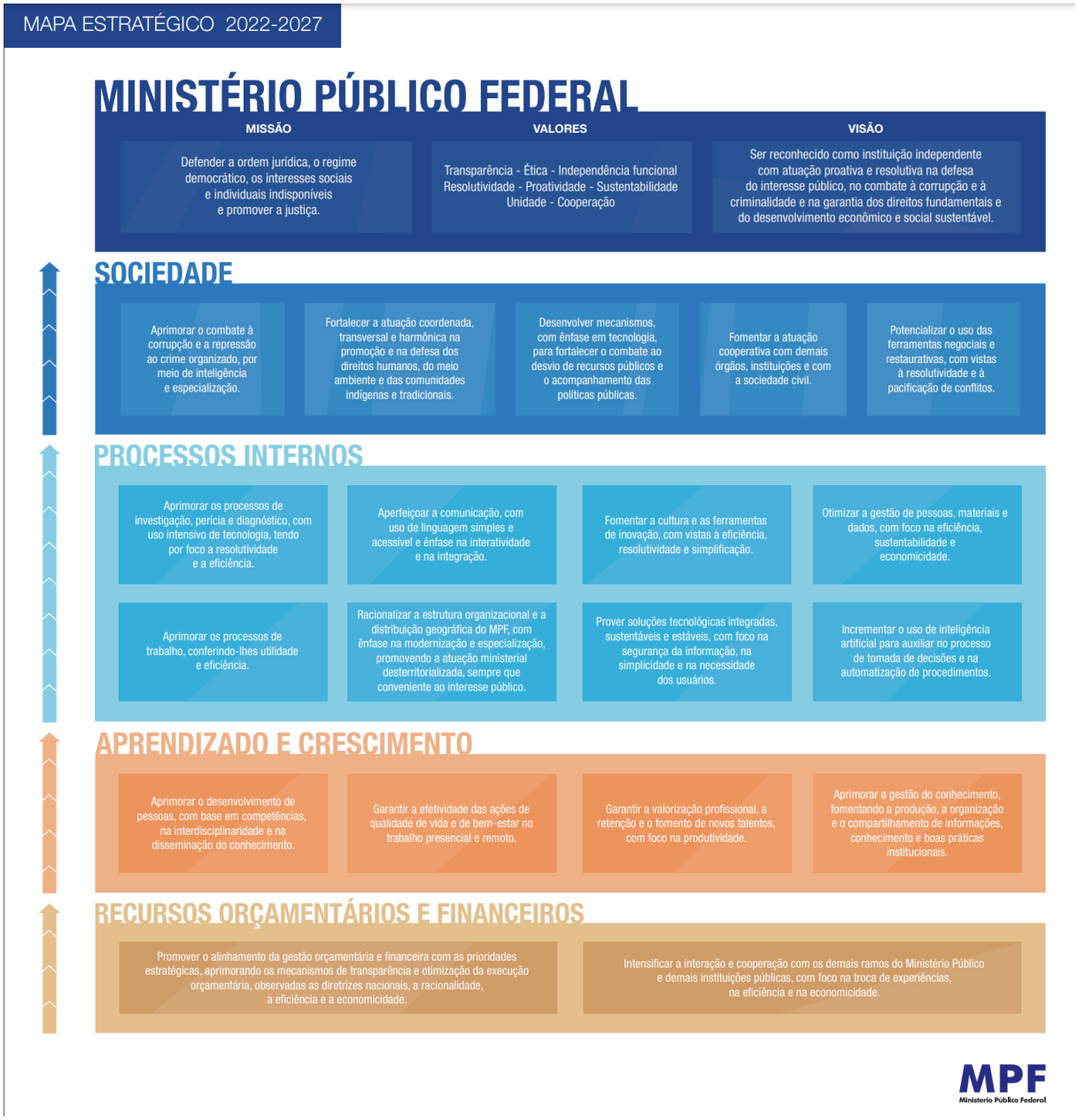

Figura 2.2: Mapa Estratégico do MPF de 2022 a 2027.

Portanto é possível afirmar que o órgão busca utilizar IA para apoiar seus processos vigentes. A solução aqui proposta pode facilitar a atividade do usuário de tarjar infor-

mações sensíveis em documentos jurídicos, garantindo a proteção de dados pessoais e o cumprimento da LGPD.

2.1.3 Dados Pessoais no MPF

Entre os dias 24 de maio e 14 de junho de 2021, as unidades administrativas do MPF participaram do questionário intitulado "Inventário de processos de trabalho para conformidade à Lei Geral de Proteção de Dados Pessoais (LGPD)" [1]. Esta iniciativa teve o objetivo de realizar um diagnóstico abrangente dos processos de trabalho institucionais que envolvem o tratamento de dados pessoais.

O Sistema de Governança do MPF registrou aproximadamente 360 processos de trabalho, distribuídos em 11 áreas estratégicas da instituição. Assim, foram identificados 210 subprocessos que envolvem o tratamento de dados pessoais, representando 58% da arquitetura dos processos institucionais. Como o questionário foi enviado a todas as unidades. A análise realizada permitiu, entre outros aspectos, identificar:

- O número de processos que tratam dados pessoais, de acordo com as respectivas áreas;
- Algumas particularidades dos tratamentos efetuados, como a presença de dados pessoais sensíveis, a utilização de operadores, e o compartilhamento desses dados com outras instituições;
- O ranqueamento das bases legais, os sistemas utilizados no tratamento de dados e os principais tipos de titulares de dados identificados.

Essas informações estão consolidadas na Figura 2.3 abaixo:

[1] `https://www.mpf.mp.br/servicos/lgpd/inventario-de-dados`

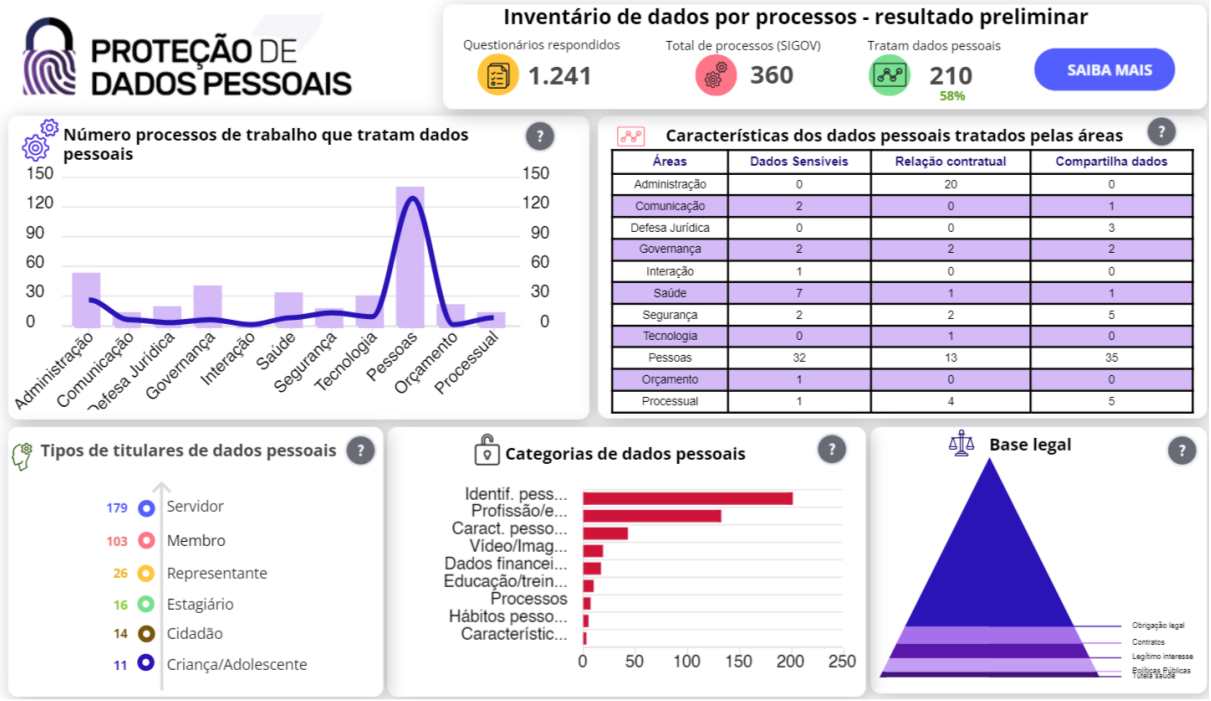

Figura 2.3: Inventário de Dados do MPF em 2021.

2.1.4 Aplicação de Anonimização de Forma Manual no MPF

O MPF é o órgão do MPU com maior capilaridade. No total, são mais de 200 unidades espalhadas em todo o país. Durante o período de 2016 e meados de 2025 (Figura 2.4), o MPF deu entrada e saída em 24 milhões de processos[2]. Isso demonstra a grande quantidade de processos que são tratados no MPF. Além disso, há demandas de vários cidadãos que solicitam que seus dados sejam anonimizados em documentos em posse do MPF. Juntos esses fatores tornam o processo de anonimização manual de documentos um processo demorado e sujeito a erros humanos.

[2] `https://www.mpf.mp.br/numeros`

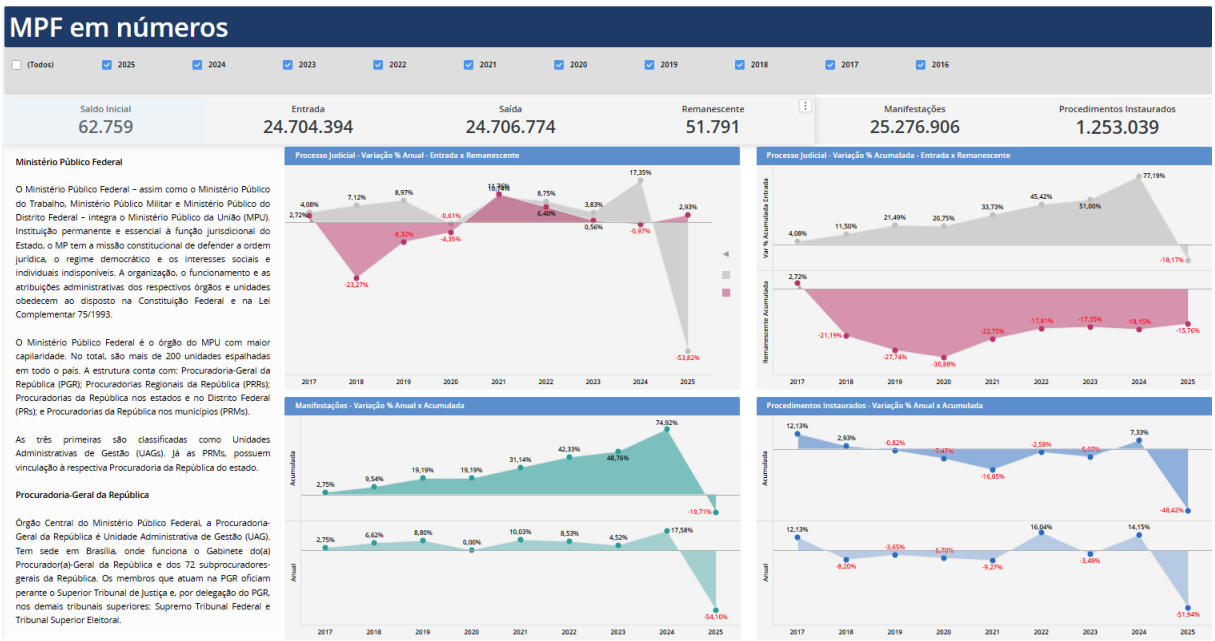

Figura 2.4: MPF em Números - Fluxo Processual x Manifestação x Procedimentos Instaurados entre 2016 e meados de 2025.

Atualmente os servidores do MPF utilizam ferramentas como o PDF24[3] para tarjar documentos PDF. A opção atual proposta é selecionar a opção "*Censurar PDF*" disponível no PDF24, conforme a Figura 2.5:

[3] `https://tools.pdf24.org/pt/`

Figura 2.5: Exemplo de uso do PDF24 para selecionar a ação de tarjar.

A anonimização desses documentos é um processo manual e demorado, que envolve a identificação e a remoção de informações sensíveis, como nomes, endereços e números de documentos, para proteger a privacidade dos indivíduos. Ainda, a aplicação de tarjas em documentos PDF não garante a proteção completa dos dados, pois as informações sensíveis ainda podem ser acessadas por meio de técnicas de recuperação de texto. Na Figura 2.6, temos um exemplo de como é aplicado.

2.2 Lei Geral de Proteção de Dados Pessoais (LGPD)

A Lei Geral de Proteção de Dados (LGPD [4], Lei nº 13.709, de 14 de agosto de 2018) é uma regulamentação brasileira que estabelece normas para a coleta, uso, processamento e armazenamento de dados pessoais. A lei concede maior controle aos indivíduos sobre suas informações pessoais, buscando equilibrar a privacidade com o avanço tecnológico. A LGPD exige que todas as operações com dados pessoais tenham uma justificativa legal e sigam princípios como finalidade, adequação e necessidade [6].

A LGPD abrange tanto o meio físico quanto o digital, estabelecendo que qualquer organização que processe dados de indivíduos em território nacional deve cumprir suas regras, independentemente da localização da sede da empresa ou dos servidores. Isso

[4] `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`

Figura 2.6: Exemplo de uso do PDF24 para tarjar.

inclui o compartilhamento de informações com organismos internacionais, desde que sejam respeitados os requisitos legais. A lei também define o conceito de dados sensíveis e impõe um cuidado especial na sua manipulação, reforçando a importância de um tratamento ético e responsável.

Um dos pilares centrais da LGPD é o consentimento do titular dos dados. Exceto em casos específicos previstos na lei, o tratamento de informações pessoais deve ser autorizado pelo titular. A LGPD garante ao cidadão o direito de solicitar a exclusão de seus dados, revogar o consentimento e até transferir suas informações para outro prestador de serviços. Além disso, o tratamento dos dados deve seguir os princípios de finalidade e necessidade, que devem ser previamente comunicados ao titular.

A fiscalização da LGPD é responsabilidade da Autoridade Nacional de Proteção de Dados (ANPD). Esta entidade tem a função de regulamentar e orientar as organizações sobre a aplicação da lei, além de aplicar sanções em casos de descumprimento. A LGPD também define os papéis de controlador, operador e encarregado dentro das empresas, estabelecendo um sistema de governança e segurança que visa prevenir falhas e minimizar riscos. Multas de até 2% do faturamento anual, limitadas a R$ 50 milhões por infração, podem ser aplicadas, destacando a seriedade da conformidade com a lei [11].

2.3 Pseudonimização

Nesta seção, veremos algumas diferenças entre anonimização e pseudonimização, bem como a recomendação do CNMP para a pseudonimização de dados pessoais.

2.3.1 Anonimização X Pseudonimização

Anonimização e pseudonimização são conceitos importantes na proteção de dados. A anonimização (LGPD, art. 5º, XI) é o processo de tratar os dados pessoais para que não possam mais ser associados a uma pessoa específica, mesmo com o uso de técnicas avançadas. Esse processo é irreversível, garantindo um alto nível de proteção à privacidade do indivíduo.

Já a pseudonimização (LGPD, art. 13º, § 4º) é uma técnica que altera os dados pessoais de modo que a identificação do titular só seja possível com o uso de informações adicionais, que são mantidas separadamente sob proteção. Ao contrário da anonimização, a pseudonimização é reversível, desde que essas informações adicionais sejam acessíveis [5]. Este trabalho adotará a pseudonimização, pois ela permite proteger os dados sem comprometer sua utilidade para análises e pesquisas.

2.3.2 Recomendação do CNMP para Pseudonimização de Dados

Consoante a norma, a Resolução CNMP nº 281, de 12 de dezembro de 2023 reforça os conceitos de anonimização e pseudonimização e, em seu art. 79, apresenta a pseudonimização com uma alternativa para a proteção dos dados pessoais no âmbito de processos que tramitam no Ministério Público:

- *Art. 79. A fim de assegurar a proteção aos dados pessoais das pessoas naturais no âmbito de procedimentos ou processos que tramitam no Ministério Público, poderá ser promovido o controle de acesso, a* ***pseudonimização*** *ou a decretação de sigilo dos autos ou de documentos específicos neles contidos, inclusive em relação às petições e aos documentos juntados pelas partes envolvidas.*

Para a Resolução, a pseudonimização é o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro (art. 4º, inciso XXVI).

[5] `https://www.cnmp.mp.br/portal/images/CALJ/resolucoes/Resoluo-281-de-2023.pdf`

2.4 NER

Nesta seção, abordaremos os conceitos fundamentais de NER, lacunas de conjuntos de dados para ner, além das métricas de avaliação.

2.4.1 Entidades e Rótulos

No contexto do PLN, uma **entidade nomeada** refere-se a um item específico ou objeto do mundo real mencionado em um texto não estruturado. Essas entidades são classificadas em tipos semânticos predefinidos, como pessoas, locais, organizações, datas e quantidades. Já o **rótulo** (ou *label*) em NER é a categoria semântica predefinida atribuída a uma entidade identificada no texto. A tarefa de NER visa localizar e categorizar essas entidades em suas categorias correspondentes. Em tarefas de NER, a saída para uma sequência de *tokens* inclui uma coleção de tuplas que especificam o início, o fim e o tipo de entidade (o rótulo) [1].

2.4.2 Ausência de Dados Pessoais Anotados no Domínio Jurídico Brasileiro

Uma das principais limitações dos estudos de NER no domínio jurídico é o foco em categorias genéricas, como *PESSOA*, *LOCAL*, *ORGANIZAÇÃO*, *LEGISLAÇÃO* e *JURISPRUDÊNCIA* [2, 3]. Documentos jurídicos frequentemente contêm dados sensíveis, como CPF, telefones e endereços, que precisam ser tratados de forma específica para atender às demandas de proteção de dados. Conforme nossa pesquisa prévia, vimos uma escassez de conjuntos de dados anotados especificamente para o domínio jurídico e dados pessoais em português brasileiro. Além disso, a criação de conjuntos de dados específicos é um processo caro e demorado, demandando colaboração entre especialistas em direito e cientistas de dados.

Existem modelos em outros idiomas que tratam de dados pessoais (Piiranha-v1[6], GLiNER PII[7]), porém não refletem o formato dos dados brasileiros. Vide o CPF, cujo formato é 000.000.000-00, ou o RG, que varia de acordo com o estado. Portanto, a adaptação de modelos existentes para o português brasileiro é um desafio adicional, que requer a criação de conjuntos de dados anotados específicos para o domínio jurídico e dados pessoais em português brasileiro.

Para superar essas limitações, é essencial o desenvolvimento de conjuntos de dados específicos e anotados com granularidade suficiente para cobrir as categorias exclusivas

[6] `https://huggingface.co/iiiorg/piiranha-v1-detect-personal-information`
[7] `https://huggingface.co/urchade/gliner_multi_pii-v1`

do domínio jurídico. Esses conjuntos de dados devem refletir a terminologia, estrutura e necessidades legais, permitindo que modelos de NER sejam treinados de maneira mais eficaz. Além disso, a incorporação de especialistas jurídicos no processo de anotação é crucial para garantir a qualidade e a precisão dos dados.

2.4.3 *Tokens* e *Chunks*

Um *token* é a unidade básica de processamento em PLN (Processamento de Linguagem Natural), geralmente correspondendo a uma palavra, pontuação ou outro elemento significativo do texto [1]. A tarefa de NER envolve a classificação de cada *token* de um texto de acordo com os tipos de entidade considerados, além de determinar se esse *token* ocorre no início, meio ou fim de uma entidade identificada. Os documentos são frequentemente divididos em sentenças e, em seguida, *tokenizados* [2].

No contexto de NER, *chunks* referem-se a sequências de *tokens* que formam uma entidade nomeada ou uma parte dela. Enquanto algumas abordagens classificam *tokens* individualmente (*token-based*), as estratégias *span-based* primeiro identificam todos os *spans* (sequências de *tokens*) menores que um limite definido e, em seguida, classificam cada um deles [12]. A tarefa de NER é, em essência, a identificação desses segmentos de texto que mencionam entidades nomeadas.

2.4.4 Métricas de Avaliação: F1, Recall e Precision

A **Precisão** mede a proporção de identificações de entidades corretas entre todas as identificações de entidades feitas pelo modelo para um determinado tipo. Uma alta precisão indica que poucos itens irrelevantes foram incluídos [12]. Já a **Revocação** (ou *Recall*) mede a proporção de entidades corretas que foram identificadas pelo modelo em relação ao total de entidades reais presentes nos dados para um determinado tipo. Um alto *Recall* significa que o modelo encontra a maioria das instâncias relevantes, embora possa retornar mais falsos positivos [13].

O **F1-Score** é a média harmônica entre a Precisão e a Revocação. É uma métrica importante porque oferece um equilíbrio entre Precisão e Revocação, sendo particularmente útil quando há um desequilíbrio de classes nos dados. Um alto F1-Score indica que o modelo tem tanto alta precisão quanto alto *recall*. O F1-Score é amplamente utilizado para avaliar o desempenho geral de modelos de NER em diversos domínios [3]. O *F1-score* é a média harmônica entre precisão e revocação, equilibrando ambas as métricas. Um valor alto de F1 só é alcançado quando precisão e revocação são simultaneamente elevadas [13].

Sua equação é dada por:

$$F1 = 2 \times \frac{\text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}}.$$

Capítulo 3

Trabalhos relacionados

Neste capítulo abordaremos os principais trabalhos e bases conceituais que sustentam esta pesquisa. Primeiro, apresentaremos estudos voltados à LGPD. Em seguida, revisaremos o estado da arte em NER, cobrindo desde avanços internacionais até corpora jurídicos brasileiros. Conjuntamente, essa análise evidencia a necessidade de soluções que conciliem conformidade à LGPD com abordagens modernas de NER, lacuna que o LEGAL-BERT-LGPD se propõe a preencher.

3.1 LGPD

Os estudos de Oliveira (2020) [14] e Rapôso et al. (2019) [15] complementam o presente trabalho ao abordarem aspectos cruciais da proteção de dados pessoais e da implementação da LGPD nas instituições públicas. Enquanto Oliveira foca na gestão de riscos e na preservação do direito à privacidade no âmbito judicial eletrônico, Rapôso et al. fornecem uma visão abrangente dos desafios tecnológicos na adaptação à LGPD. Ambos os trabalhos reforçam a relevância de soluções baseadas em IA, como o NER, para assegurar o tratamento adequado de dados pessoais, alinhando-se aos objetivos deste estudo.

3.1.1 Gestão de Riscos no Processo Judicial Eletrônico

O trabalho de Oliveira (2020) [14] aborda os desafios da gestão de dados no ambiente digital, especialmente em processos judiciais eletrônicos, em que a exposição de dados pessoais pode comprometer a privacidade das partes envolvidas. Trabalhos como o de Oliveira (2020) investigam o potencial de violações de privacidade, destacando a vulnerabilidade dos sistemas de justiça digital, que podem expor dados das partes envolvidas, mesmo em processos que tramitam sob sigilo de justiça.

Uma abordagem frequentemente citada para mitigação de riscos é o uso de pseudonimização e de-identificação dos dados, conforme o modelo proposto pela legislação norte-americana HIPAA. Oliveira sugere que, no contexto judicial, as iniciais das partes processuais não são suficientes para garantir a anonimidade, especialmente quando combinadas com outros dados, como cidade ou profissão, o que facilita a reidentificação.

3.1.2 Desafios Tecnológicos na Adaptação a LGPD

O artigo "*LGPD - Uma visão de tecnologia e agnóstica*" de Nunes e Santos (2023) [15] analisa a LGPD sem dependência de tecnologias específicas. O estudo aborda que o foco é garantir a conformidade com a lei através de abordagens técnicas e organizacionais que se adaptem ao cenário digital em constante evolução. O trabalho também destaca a importância de políticas que protejam dados pessoais, em conformidade com os princípios estabelecidos pela LGPD.

Além disso, um dos principais tópicos abordados é a definição de dados pessoais e as técnicas de *anonimização* e *pseudonimização*, essenciais para garantir a privacidade dos indivíduos. O artigo discute também o uso de tecnologias emergentes, como *Internet das Coisas* (IoT), *blockchain*, e *Machine Learning*, para aprimorar a conformidade com a LGPD. O conceito de *privacidade por design* é destacado, no qual a privacidade é integrada desde a concepção dos sistemas. Este princípio é essencial para que as empresas lidem com os desafios legais e tecnológicos de forma proativa, promovendo a proteção dos dados em ambientes digitais.

Outro ponto relevante do trabalho é a relação entre LGPD e *eDiscovery*, que envolve a coleta e análise de dados eletrônicos em processos judiciais. O eDiscovery precisa garantir que as informações coletadas respeitem os direitos dos titulares dos dados, equilibrando a necessidade de informações com as exigências de privacidade. Isso exige um alinhamento cuidadoso entre as tecnologias de coleta de dados e as regulamentações de proteção de dados.

3.1.3 Privacidade de dados pessoais em Infraestruturas de Terceiros

O uso de IA no MPF também envolve questões de proteção de privacidade e direitos autorais em sistemas de IA generativa. O estudo de Ferreira Zhang et al. [16] discutem a importância de um ciclo de vida que contemple a proteção de dados em sistemas de IA, desde a geração de dados até seu armazenamento e uso. Esse trabalho é relevante para nosso trabalho, pois ao implementar soluções de IA para pseudonimização de dados, é

essencial garantir que os dados gerados ou manipulados sigam rigorosamente as diretrizes de proteção de privacidade.

Ferreira [17], por sua vez, concentra-se nos desafios de aplicar a LGPD em ambientes de nuvem pública, um cenário cada vez mais comum para armazenamento e processamento de dados. O autor destaca a complexidade de garantir a segurança e a privacidade de dados pessoais em infraestruturas de terceiros, onde a instituição responsável pelos dados não possui controle total sobre os mecanismos de segurança. A necessidade de gerenciar chaves criptográficas em locais distintos, a fim de garantir a anonimização irreversível dos dados, adiciona uma camada extra de complexidade à implementação da LGPD em ambientes de nuvem.

A pesquisa de Ferreira [17] destaca a importância da anonimização e pseudonimização de dados como estratégias fundamentais para proteger Dados Pessoais Identificáveis (DPI). A adoção de modelos de IA que automatizam esse processo melhora a eficiência e reduz a margem de erro humano, tornando-se uma solução necessária frente ao volume de dados processados diariamente no MPF.

Os artigos de Ferreira [17] e Zhang et al. [16] complementam-se ao abordar tanto o aspecto prático da implementação da LGPD com o uso de IA quanto as implicações mais amplas relacionadas à privacidade e direitos autorais no uso de modelos de IA generativa. Esses estudos reforçam a importância de adotar uma abordagem holística para a proteção de dados, desde a concepção de modelos de IA até a sua aplicação no contexto jurídico e institucional, como no MPF.

3.2 NER

3.2.1 Estado da Arte em NER

O NER tem experimentado avanços significativos impulsionados por técnicas de aprendizado profundo, como descrito em *"A Survey on Recent Advances in Named Entity Recognition"*, trabalho de Keraghel, I., Morbieu, S., Nadif, M. (2024) [1].

Arquiteturas baseadas em transformers, como BERT e suas variantes, demonstraram desempenho superior em conjuntos de dados extensos. Sua capacidade de capturar relações complexas entre palavras em sequências longas e a capacidade de serem pré-treinados em grandes corpus de texto os tornam particularmente eficazes para NER. Contudo, a pesquisa destaca que, apesar do sucesso geral dos transformers, modelos específicos como o GPT enfrentam desafios na desambiguação e detecção precisa de entidades nomeadas compostas, um aspecto crucial em tarefas de NER.

Em outro ponto, o estudo destaca que um desafio significativo em NER é a escassez de dados anotados, especialmente para domínios especializados ou línguas com poucos recursos. Para mitigar essa limitação, técnicas como *Transfer Learning*, aumento de dados, aprendizado ativo e aprendizado com poucos exemplos (*few-shot learning*) têm sido exploradas. Essas técnicas permitem que os modelos aprendam de forma eficiente, mesmo quando o volume de dados rotulados é limitado.

Os autores destacaram que diversos *frameworks* e ferramentas têm facilitado a implementação de sistemas de NER. Entre os mais populares, o *spaCy*, *NLTK*, *Apache OpenNLP*, *Stanford CoreNLP*, e *Flair*, que oferece suporte para múltiplos idiomas. Além disso, bibliotecas baseadas em transformadores, como *Hugging Face Transformers*, têm ganhado popularidade por fornecerem modelos pré-treinados que podem ser facilmente ajustados para diferentes tarefas de NER.

Por fim, verificaram que a avaliação de sistemas de NER é realizada com base em métricas como precisão, *recall* e F1-score. Além disso, diferentes esquemas de anotação, como BIO e IOBES, são utilizados para definir os limites das entidades nomeadas, influenciando diretamente o desempenho dos modelos. As avaliações podem seguir tanto estratégias exatas quanto relaxadas, dependendo da precisão desejada na identificação e classificação das entidades.

A Figura 3.1 constante no estudo citado exibe uma comparação dos frameworks de NER em termos de F1-Score Macro-médio, destacando o desempenho de modelos como BERT, RoBERTa, DistilBERT, entre outros. Esses resultados evidenciam a eficácia dos transformers em tarefas de NER, especialmente quando treinados em grandes conjuntos de dados.

Frameworks	Algorithms	CoNLL-2003	OntoNotes	WNUT2017	FIN	BioNLP2004	NCBI Disease	BC5CDR	MITRestaurant	Few-NERD	MultiCoNER
		Macro-averaged F1-score									
Apache OpenNLP	Maximum Entropy	80.00	67.83	-	63.24	-	-	-	-	-	-
Stanford CoreNLP	CRF	85.18	63.87	8.34	55.25	**73.26**	86.10	85.22	70.57	45.13	19.39
Flair	LSTM-CRF	90.35	80.10	38.07	**74.23**	71.64	86.21	**90.27**	78.33	60.03	56.27
spaCy	CNN-small	81.26	69.30	9.01	55.12	65.92	77.92	80.83	75.62	40.55	35.63
	CNN-large	85.64	69.60	9.78	54.71	66.17	79.15	79.66	76.39	40.01	35.82
	roberta-base	89.92	81.04	41.84	63.18	66.56	**87.05**	87.08	79.09	59.15	55.21
Hugging Face	xlm-roberta-large	**91.46**	**81.57**	**43.92**	48.68	71.43	85.25	87.41	**80.12**	**61.59**	58.15
	distilbert-base-cased	88.12	77.63	25.45	43.74	69.63	84.42	84.03	77.67	58.62	55.17
	bert-base-uncased	88.89	76.99	32.12	46.84	70.50	85.64	85.78	79.18	58.16	**59.96**
	bert-base-cased	90.09	79.55	33.32	39.53	69.46	85.27	85.14	78.48	59.48	56.64
OpenAI	GPT-4	62.74	33.61	18.82	36.70	41.32	57.46	55.67	41.38	44.96	33.61

Figura 3.1: Comparação dos frameworks de NER em termos de F1-Score Macro-médio. Os melhores e os segundos melhores resultados estão, respectivamente, em negrito e sublinhados. Os traços representam valores ausentes, pois o Apache OpenNLP não permite a adição de outras categorias de entidades nomeadas além dos quatro tipos padrão: pessoas, organizações, locais e diversos. Adaptado de [1].

3.2.2 Corpus Brasileiros no Domínio jurídico

LeNER-Br (2018)

O uso de corpora especializados em textos jurídicos no Brasil tem se mostrado uma área de crescente interesse na pesquisa de IA e Processamento de Linguagem Natural (PLN). Um exemplo significativo é o *LeNER-Br*, um conjunto de dados desenvolvido para o NER em textos legais brasileiros [2]. Esse corpus foi anotado manualmente e projetado especificamente para o domínio jurídico.

O conjunto de dados LeNER-Br é composto por 66 documentos jurídicos provenientes de vários tribunais brasileiros. Além disso, foram coletados quatro documentos legislativos, como a "Lei Maria da Penha", totalizando 70 documentos no conjunto de dados. A anotação dos documentos foi realizada manualmente com auxílio de uma ferramenta chamada WebAnno [18]. Foram utilizadas as seguintes categorias de entidades: *ORGANIZAÇÃO*, *PESSOA*, *TEMPO*, *LOCAL*, *LEGISLAÇÃO* e *JURISPRUDÊNCIA*. As categorias de *LEGISLAÇÃO* e *JURISPRUDÊNCIA* foram incluídas para representar especificamente leis e decisões jurídicas.

Para rotular as entidades, os autores utilizaram a notação IOB, onde "B-" indica o início de uma entidade nomeada, "I-" indica que o token está dentro de uma entidade, e "O-" indica que o token não pertence a nenhuma entidade nomeada. O conjunto de dados foi dividido em conjuntos de treinamento, desenvolvimento e teste, com 50, 10 e 10 documentos, respectivamente. No total, o LeNER-Br contém aproximadamente 318.073 tokens. A tabela 3.1 apresenta o número de documentos, sentenças e tokens em cada conjunto.

Tabela 3.1: Contagem de palavras de entidades nomeadas para cada conjunto. Fonte [2].

Categoria	**Treinamento**	**Desenvolvimento**	**Teste**
Pessoa	4.612	894	735
Casos Jurídicos	3.967	743	660
Tempo	2.343	543	260
Local	1.417	244	132
Legislação	13.039	2.609	2.669
Organização	6.671	1.608	1.367

CDJUR-BR (2023)

Outro trabalho relevante é o *CDJUR-BR*, uma coleção de documentos legais brasileiros com anotações detalhadas de entidades nomeadas [3]. Essa coleção se distingue por sua granularidade na identificação de entidades, o que a torna valiosa para análises mais profundas e refinadas do conteúdo jurídico. A coleção *CDJUR-BR* foi desenvolvida para atender à demanda crescente por bases de dados específicas e anotadas no domínio jurídico em português, uma vez que os textos legais possuem terminologia técnica e complexa, como leis, réus e normas.

Embora o trabalho discuta a criação e a metodologia utilizada para o desenvolvimento do corpus CDJUR-BR, não conseguimos obter os dados do conjunto de dados. Percebemos que para notações, provavelmente foi utilizado o formato IOB (In, Out, Begin), que é comum em tarefas de NER. No tópico "*4.3 Resultados e Discussões*", vemos que há uma análise de erros e há uma indicação do formato IOB.

O estudo destaca a dificuldade de utilizar modelos genéricos de NER, dada a ausência de corpora específicos com entidades nomeadas refinadas, como pessoas, endereços e penas. Para resolver essas limitações, o estudo propôs um corpus de 1.216 documentos oriundos do Tribunal de Justiça do Ceará (TJCE), abrangendo várias classes processuais, como inquéritos e denúncias. O estudo também avaliou três modelos de aprendizado de máquina para a tarefa de NER, com destaque para o *BERT*, que alcançou um F1-Score de 0,58, superando modelos como *BI-LSTM+CRF* e *SpaCy*.

Neste estudo, foi desenvolvida uma metodologia própria para a anotação manual de documentos jurídicos, resultando na criação da CDJUR-BR, uma coleção padrão-ouro com 44.526 anotações de 21 entidades nomeadas, conforme vemos na Tabela 3.2. O processo de anotação foi rigorosamente avaliado. Experimentos com modelos de aprendizado de máquina, como SPACY, BI-LSTM+CRF e BERT, demonstraram a viabilidade da CDJUR-BR, com o modelo BERT obtendo o melhor desempenho (Média Macro F1 de 0,58).

Tabela 3.2: Resultados de F1-score para o Reconhecimento das Entidades Específicas (C1) utilizando os modelos BI-LSTM+CRF, SPACY e BERT. Fonte [3].

Entidade Nomeada	**BI-LSTM+CRF**	**SPACY**	**BERT**	**Suporte**
END-AUTOR	0.56	0.31	0.33	18
END-DELITO	0.72	0.45	0.73	61
END-OUTROS	0.00	0.02	0.16	81
END-REU	0.55	0.59	0.71	152
END-TESTEMUNHA	0.27	0.26	0.67	68
END-VITIMA	0.06	0.00	0.22	27
NOR-ACESSÓRIA	0.79	0.79	0.82	990
NOR-JURISPRUDÊNCIA	0.90	0.87	0.89	333
NOR-PRINCIPAL	0.67	0.71	0.77	791
PENA	0.56	0.39	0.50	82
PES-ADVOG	0.54	0.22	0.63	122
PES-AUTOR	0.59	0.59	0.56	169
PES-AUTORID-POLICIAL	0.87	0.66	0.90	300
PES-JUIZ	0.79	0.50	0.78	83
PES-OUTROS	0.54	0.44	0.58	1.210
PES-PROMOTOR-MP	0.81	0.27	0.88	57
PES-REU	0.64	0.57	0.71	1.503
PES-TESTEMUNHA	0.57	0.45	0.64	519
PES-VÍTIMA	0.33	0.23	0.46	405
PROVA	0.47	0.29	0.34	461
SENTENÇA	0.00	0.29	0.00	11
F1-micro avg	0.64	0.55	0.67	7.443
F1-macro avg	0.53	0.42	0.58	7.443
F1-weighted avg	0.62	0.54	0.67	7.443

A coleção *CDJUR-BR* foi comparada ao *LeNER-BR* (Tabela 3.3), outra base de dados jurídica, e demonstrou uma maior capacidade de generalização ao reconhecer entidades em outros corpora. A tabela mostra os diferentes cenários executados. Devido à variação significativa no número de anotações entre as categorias de entidades nomeadas, no estudo da coleção *CDJUR-BR*, foi elaborada uma heurística para manter a proporção equilibrada

de exemplos ao dividir os dados em treino, validação e teste. Isso evitou a escassez de exemplos nas categorias. Os cenários foram:

- **Cenário C1**: Reconhecimento das entidades específicas da CDJUR-BR
- **Cenário C2**: Reconhecimento das categorias da CDJUR-BR
- **Cenário C3**: Reconhecimento das categorias de entidades da LENER-BR a partir de modelo treinado com LENER-BR
- **Cenário C4**: Reconhecimento das entidades do LENER-BR a partir de modelo treinado com CDJUR-BR
- **Cenário C5**: Reconhecimento das categorias de entidades da CDJUR-BR a partir de modelo treinado com LENER-BR

Apesar de resultados inferiores ao LENER-BR em exatidão, a CDJUR-BR se mostrou mais adaptável na tarefa de reconhecer entidades no corpus LENER-BR. O estudo sugere, como melhorias futuras, o balanceamento de entidades minoritárias e o desenvolvimento de classificadores especializados para entidades jurídicas refinadas.

Tabela 3.3: Resultados de F1-score para o NER na CDJUR-BR e LENER-BR (C2, C3, C4 e C5) utilizando o modelo BERT. Adaptado de Fonte [3].

Entidade	**Cenário C2**	**Cenário C3**	**Cenário C4**	**Cenário C5**
JURISPRUDÊNCIA	0.89	0.96	0.79	0.48
LEGISLAÇÃO	0.92	0.97	0.92	0.86
LOCAL	0.77	0.77	0.32	0.15
PESSOA	0.83	0.97	0.69	0.76
F1-micro avg	0.85	0.96	0.81	0.60
F1-macro avg	0.85	0.92	0.68	0.56
F1-weighted avg	0.85	0.96	0.79	0.74

Corpus Carolina (2022)

Por fim, o *Corpus Carolina*, um projeto aberto voltado para a linguística e inteligência artificial, oferece um grande conjunto de dados com diversas anotações [4]. O artigo *"Carolina's Methodology: building a large corpus with provenance and typology information"* [19] apresenta a metodologia **WaC-wiPT** desenvolvida para o *Carolina Open Corpus*.

Esse corpus está sendo criado no Centro de Inteligência Artificial da Universidade de São Paulo (C4AI-USP) e tem como objetivo fornecer uma ampla base de dados para pesquisas em Linguística e Ciência da Computação. A inovação do *Carolina Open Corpus* está na garantia de procedência e tipologia das fontes, uma característica pouco abordada por outras abordagens de construção de corpora.

A metodologia *WaC-wiPT* combina técnicas automáticas de coleta de textos na web com uma rigorosa curadoria de metadados. O foco é priorizar textos de acesso aberto, com fontes que variam entre textos jurídicos, jornalísticos, redes sociais e domínios acadêmicos. Um diferencial importante é o uso de cabeçalhos XML com metadados detalhados, que seguem os padrões da *Text Encoding Initiative* (TEI), garantindo a rastreabilidade e a transparência do conteúdo extraído, algo que se destaca frente a outras metodologias.

Outro ponto chave do artigo é a discussão sobre os desafios enfrentados para assegurar a qualidade e os direitos de uso do corpus. Mesmo com a automação, a inspeção humana foi necessária para identificar e corrigir problemas que não podem ser facilmente detectados por algoritmos. A fase prototípica do *Carolina Open Corpus* já extraiu mais de 1 bilhão de tokens, e a primeira versão pública, denominada Carolina 1.0 (Ada), foi planejada para março de 2022.

A equipe planeja expandir o corpus, assim como feito com a versão 1.1 [4], aprimorando a coleta de metadados, balanceando a tipologia textual e desenvolvendo ferramentas de análise linguística mais sofisticadas. O conjunto de dados está disponível no repositório do Huggingface [1] e contém o conjunto de dados conforme a Tabela 3.4:

Tabela 3.4: Dados do *Carolina Open Corpus*. Fonte [4].

Código	**Taxonomia**	**Instâncias**	**Tamanho**
dat	Conjuntos de Dados e Outros Corpora	1.102.049	4,4 GB
wik	Wikis	960.139	5,2 GB
jud	Poder Judiciário	40.464	1,5 GB
leg	Poder Legislativo	13	25 MB
soc	Mídia Social	3.413	17 MB
uni	Domínios Universitários	941	10 MB
pub	Obras de Domínio Público	26	4,5 MB
Total	-	2.107.045	11 GB

[1] `https://huggingface.co/datasets/carolina-c4ai/corpus-carolina`

O *Carolina Open Corpus* ainda não foi anotado para tarefas de NER. No entanto, o corpus inclui uma quantidade significativa de dados provenientes do Poder Judiciário, o que oferece um potencial interessante para aplicações de NER, especialmente no contexto de textos jurídicos. A presença de dados judiciais pode facilitar o desenvolvimento de modelos de NER voltados à pseudonimização de informações sensíveis, atendendo às necessidades de conformidade com legislações de privacidade como a Lei Geral de Proteção de Dados Pessoais (LGPD).

3.2.3 Engenharia de Prompt

A engenharia de *prompt* tem se destacado como uma área crucial na pesquisa em Processamento de Linguagem Natural (PNL), especialmente com o advento de modelos de linguagem extensos como o GPT-3. No contexto de NER, a técnica consiste em formular *prompts* eficazes que guiam o modelo a identificar e classificar entidades em um texto. Oliveira, Vitor, et al. [20], por exemplo, demonstraram como a engenharia de *prompt* pode ser utilizada para gerar dados rotulados automaticamente para o treinamento de modelos de NER em documentos jurídicos. Os autores exploraram o uso de *prompts* que forneceram ao GPT-3 exemplos de textos anotados, permitindo que o modelo aprendesse a estrutura e os padrões de rotulagem para, em seguida, aplicar esse conhecimento a novos textos.

Um dos principais desafios da engenharia de *prompt* reside na necessidade de encontrar o equilíbrio entre *prompts* informativos e concisos. *Prompts* excessivamente longos podem confundir o modelo, enquanto *prompts* muito curtos podem não fornecer informações suficientes para uma rotulagem precisa. O trabalho de Oliveira, Vitor, et al [20]. destacou essa dificuldade, mencionando as limitações de tamanho de *prompt* do GPT-3 e a necessidade de segmentar os documentos e utilizar exemplos de *prompts* cuidadosamente selecionados para garantir a presença de todas as entidades relevantes. Os autores também observaram que a escolha de exemplos representativos e a ordem de apresentação dos *prompts* podem influenciar significativamente o desempenho do modelo.

A pesquisa em engenharia de *prompt* para NER ainda está em desenvolvimento, com diversas oportunidades para futuras investigações. Oliveira, Vitor, et al [20]. sugeriram a exploração de técnicas de aprendizado ativo para a seleção de *prompts* mais informativos, bem como a integração de métodos de *weak supervision* para complementar a rotulagem baseada em *prompts*. A busca por estratégias mais sofisticadas para a formulação de *prompts*, considerando o contexto específico da tarefa e do domínio, também foi apresentada como uma área promissora para futuras pesquisas.

Um outro trabalho de relevância é o de Villena et al. [21], que exploraram como LLMs, com capacidades de reconhecimento de entidade em *zero-shot* ou *few-shot*, podem ser

aplicados a essa tarefa com excelente desempenho. Essa abordagem reduziu a necessidade de grandes quantidades de dados anotados manualmente, o que foi crucial no contexto de dados jurídicos, em que a exatidão e a privacidade são primordiais.

Além disso, o uso de LLMs para NER facilita a adaptação a novos contextos e entidades, tornando o processo mais escalável e flexível. No caso do MPF, onde milhões de registros são processados diariamente, essa capacidade de adaptação é fundamental para garantir que as soluções de IA acompanhem a evolução das necessidades institucionais e as regulamentações legais, como a LGPD [21].

3.2.4 Modelo BERT para NER

O modelo BERT (Bidirectional Encoder Representations from Transformers), apresentado por Devlin et al. (2019) [22], revolucionou o processamento de linguagem natural, incluindo a área de NER. Sua capacidade de analisar um texto bidirecionalmente, considerando o contexto completo de cada palavra, o torna extremamente eficaz na identificação e classificação de entidades. Andrade, Claudio MV, et al. [12] exploraram o potencial do BERT em um cenário desafiador: NER em dados sensíveis de reclamações de consumidores. Os autores destacaram que a aplicação direta de modelos como o BERT em dados sensíveis pode apresentar problemas de privacidade e alto custo de infraestrutura, especialmente ao utilizar APIs externas.

Para contornar essas limitações, Andrade, Claudio MV, et al. [12] propuseram uma abordagem que combina o poder de inferência do BERT com a eficiência e segurança de modelos mais simples, como o SpERT (Span-based Entity and Relation Transformer). A ideia central foi utilizar o BERT, através de técnicas de *prompt learning*, para rotular automaticamente um conjunto inicial de dados. Esse conjunto rotulado serviu então para o *fine-tuning* do modelo SpERT, que, por ser mais leve e poder ser executado localmente, garantiu a privacidade dos dados sensíveis e reduziu a dependência de APIs externas.

Os autores demonstraram a eficácia da abordagem proposta em um conjunto de dados reais de reclamações de consumidores, extraídas da plataforma Consumidor.gov.br [12]. Os resultados indicaram que o modelo SpERT, após o *fine-tuning* com os dados rotulados pelo BERT, apresentou um desempenho significativamente superior em comparação ao modelo treinado apenas com dados rotulados manualmente. Essa estratégia se mostrou promissora para cenários com escassez de dados rotulados, como é o caso de muitos problemas do mundo real, e abriu caminho para a aplicação de NER em domínios específicos com restrições de privacidade, como saúde, jurídico e governamental.

Monteiro e Zanchettin [23] analisaram as estratégias de otimização do BERT para o NER, explorando métodos como adaptação de domínio e técnicas de modelagem, como

model soups. Eles mostraram que a adaptação de domínio para o idioma português brasileiro melhora substancialmente o desempenho em tarefas de NER.

Para aprimorar o desempenho do NER, Monteiro e Zanchettin [23] aplicaram técnicas como *model soups*, em que modelos pré-treinados com diferentes conjuntos de hiperparâmetros têm seus pesos combinados. A abordagem reduziu a necessidade de retreinamento e custos computacionais típicos de *ensembles*, mostrando-se eficaz na melhora da precisão e na classificação de entidades em corpora de baixa disponibilidade. Além disso, a adaptação de domínio também foi investigada, com o modelo BERTimbau sendo ajustado para o domínio de auditoria pública no Brasil.

3.2.5 Modelos BERT e Large Language Models (LLMs)

A pseudonimização com IA evoluiu significativamente nos últimos anos, combinando modelos inteligentes, *frameworks* adaptáveis e diretrizes jurídicas para proteger dados sem comprometer a análise. Apesar dos desafios, as soluções estão mais maduras, e ferramentas modernas permitem tratar dados com cuidado e inteligência. A seguir, destacamos alguns trabalhos relacionados que abordam a pseudonimização de dados pessoais e como eles se comparam à nossa proposta do LEGAL-BERT-LGPD.

Yermilov et al. (2023) [24] investigaram a eficácia de diferentes técnicas de pseudonimização, desde substituições baseadas em regras até o uso de Modelos de Linguagem de Grande Porte (LLMs). Nossa proposta do LEGAL-BERT-LGPD se alinha a essa abordagem, porém empregando um modelo BERT especializado para a tarefa de NER, em vez da tarefa de classificação e sumarização explanados no trabalho de Yermilov. Outra diferença é por focar especificamente na conformidade com a LGPD, treinando o modelo com um corpus anotado do domínio jurídico brasileiro para aprimorar a identificação de dados relevantes para a legislação.

Recentemente, o uso de LLMs também tem sido investigado para a tarefa específica de pseudonimização. Os autores do estudo sobre *Cloaked Classifiers* Riabi et al. (2024) [25] exploraram estratégias de pseudonimização em tarefas de classificação, compartilhando um método para pseudonimizar manualmente um conjunto de dados multilíngue de radicalização, garantindo desempenho comparável aos dados originais. O estudo utilizou algumas estratégias de substituição, entre elas, a "*Category-Specific Placeholder (S2)*", que substitui entidades por rótulos específicos de categoria. Essa abordagem é aprofundada em nosso estudo no próximo capítulo, utilizando o nosso modelo treinado e o LLM para comparar o uso dessa estratégia. Porém nos diferenciamos ao utilizar mais rótulos no domínio brasileiro e conforme a LGPD, uma lei brasileira, análoga à GDPR europeia citada no artigo.

Além disso, a literatura existente enfatiza a importância de um *framework geral de pseudonimização*. Hou et al. (2025) [26] propuseram um framework com três componentes: detecção, geração e substituição. Nossa abordagem poderia se integrar a esse framework, com o LEGAL-BERT-LGPD atuando como o componente principal para a detecção de informações sensíveis no contexto da LGPD.

Vakili et al. (2024) [27] avaliaram os efeitos da pseudonimização de ponta a ponta em modelos clínicos BERT ajustados para cinco tarefas de PLN, revelando impacto mínimo no desempenho preditivo ao usar dados ajustados e pseudonimizados. Embora o domínio médico possua suas próprias particularidades e categorias de dados sensíveis, as lições aprendidas e os métodos de avaliação desenvolvidos no contexto clínico são relevantes para o nosso trabalho. Os autores mencionaram ataques de privacidade em modelos de linguagem, destacando que é possível recuperar informações confidenciais de dados de treinamento do modelo. Com isso, para o modelo LEGAL-BERT-LGPD, nós treinamos em um conjunto de dados público (Carolina conjunto de dados [4]), o qual contém dados revisados por estudos anteriores. Esse etapa será detalhada no próximo capítulo.

Sobre a questão da definição e categorização de informações pessoais, Szawerna et al. (2024) [28] argumentam pela necessidade de um sistema de "*tags*" universal para categorizar dados pessoais em diferentes domínios. O termo "*Tagset*" utilizado no estudo de Szawerna é equivalente ao termo "*rótulo*" utilizado neste trabalho. Os autores citaram os Tagsets mais comuns entre diversos domínios. Além disso, citaram também o trabalho de Pilan et al. (2021) [29] no domínio legal, que por sua vez citou alguns desafios de anonimização de texto. Nosso trabalho complementa essa proposta, iniciando uma discussão sobre a categorização de dados pessoais no contexto do Brasil e um conjunto de rótulo para pseudonimização.

O estudo de Lothritz et al. [30] de 2023 avaliou o impacto da desidentificação de texto no desempenho de nove tarefas de PNL, com foco na anonimização e pseudonimização de nomes de pessoas. Os autores compararam seis estratégias diferentes de anonimização, utilizando dois modelos pré-treinados de última geração: BERT e ERNIE. Apesar da relevância do estudo, eles não abordaram a tarefa de NER. Nosso estudo foca especificamente na tarefa de NER e também abordamos mais rótulos, além do rótulo relacionado a pessoa.

Em suma, o LEGAL-BERT-LGPD se diferencia dos métodos existentes ao propor um modelo BERT especializado e treinado no **domínio jurídico brasileiro** para a tarefa de pseudonimização, com o objetivo primordial de garantir a **conformidade com a LGPD**.

Capítulo 4

Desenvolvimento da Pesquisa e Resultados

Neste capítulo, abordaremos o desenvolvimento do modelo LEGAL-BERT-LGPD, assim como já demonstrado na Figura 1.2. Utilizaremos um Modelo NER Fundacional mais um Conjunto de Dados Jurídico em Português, e em seguida, vamos comparar o modelo LEGAL-BERT-LGPD e os LLMs em cenários práticos de utilização.

4.1 Dados Pessoais Utilizados

Para os Dados Pessoais, nós nos baseamos em alguns termos exibidos no já citado trabalho de Pilan et al. [29] no domínio legal. Os autores abordaram alguns termos como o "*Direct Identifier*" (Identificador Direto). O Identificador Direto é uma variável única para um indivíduo (um nome, endereço, número de telefone ou conta bancária) que pode ser utilizada para identificar diretamente o sujeito. Alinhado a isso, buscamos os itens mencionados no site [1] do MPF, e então, **selecionamos de forma livre alguns termos** para identificação de Dados Pessoais. Os termos utilizados e seus exemplos estão dispostos na Tabela 4.1.

4.2 Modelo Fundacional NER

Utilizamos um modelo BERT pré-treinado para a tarefa de NER em português, disponibilizado por Pierre [2]. Este modelo utilizou o conjunto de dados LenerBR derivado do trabalho [2] e foi aprimorado através de um *fine-tuning* do modelo *BERTimbau*. Porém nosso modelo disposto neste estudo alterou o nome de alguns rótulos e excluiu outros.

[1] `https://www.mpf.mp.br/servicos/lgpd/politicas/privacidade/aviso-de-privacidade-1`
[2] `https://huggingface.co/pierreguillou/ner-bert-large-cased-pt-lenerbr`

Tabela 4.1: Rótulos selecionados para identificação de Dados Pessoais. Ao lado, estão seus respectivos exemplos. CPF é um número de identidade brasileiro.

Rótulo	**Exemplo de Entidade**
NOME	*Francis Pantele*
DATA	*12 de Janeiro de 2013*
ENDERECO	*Campo Grande, MS*
CPF	*049.567.041-22*
TELEFONE	*(61) 9412 3333*
EMAIL	*fran@bol.com*
DINHEIRO	*5.534,00*
CEP	*59123-222*

Fizemos essas alterações para melhor adequar ao rótulos deste trabalho. As alterações mencionadas estão dispostas na Tabela 4.2 abaixo.

Tabela 4.2: Alterações dos rótulos para o novo modelo NER

Modelo Original	**Modelo Atual**	**Situação**
PESSOA	NOME	*Alterado*
TEMPO	DATA	*Alterado*
LOCAL	ENDERECO	*Alterado*
JURISPRUDENCIA	-	*Removido*
LEGISLACAO	-	*Removido*
ORGANIZACAO	-	*Removido*

4.3 Conjunto de dados Jurídico e em Português

O *Corpus Carolina* foi escolhido para este trabalho devido ao seu volume significativo de textos do domínio jurídico, altamente relevantes para a tarefa de pseudonimizar informações sensíveis em conformidade com a LGPD.

4.3.1 Extração e Transformação do conjunto de dados

Do corpus Carolina, utilizamos apenas a seção "*jud*" do conjunto de dados. Dividimos os dados em trechos de texto menores e extraímos entidades usando expressões regulares. Os passos abaixo descrevem o processo de extração de entidades:

1. **Passo 1**: Analisamos a seção "*jud*" do conjunto de dados (40.464 registros), que variava de 1.043 caracteres (325 *tokens*) a 26.012 caracteres (8.231 *tokens*).

2. **Passo 2**: Dividimos os registros em fragmentos de 512 *tokens* (para alinhar com os parâmetros do modelo fundacional), resultando em um total de 2.705.591 segmentos. Isso permitiu treinar o novo modelo, ajustado para 512 *tokens*.

3. **Passo 3**: Utilizamos o modelo NER Fundamental (Seção 4.2) para rotular os 2.705.591 segmentos.

4. **Passo 4**: Aplicamos também expressões regulares (Regex) para rotular "*CEP*", "*CPF*", "*TELEFONE*", "*EMAIL*", "*DINHEIRO*" e "*DATA*". Para "*DATA*", combinamos entidades encontradas tanto pelo modelo NER quanto pelo Regex.

Para balancear os 2.705.591 segmentos, extraímos uma amostra menor. Amostramos apenas o "*TELEFONE*", "*EMAIL*" e "*DINHEIRO*", resultando em um total de 25.145 entradas, enquanto mantivemos o restante, pois haviam poucos dados. O conjunto de dados final para treinamento é mostrado na Tabela 4.3.

Tabela 4.3: Balanceamento do Conjunto de Dados para Treinamento

Rótulo	**Conjunto Completo**	**Amostra**
NOME	3726	3726
DATA	9979	9979
ENDERECO	4857	4857
CEP	750	750
CPF	381	381
TELEFONE	25614	2559
EMAIL	8520	1183
DINHEIRO	182246	1710
Total	236073	25145

4.4 Treinamento do Modelo

O ambiente de treinamento foi uma máquina local com GPU de 24 GB de VRAM e processador Intel I9 de 10ª geração. Utilizamos a biblioteca Transformers (v4.44.2) e PyTorch (v2.4.0). Com 5 épocas, o modelo foi treinado em aproximadamente 2 min e 30 s. O modelo treinado possui cerca de 1,5 GB e está disponível no Hugging Face[3]. Ele pode ser usado sem exigir uma GPU dedicada, embora o tempo de execução possa aumentar.

Realizamos treinamento com diferentes valores de parâmetros, ajustando principalmente *logging_steps* (valores entre 200 e 300) e *num_train_epochs* (valores entre 3 e 5). Utilizamos "*seed=42*". Assim, os principais parâmetros e seus respectivos valores foram:

[3] `https://huggingface.co/celiudos/legal-bert-lgpd`

token_size=512, *num_train_epochs=5*, *logging_steps=200* e *per_device_train_batch_size=16*. Após vários testes, selecionamos os valores mostrados na Tabela 4.4.

Tabela 4.4: Parâmetros utilizados no treinamento do modelo

Parâmetro	**Valor**
token_size	512
logging_strategy	steps
num_train_epochs	5
logging_steps	200
save_total_limit	3
per_device_train_batch_size	16
per_device_eval_batch_size	32
learning_rate	2e-5
gradient_accumulation_steps	1
weight_decay	5

4.5 Resultados do treinamento do modelo

Com base no treinamento do novo modelo, geramos os resultados utilizando o F1-Score. Dividimos o conjunto de dados em treinamento (80%) e teste (20%) para verificar o desempenho. A validação ficou como parte do próximo capítulo, quando compararemos com os LLMs e a partir de um novo Dataset anotado manualmente. A Tabela 4.5 apresenta os resultados.

Tabela 4.5: Resultados de desempenho do modelo. Todos os valores foram arredondados para duas casas decimais. A Média é simples e por Entidade.

Entidade	**Precisão**	**Recall**	**F1-Score**	**Suporte**
NOME	0.95	0.95	0.95	1743
DATA	0.98	0.97	0.98	2024
ENDERECO	0.80	0.84	0.82	1323
CPF	0.98	1.00	0.99	144
TELEFONE	0.96	0.97	0.97	983
EMAIL	0.96	0.98	0.97	567
DINHEIRO	0.95	0.95	0.95	348
CEP	0.98	0.98	0.98	447
Média	0.94	0.96	0.95	947

Observamos que o modelo alcançou F1-Score geral de 95%, resultado satisfatório. A Tabela 4.5 mostra bom desempenho para todas as entidades, exceto *ENDERECO*, que obteve F1-Score de 82%. O modelo apresentou desempenho excelente para entidades como *DATA*, *CPF* e *CEP*, todas com F1-Scores acima de 98%.

Para a entidade *CPF*, que atingiu F1-Score de 99%, isso era esperado, pois CPF segue um padrão fixo e é mais fácil de identificar. Em contraste, o desempenho inferior na entidade *ENDERECO* pode ser atribuído à complexidade e variabilidade de endereços, que podem incluir diferentes formatos e estilos de escrita (por exemplo, "*Rua das Flores, n. 23, Jardim Rony, Guaratinguetá-SP*").

4.6 Comparações entre o modelo treinado e LLMs

Nesta seção, validamos o modelo a partir de um conjunto de dados inédito. Criamos um cenário de teste utilizando 105 arquivos. Baseados em documentos reais, substituímos seus dados pessoais por valores sintéticos. Os conjuntos de dados com os arquivos podem ser encontrados no Hugging Face[4]. Utilizamos esses arquivos em vez de conjuntos de dados públicos devido à falta de conjuntos de dados contendo dados pessoais alinhados aos rótulos definidos neste estudo. Agrupamos os arquivos em 11 categorias da área do Direito, seguindo a taxonomia utilizada pelo CNMP (Conselho Nacional do Ministério Público)[5]. Esses arquivos foram selecionados aleatoriamente e envolveram uma variedade de tópicos, incluindo casos cíveis, criminais e de saúde. A anotação de entidades foi realizada manualmente. Os arquivos de teste contêm informações sensíveis sintéticas para garantir a robustez do modelo. A Tabela 4.6 apresenta as características dos arquivos utilizados nos testes.

Tabela 4.6: Categorias de arquivos de teste e suas características. A coluna **Qtd. (n)** se refere à quantidade de arquivos por categoria.

ID	**Categorias**	**Qtd. (n)**	**Caracteres**	**Entidades**
1	ADMINISTRATIVO	16	59842	261
2	CIVIL	24	81235	665
3	CRIANÇA E DO ADOLESCENTE	7	25946	237
4	CONSUMIDOR	10	39294	276
5	TRABALHO	11	43005	211
6	ELEITORAL	2	8071	58
7	PENAL	7	27057	223
8	PREVIDENCIÁRIO	9	35929	152
9	PROCESSUAL CIVIL	9	38691	282
10	PROCESSUAL PENAL	7	28901	178
11	TRIBUTÁRIO	3	8643	42
	Total	105	396614	2585

[4] `https://huggingface.co/datasets/celiudos/corpus-synthetic-lgpd`
[5] `https://sgt.cnmp.mp.br/consulta_publica_assuntos.php`

Com os arquivos definidos, realizamos testes utilizando o modelo BERT (LEGAL-BERT-LGPD) e os LLMs GPT-4o mini (versão gpt-4o-mini-2024-07-18)[6] e GPT-4.1 (versão gpt-4.1-2025-04-14)[7]. Selecionamos o GPT-4o mini por ser um modelo pequeno, rápido e acessível, projetado para tarefas específicas. O GPT-4.1, por sua vez, é um modelo mais avançado no momento da escrita deste estudo, com maior capacidade de processamento e compreensão de linguagem natural. Para os testes remotos com modelos GPT, utilizamos *prompt engineering* com técnicas de *few-shot learning*. Os *prompts* utilizados para o GPT estão no Apêndice I.1. Os testes GPT foram conduzidos via API REST da plataforma, com os parâmetros "*temperature=0.6*" e "*seed=42*" para garantir reprodutibilidade.

Também realizamos comparações entre o modelo BERT e o LLM DeepSeek-R1[8]. O DeepSeek-R1 é um modelo de linguagem de código aberto. Utilizamos as versões "*DeepSeek-R1-Distill-Qwen-8B*" e "*DeepSeek-R1-Distill-Qwen-32B*" devido a limitações de hardware local. Os modelos utilizaram técnicas de quantização para reduzir o tamanho do modelo alocado em memória[9]. Além disso, foram executados utilizando o Ollama[10] com os parâmetros "*temperature=1.0*" e "*seed=42*". Para os testes, utilizamos *prompt engineering* com técnicas de *few-shot learning*. Os *prompts* utilizados para o DeepSeek-R1 estão no Apêndice I.2. Os testes locais com BERT e DeepSeek-R1 foram realizados em uma máquina com GPU de 24 GB de VRAM e processador Intel I9 de 10ª geração. O BERT consumiu 1,5 GB de VRAM, enquanto o DeepSeek-R1 consumiu 5 GB de VRAM (8B) e 21 GB de VRAM (32B).

Conforme mostrado na Tabela 4.7, em relação ao tempo de execução, o modelo BERT (GPU) superou todos os LLMs, com média de 0,138 segundos contra 2,754 s do LLM mais rápido (DeepSeek-R1 8B).

Já as Tabelas 4.8 e 4.9 apresentam os resultados de teste de Precisão, Recall e F1. Os resultados foram avaliados pela média dos valores, comparando o modelo BERT com os modelos DeepSeek-R1 e GPT respectivamente.

A Tabela 4.8 apresenta alguns padrões centrais. Primeiro, o *LEGAL-BERT-LGPD* apresenta desempenho bastante equilibrado: sua média de Precisão (0,693), Recall (0,725) e F1 (0,702) indica constância, com leve ênfase em Recall em relação a Precisão. Segundo, o *DeepSeek-R1 8B* mostra um viés oposto: atinge a maior Precisão média (0,804) mas sacrifica severamente o Recall (0,435), resultando no pior F1 global (0,538) — nota-se o caso extremo do ID 6, onde a Precisão chega a 0,981 enquanto o Recall desaba para

[6] `https://platform.openai.com/docs/models/gpt-4o-mini`
[7] `https://platform.openai.com/docs/models/gpt-4.1`
[8] `https://github.com/deepseek-ai/DeepSeek-R1`
[9] `https://apxml.com/posts/gpu-requirements-deepseek-r1`
[10] `https://ollama.com/library/deepseek-r1`

Tabela 4.7: Resultados de teste com LEGAL-BERT-LGPD (GPU e CPU) e LLMs para tempo de execução em segundos. A coluna **ID** se refere às categorias da Tabela 4.6. A coluna ***n*** (número de arquivos) foi utilizada como peso para médias ponderadas. As médias são ponderadas com base na quantidade n de arquivos por categoria.

ID	**n**	**LEGAL-BERT-LGPD**		**DeepSeek-R1**		**GPT**	
		GPU	**CPU**	**8B**	**32B**	**4o-mini**	**4.1**
1	16	0.135	1.523	2.549	9.232	2.787	2.391
2	24	0.125	1.389	3.106	11.002	3.447	3.348
3	7	0.133	1.429	2.773	10.549	3.220	2.915
4	10	0.146	1.670	3.481	12.976	3.771	3.029
5	11	0.155	1.595	2.319	10.226	2.535	2.238
6	2	0.130	1.342	2.856	11.973	3.088	4.370
7	7	0.141	1.509	2.876	12.603	3.901	2.892
8	9	0.139	1.510	2.422	10.230	2.899	2.472
9	9	0.158	1.784	2.717	9.718	3.159	2.885
10	7	0.140	1.529	2.231	7.981	2.916	2.427
11	3	0.113	1.368	2.128	8.902	3.616	1.741
Média	105	0.138	1.521	2.754	10.496	3.188	2.793

0,369. Por fim, o *DeepSeek-R1 32B* comprova o ganho de escala: mantém Precisão elevada (0,846) e, sobretudo, recupera parte do Recall perdido (0,654), superando levemente o *LEGAL-BERT-LGPD* no F1 médio (0,721 x 0,702). Porém mesmo sem igualar o pico de F1 do *DeepSeek-R1 32B*, o *LEGAL-BERT-LGPD* oferece um balanço ótimo entre cobertura, estabilidade e eficiência, o que o torna particularmente atraente para aplicações produtivas em conformidade com a legislação de proteção de dados.

Conforme exposto na Tabela 4.9, os modelos *GPT-4* apresentam ganhos substanciais de desempenho — F1 médio de 0,817 para o *GPT-4o-mini* e 0,825 para o *GPT-4.1*, contra 0,702 do *LEGAL-BERT-LGPD*. Já o *GPT-4o-mini* alcança a maior precisão média (0,931) e um F1 sólido (0,817), sacrificando relativamente pouco Recall (0,747). Por fim, o *GPT-4.1* entrega o melhor F1 global (0,825) ao equilibrar precisão (0,885) com o maior Recall (0,797), embora exija hardware ou infraestrutura em nuvem mais potente, o que pode aumentar custos e latência.

Todavia, alguns fatores atestam a competitividade (e, em cenários regulatórios, a vantagem prática) do *LEGAL-BERT-LGPD*. Primeiro, a cobertura normativa dedicada: mesmo exibindo aproximadamente 12 pontos percentuais a menos em F1, o modelo jurídico atinge Recall médio de 0,725 sem recorrer a instruções externas ou engenharia de*prompt*(um valor apenas 2 pontos percentuais inferior ao *GPT-4o-mini*), sugerindo que seu pré-treino específico em corpus da LGPD já internaliza boa parte da semântica regulatória necessária. Segundo, a eficiência e soberania de dados: a arquitetura BERT pode ser executada localmente em GPUs de baixo custo, ou até mesmo CPUs, elimi-

Tabela 4.8: Resultados de pontuação com LEGAL-BERT-LGPD e modelos DeepSeek-R1. A coluna **ID** se refere às categorias da Tabela 4.6. A coluna ***n***(número de arquivos) foi utilizada como peso para médias ponderadas. As colunas **Prec.** e **Rec.** se referem respectivamente à Precisão e Recall. As médias são ponderadas com base na quantidade n de arquivos por categoria.

ID	**n**	**LEGAL-BERT-LGPD**			**DeepSeek-R1 8B**			**DeepSeek-R1 32B**		
		Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
1	16	0.587	0.746	0.650	0.748	0.392	0.458	0.762	0.668	0.686
2	24	0.713	0.713	0.710	0.833	0.543	0.650	0.848	0.657	0.729
3	7	0.755	0.801	0.775	0.736	0.294	0.410	0.879	0.611	0.714
4	10	0.707	0.717	0.710	0.884	0.512	0.630	0.867	0.732	0.784
5	11	0.693	0.588	0.621	0.776	0.368	0.467	0.877	0.596	0.692
6	2	0.741	0.827	0.781	0.981	0.369	0.470	0.939	0.740	0.828
7	7	0.710	0.778	0.741	0.929	0.442	0.595	0.881	0.705	0.780
8	9	0.668	0.707	0.682	0.706	0.396	0.481	0.749	0.644	0.679
9	9	0.675	0.724	0.691	0.840	0.426	0.551	0.914	0.576	0.698
10	7	0.797	0.753	0.772	0.848	0.397	0.522	0.902	0.552	0.665
11	3	0.723	0.868	0.787	0.547	0.385	0.421	0.839	0.925	0.876
Média	105	0.693	0.725	0.702	0.804	0.435	0.538	0.846	0.654	0.721

nando latência de API, dependência de fornecedor e exposição de informações sensíveis, requisitos frequentes em fluxos legais. E por último, a estabilidade operacional: como já dito anteriormente, o *LEGAL-BERT-LGPD* possui uma estabilidade em seus resultados, indicando comportamento previsível entre documentos, um atributo valioso quando se busca consistência auditável.

Após os testes, realizamos análise estatística para verificar a significância dos resultados. Criamos um ranking de 1 a 5 (para cinco modelos) baseado nos F1s de cada modelo nos arquivos de teste. Aplicamos o teste de Friedman, seguido do teste de Nemenyi para comparações múltiplas. O teste de Nemenyi foi escolhido pois é adequado para comparações pós-hoc múltiplas sem grupo de controle explícito. A Figura 4.1 apresenta os resultados.

Com base na Figura 4.1, percebemos que os resultados do teste pós-hoc revelaram três faixas de desempenho. **(i)** O *GPT-4.1* lidera o ranking e apresenta diferença estatisticamente significativa em relação a todos os demais modelos. **(ii)** O *GPT-4o-mini* compartilha a elite com o *GPT-4.1*, mas difere do *DeepSeek-R1 8B* e do *LEGAL-BERT-LGPD*. **(iii)** Forma-se ainda um patamar intermediário composto por *DeepSeek-R1 32B* e *LEGAL-BERT-LGPD*, cuja discrepância interna não excede a diferença crítica. Ambos, entretanto, superam o *DeepSeek-R1 8B*, último colocado no ranking.

A partir das análises anteriores, observa-se que o *LEGAL-BERT-LGPD* destaca-se em ambientes locais pela eficiência, exigindo apenas 1,5 GB de VRAM contra 21 GB

Tabela 4.9: Resultados de pontuação com LEGAL-BERT-LGPD e modelos GPT. A coluna **ID** se refere às categorias da Tabela 4.6. A coluna ***n*** (número de arquivos) foi utilizada como peso para médias ponderadas. As colunas **Prec.** e **Rec.** se referem respectivamente à Precisão e Recall. As médias são ponderadas com base na quantidade n de arquivos por categoria.

ID	**n**	**LEGAL-BERT-LGPD**			**GPT-4o-mini**			**GPT-4.1**		
		Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
1	16	0.587	0.746	0.650	0.845	0.771	0.793	0.784	0.807	0.774
2	24	0.713	0.713	0.710	0.955	0.774	0.846	0.899	0.821	0.849
3	7	0.755	0.801	0.775	0.935	0.698	0.787	0.917	0.769	0.827
4	10	0.707	0.717	0.710	0.955	0.749	0.830	0.906	0.826	0.858
5	11	0.693	0.588	0.621	0.960	0.648	0.759	0.956	0.656	0.764
6	2	0.741	0.827	0.781	0.946	0.768	0.848	0.944	0.797	0.864
7	7	0.710	0.778	0.741	0.949	0.798	0.864	0.957	0.840	0.893
8	9	0.668	0.707	0.682	0.885	0.713	0.780	0.781	0.768	0.757
9	9	0.675	0.724	0.691	0.937	0.750	0.824	0.873	0.845	0.853
10	7	0.797	0.753	0.772	0.977	0.719	0.817	0.918	0.759	0.820
11	3	0.723	0.868	0.787	0.975	0.893	0.926	0.975	0.978	0.976
Média	105	0.693	0.725	0.702	0.931	0.747	0.817	0.885	0.797	0.825

necessários pelo *DeepSeek-R1 32B*, tornando-o viável em máquinas comuns. Além dos requisitos reduzidos de hardware, o *LEGAL-BERT-LGPD* não apresenta tendência a alucinações como os modelos generativos (*DeepSeek* e *GPT*). Em termos de desempenho, também oferece tempos de resposta significativamente menores.

4.7 Limitações

É importante enfatizar que os arquivos de teste criados, conforme definidos na Tabela 4.6, foram analisados sem rigor jurídico estrito e sem validação por especialista do domínio, podendo conter vieses. Portanto, os resultados devem ser interpretados com cautela. Além disso, devido a restrições de custo e tempo, a análise foi realizada com número limitado de arquivos, que pode não representar todo o domínio jurídico brasileiro. Também não estendemos a avaliação para validar a efetividade da pseudonimização; apenas validamos a acurácia na identificação de entidades sensíveis e a comparamos com outros modelos.

Embora este estudo tenha usado as versões *DeepSeek-R1* (8B e 32B), reconhece-se que variantes mais recentes, como as versões 70B e 671B, fornecem respostas de maior qualidade. Entretanto, exigem infraestrutura mais robusta, como GPUs com mais de 24 GB de VRAM ou clusters, impraticáveis para aplicações locais de baixo custo. Assim, apesar de possíveis melhorias de desempenho, o uso desses modelos permanece limitado por barreiras operacionais e de infraestrutura.

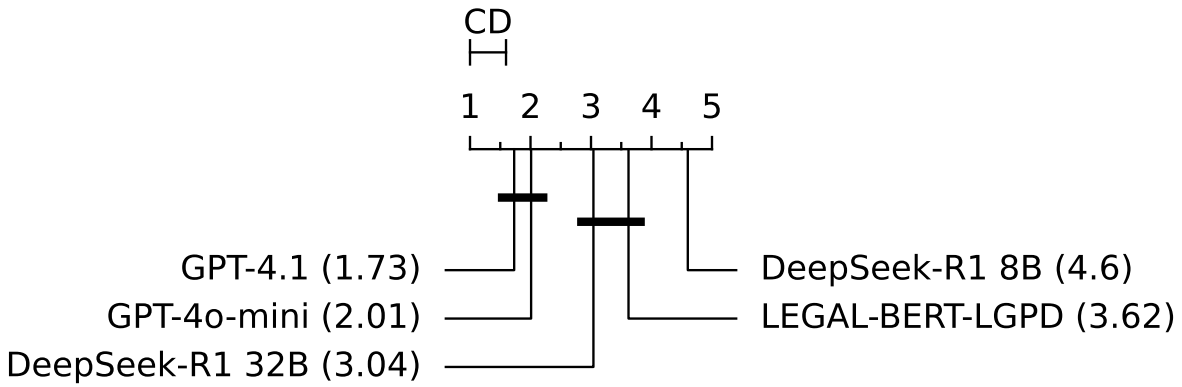

Figura 4.1: Teste pós-hoc de Nemenyi para os 5 modelos e 105 arquivos quanto ao ranking de F1. O eixo horizontal que indica a posição média (quanto menor, melhor), enquanto a barra superior representa a Diferença Crítica (CD = 0,595; alpha = 0,05). Segmentos pretos que não se sobrepõem identificam pares de modelos com desempenho estatisticamente distinto. Os valores entre parênteses indicam os valores das posições do ranking

Entendemos que o *DeepSeek-R1 32B* opera em modo de raciocínio, o que impacta diretamente seu tempo de resposta. Reconhecemos, também, que existem modelos de raciocínio da OpenAI, como *o4-mini* e *o3*, que não foram incluídos neste estudo. Optamos pelo melhor modelo DeepSeek-R1 que poderia ser executado com o hardware disponível, bem como pelos modelos de melhor custo-benefício da OpenAI disponíveis no momento. Pode-se inferir que variantes sem raciocínio do DeepSeek, como *DeepSeek V3*, tenderiam a apresentar desempenho pior que o *DeepSeek-R1 32B* em nossos testes, enquanto modelos de raciocínio da OpenAI tenderiam a superar os modelos usados.

Conforme mostrado na Tabela 4.7, a pontuação definida para os modelos *DeepSeek-R1* levam em consideração o hardware disponível, o que não é necessariamente amplamente representativo. Da mesma forma, os tempos para os LLMs *GPT Models* na nuvem consideram o tempo de resposta HTTP e de execução do modelo. Portanto, os tempos dos LLMs podem variar de acordo com a latência de rede e o tempo de resposta do servidor.

A *prompt engineering* utilizada nos LLMs testados para obter os resultados apresentados nas Tabelas 4.8 e 4.9 buscou resultados satisfatórios, mas não foi otimizada para obter o melhor desempenho possível. Logo, os resultados devem ser considerados como uma referência inicial, e não como uma avaliação definitiva. Reconhecemos que há outros LLMs que potencialmente entregariam resultados melhores. Todavia, o objetivo deste artigo não é comparar todos os modelos disponíveis, mas sim apresentar uma comparação entre um modelo especializado e modelos de uso geral.

Capítulo 5

Conclusão e Trabalhos Futuros

Neste estudo, comparamos dois paradigmas populares para reconhecimento de entidades nomeadas (NER): os modelos baseados em BERT e os Large Language Models (LLMs). O primeiro exige fine-tuning supervisionado, mas oferece boa exatidão em domínios específicos e baixo custo de inferência, enquanto o último utiliza engenharia de prompts, podendo mapear diferentes rótulos, mesmo sem dados anotados, mas com maior custo computacional. A escolha entre estes dois depende do cenário de uso: tarefas especializadas se beneficiam de modelos menores treinados, enquanto contextos com múltiplas tarefas ou dados escassos favorecem os LLMs.

O custo-benefício pesa a favor do LEGAL-BERT-LGPD em muitos casos. Ele pode ser treinado e executado localmente, mesmo em *hardware* com bom custo-benefício, enquanto LLMs exigem infraestrutura com custos mais elevados. Isso torna o modelo atrativo para **projetos com orçamento limitado** ou que exigem **privacidade** e **controle local** dos dados. O LEGAL-BERT-LGPD domina em velocidade e dispensa custos de API, além de eliminar a necessidade de transmitir dados sensíveis para servidores de terceiros, uma exigência que se alinha a LGPD. LLMs como o GPT-4.1 agregam ganhos práticos de F1 e maior versatilidade de linguagem, mas impõem maior latência, custos variáveis e um vetor adicional de exposição de dados. Para fluxos críticos, especialmente em tribunais ou órgãos públicos, o LEGAL-BERT-LGPD local surge como a solução mais segura e economicamente previsível.

Em resumo, embora os LLMs sejam poderosos e flexíveis, modelos como o BERT continuam sendo altamente relevantes em 2025. Sua eficiência, interpretabilidade e adaptabilidade os tornam ideais para tarefas de classificação de tokens em domínios específicos. Em muitos casos, a melhor estratégia é usar os dois tipos de modelo de forma complementar, aproveitando a especialização do BERT e a generalização dos LLMs para maximizar desempenho e reduzir esforço de desenvolvimento.

5.1 Trabalhos Futuros

Para trabalhos futuros, sugerimos explorar a combinação de BERT e LLMs em um único processo, assim como no trabalho de Hou et al., em que o BERT poderia ser usado para tarefas específicas e os LLMs para tarefas mais gerais. Além disso, a pesquisa sobre técnicas de engenharia de prompts para melhorar a precisão dos LLMs em tarefas específicas é uma área promissora. Outras linhas possíveis também seriam o fine-tuning de modelos BERT de outras variantes, como o XLM-ROBERTA, que vem apresentando bons resultados. Assim como utilizar mais dados para treinamento e avaliação, para com isso, aumentar a precisão do modelo.

Referências

[1] Keraghel, Imed, Stanislas Morbieu e Mohamed Nadif: *A survey on recent advances in named entity recognition*. arXiv preprint arXiv:2401.10825, 2024. x, 14, 15, 19, 21

[2] Araujo, Pedro Henrique Luz de, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto e Paulo Bermejo: *Lener-br: a dataset for named entity recognition in brazilian legal text*. Em *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, páginas 313–323. Springer, 2018. xi, 14, 15, 21, 22, 30

[3] Mauricio, Antonio, Vladia Pinheiro, Vasco Furtado, João Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André Câmara Ferreira da Costa, Raquel Silveira e Nilsiton Aragão: *Cdjur-br–a golden collection of legal document from brazilian justice with fine-grained named entities*. arXiv preprint arXiv:2305.18315, 2023. xi, 14, 15, 22, 23, 24

[4] Finger, Marcelo, Maria Clara Paixão de Sousa, Cristiane Namiuti, Vanessa Martins do Monte, Aline Silva Costa, Felipe Ribas Serras, Mariana Lourenço Sturzeneker, Raquel de Paula Guets, Renata Morais Mesquita, Guilherme Lamartine de Mello, Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Patrícia Brasil, Mariana Marques da Silva e Mayara Feliciano Palma: *Carolina: The open corpus for linguistics and artificial intelligence*. `https://sites.usp.br/corpuscarolina/corpus`, 2022. Version 1.1 (Ada). xi, 25, 29

[5] Caseli, Helena de Medeiros e Maria das Graças Volpe Nunes (editores): *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3ª edição, 2024, ISBN 978-65-01-20581-6. `https://brasileiraspln.com/livro-pln/3a-edicao/`. 1

[6] Dados Pessoais, Proteção de: *Lgpd*. 2020. 1, 11

[7] Nunes, Luiz Fernando Pereira e José Carlos Francisco dos Santos: *Lgpd–uma visão de tecnologia e agnóstica*. Revista Direito & Paz, 2(49):218–237, 2023. 1

[8] MPF: *Sobre o mpf*, 2023. `http://www.mpf.mp.br/o-mpf/sobre-o-mpf`, Acesso em: 1 de Fevereiro de 2023. 1, 5

[9] MPF: *Conheça a estrutura do mpf*, 2023. `https://www.mpf.mp.br/o-mpf/sobre-o-mpf/conheca-o-mpf-1`, Acesso em: 1 de Fevereiro de 2023. 6

[10] CNMP: *Mapa estratégico*, 2023. https://www.mpf.mp.br/o-mpf/sobre-o-mpf/gestao-estrategica-e-modernizacao-do-mpf/planejamento-estrategico/planejamento-estrategico-2022-2027/mapaestrategicoMPF20222027.pdf, Acesso em: 1 de Fevereiro de 2023. 7

[11] MPF: *O que é a lgpd?*, 2024. https://www.mpf.mp.br/servicos/lgpd/o-que-e-a-lgpd, Acesso em: 1 de Fevereiro de 2024. 12

[12] Andrade, Claudio MV de, Celso França, Fabiano Belém, Gabriel Jallais, Marcelo AS Ganem, Gabriel Texeira, Alberto HF Laender e Marcos A Gonçalves: *Promptner: Uma abordagem para reconhecimento de entidades nomeadas em dados sensíveis a partir de instâncias rotuladas automaticamente.* Em *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, páginas 269–281. SBC, 2023. 15, 27

[13] Powers, David M. W.: *Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation.* Journal of Machine Learning Technologies, 2(1):37–63, 2011. 15

[14] Oliveira, Frank Ned Santa Cruz de: *Gestão de riscos no direito fundamental à privacidade de dados pessoais no processo judicial eletrônico/diário de justiça eletrônico.* 2020. 17

[15] Rapôso, Cláudio Filipe Lima, Haniel Melo de Lima, Waldecy Ferreira de Oliveira Junior, Paola Aragão Ferreira Silva e Elaine Elaine de Souza Barros: *Lgpd-lei geral de proteção de dados pessoais em tecnologia da informação: Revisão sistemática.* RACE-Revista de Administração do Cesmac, 4:58–67, 2019. 17, 18

[16] Zhang, Dawen, Boming Xia, Yue Liu, Xiwei Xu, Thong Hoang, Zhenchang Xing, Mark Staples, Qinghua Lu e Liming Zhu: *Privacy and copyright protection in generative ai: A lifecycle perspective.* Em *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, páginas 92–97, 2024. 18, 19

[17] Ferreira, Juliano Rodrigues: *Aplicação da lei geral de proteção de dados com utilização de modelos de anonimização de.* 2023. http://repositorio.unb.br/handle/10482/47940, Acesso em: 1 de Setembro de 2024. 19

[18] Castilho, Richard Eckart de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank e Chris Biemann: *A web-based tool for the integrated annotation of semantic and syntactic structures.* Em Hinrichs, Erhard, Marie Hinrichs e Thorsten Trippel (editores): *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, páginas 76–84, Osaka, Japan, dezembro 2016. The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4011/. 21

[19] Sturzeneker, Mariana, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte e Cristiane Namiuti: *Carolina's methodology: building a large corpus with provenance and typology information.* Em *DHandNLP@ PROPOR*, páginas 53–58, 2022. 25

[20] Oliveira, Vitor, Gabriel Nogueira, Thiago Faleiros e Ricardo Marcacini: *Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents.* Artificial Intelligence and Law, páginas 1–21, 2024. 26

[21] Villena, Fabián, Luis Miranda e Claudio Aracena: *llmner:(zero/ few)-shot named entity recognition, exploiting the power of large language models.* arXiv preprint arXiv:2406.04528, 2024. 26, 27

[22] Devlin, Jacob: *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018. 27

[23] Monteiro, Monique e Cleber Zanchettin: *Optimization strategies for bert-based named entity recognition.* Em *Brazilian Conference on Intelligent Systems*, páginas 80–94. Springer, 2023. 27, 28

[24] Yermilov, Oleksandr, Vipul Raheja e Artem Chernodub: *Privacy-and utility-preserving nlp with anonymized data: a case study of pseudonymization.* arXiv preprint arXiv:2306.05561, 2023. 28

[25] Riabi, Arij, Menel Mahamdi, Virginie Mouilleron e DjamÃŠ Seddah: *Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks.* arXiv preprint arXiv:2406.17875, 2024. 28

[26] Hou, Shilong, Ruilin Shang, Zi Long, Xianghua Fu e Yin Chen: *A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation.* arXiv preprint arXiv:2502.15233, 2025. 29

[27] Vakili, Thomas, Aron Henriksson e Hercules Dalianis: *End-to-end pseudonymization of fine-tuned clinical bert models: Privacy preservation with maintained data utility.* BMC Medical Informatics and Decision Making, 24(1):162, 2024. 29

[28] Szawerna, Maria Irena, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan Son Vu e Elena Volodina: *Pseudonymization categories across domain boundaries.* Em *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, páginas 13303–13314, 2024. 29

[29] Lison, Pierre, Ildikó Pilán, David Sánchez, Montserrat Batet e Lilja Øvrelid: *Anonymisation models for text data: State of the art, challenges and future directions.* Em *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 4188–4203, 2021. 29, 30

[30] Lothritz, Cedric, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre e Anne Goujon: *Evaluating the impact of text de-identification on downstream nlp tasks.* Em *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, páginas 10–16, 2023. 29

Apêndice A

Fichamento de Artigo Científico

Pseudonymization in Legal Texts according to the LGPD: A Named Entity Recognition Approach

Marcelo Anselmo[1][0009−0005−1145−3501] and Bruno César Ribas[2][0000−0001−6314−1511]

[1] Department of Computer Science, University of Brasilia, Brasilia, Brazil
marcelo.anselmo@aluno.unb.br
[2] Department of Computer Science, University of Brasilia, Brasilia, Brazil
bruno.ribas@unb.br

Abstract. This study explores the application of Named Entity Recognition (NER) for the pseudonymization of data in legal texts, aiming to protect Personally Identifiable Information (PII) in compliance with Brazil's General Data Protection Law (LGPD). The research highlights the challenge of balancing data privacy and utility, presenting a methodology that uses artificial intelligence technologies to effectively identify and obscure PII in legal documents. In this study, we propose a Transformer model along with Regex techniques to identify entities in a text. To test the model, we created a new dataset from the existing LenerBR. We also used a function and prompt engineering applied to the Llama 8B version 3 model to generate synthetic data. Tests showed the need for further adjustments in the proposed new model. Future work will focus on improving the model's accuracy and efficiency, as well as enhancing the identification of sensitive data and learning from user interactions.

Keywords: Data Pseudonymization · Named Entity Recognition · Information Privacy · Legal Texts · LGPD · Transformer.

1 Introduction

In today's digital age, where technology and computing power have grown exponentially, personal data has become an extremely valuable asset. Companies capture this information from online interactions, such as websites and social networks, to create detailed user profiles. These profiles range from consumer habits to sensitive information like health conditions and personal preferences. This practice not only allows for the personalization of online experiences, but it also exposes users to privacy risks, including the illegal sale of such data in dark markets, potentially for use in criminal activities [3].

This scenario highlights the critical importance of protecting personal data, not only for individual privacy, but also for the overall security of society. The solution to these issues has manifested through regulations and laws dedicated to data protection, such as the General Data Protection Regulation (GDPR) in the European Union and the General Data Protection Law (LGPD) in Brazil [9].

The main problem addressed in this work is the difficulty of balancing individual privacy with the need to use their information in legal and administrative texts. This issue deepens when dealing with identifying and hiding PII without compromising data integrity and utility. Using resources, such as Transformer models and Regex techniques, an effective balance between privacy and utility is sought, as already accomplished in other studies, such as the work of G. M. GR, S. Abhi, and R. Agarwal [5].

The aim of this study is to investigate and demonstrate the effectiveness of Named Entity Recognition (NER) applied to PII pseudonymization in legal documents. It proposes to develop and test a model capable of efficiently identifying and hiding named entities that correspond to PII. The Figure 1 illustrates the pipeline adopted in our methodology. Another important note is the CNMP Resolution No. 281, dated December 12, 2023 [3], in its Article 79, presents pseudonymization as an alternative to protecting the personal data of natural persons in procedures or processes within the Brazilian Public Prosecutor's Office.

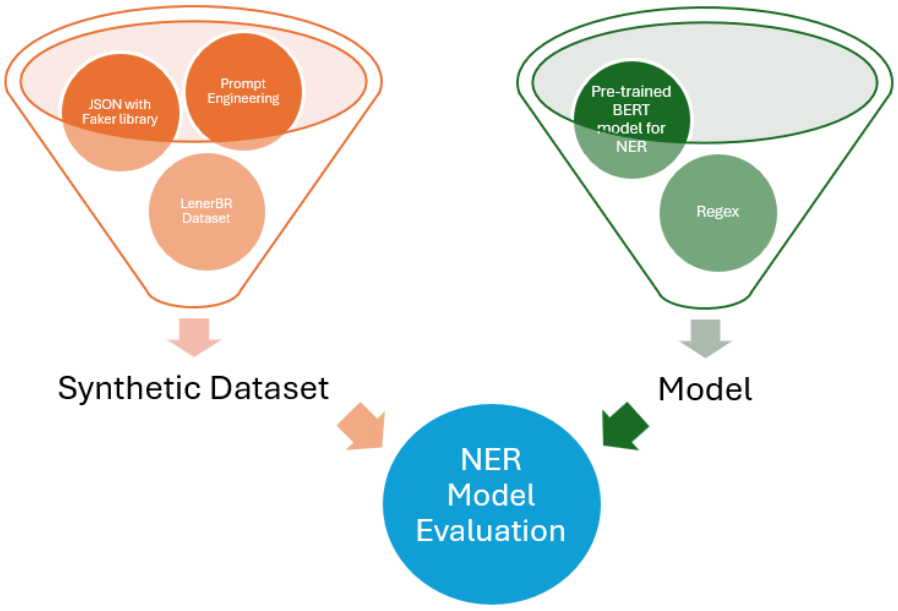

Fig. 1. Pipeline adopted in our methodology

Legal texts are often lengthy and complex, prioritizing formality over readability [12]. However, the implementation of advanced techniques, such as ar-

[3] https://www.cnmp.mp.br/portal/images/CALJ/resolucoes/Resoluo-281-de-2023.pdf

tificial intelligence models, often relies on large volumes of annotated data for training, which can be challenging in terms of cost and time [13]. With this in mind, we face a fundamental challenge: the need for data to be reliably and efficiently labeled. Traditionally, this labeling is done manually, a process that, despite improving the accuracy of artificial intelligence models, is notoriously time-consuming and expensive [4]. This scenario imposes a high cost in terms of time, money, and effort, limiting the scalability of solutions.

This paper is structured as follows: Section 2 presents the background concepts, covering fundamental concepts, like the related to personal data protection and pseudonymization. Section 3 describes related work, highlighting previous studies that addressed similar issues. Section 4 presents the proposed methodology, detailing the data used, tools, and terms to be found. Section 5 discusses the expected results for each test set. Finally, Section 6 presents the conclusions and suggestions for future work.

2 Background Concepts

2.1 General Data Protection Law (LGPD)

The General Data Protection Law (LGPD[4], Law No. 13,709, dated August 14, 2018) of Brazil is a regulation that establishes guidelines for the collection, use, processing, and storage of personal data. The law grants individuals more control over their personal information while balancing privacy rights with technological advancement. The LGPD requires that any operation involving personal data be based on clear legal justifications and respect strict principles, such as purpose, adequacy, and necessity [3].

2.2 Personal Data (PD)

Personal data (LGPD, Art. 5, I) refers to information that identifies or can directly or indirectly identify a natural person. This definition is broad and includes a variety of types of information, ranging from names and digital identities to physical and electronic characteristics, such as photos and location data. Such data is crucial in digital society, being used in various contexts, from identity verification to service personalization. However, handling it involves significant privacy risks, which requires robust protection measures to ensure data security and confidentiality [5].

2.3 Anonymization vs. Pseudonymization

Anonymization and pseudonymization are two fundamental concepts in data privacy management. Anonymization (LGPD, Art. 5, XI) refers to the process

[4] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
[5] https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-pessoais-lgpd

whereby personal data is processed to remove the ability to permanently associate this data with a specific individual, without using additional information. This means that once anonymized, the data cannot be linked back to the owner, even with the use of reasonable technical means. This process is irreversible, ensuring a high level of protection for the individual's identity.

On the other hand, pseudonymization (LGPD, Art. 13, § 4) is a technique that modifies personal data so that identification of the owner cannot be done without using additional information, which is maintained separately and under strict security measures by the data controller. Unlike anonymization, pseudonymization is reversible, provided that the additional information is made available [6]. This method will be adopted in this work as it allows the data to be protected without losing its utility for analysis and research.

2.4 Retrieval-Augmented Generation (RAG)

RAG is an advanced approach in natural language processing (NLP) models that integrates information retrieval with text generation [6]. This technique allows the model to retrieve relevant documents that are used to inform subsequent text generation. The unique ability of RAG to combine parametric and non-parametric memory enables it to adapt both retrieval and content generation for specific NLP tasks, improving the accuracy and relevance of the generated information [7]. By utilizing RAG, we can incorporate a vast amount of pre-existing legal information to enhance the quality of our generated dataset.

3 Related Work

In the article by Patsakis and Lykousas (2023) [11], the authors explore the challenge of effectively anonymizing text in the age of large language models. This study questions the effectiveness of these approaches and assesses their ability to prevent identification, especially with the use of AI on large datasets. An experiment is conducted using GPT on anonymized texts of well-known personalities to verify whether these language models can re-identify individuals. They argue that despite technological advances, there are still significant obstacles in protecting data privacy in texts processed by these tools.

In the work of de Andrade et al. (2023) [1], the authors present an approach called PromptNER, which focuses on NER in sensitive data using automatically labeled instances. The authors propose an approach that employs LLMs to identify entities in complaints and then trains simpler models like the SpERT method. The improved NER model shows substantial improvements in F-score, ranging from 41% to 129% compared to models using only manually labeled data. This study is crucial because it combines artificial intelligence methods to improve efficiency in identifying and handling personal data.

[6] https://www.cnmp.mp.br/portal/images/CALJ/resolucoes/Resoluo-281-de-2023.pdf

[7] https://huggingface.co/transformers/model_doc/rag.html

Mota et al. (2021) [7] investigate the use of neural networks for named entity recognition in legal documents in Portuguese. The authors used Spacy and FLAIR libraries. Their results demonstrate the ability of these advanced technologies to handle the linguistic complexity of Portuguese, providing a technical foundation for the development of more effective solutions.

Nunes and dos Santos (2023) [9] discuss the application and impact of the LGPD in the Brazilian context, focusing on a technological and agnostic approach. They emphasize the need for organizations to adapt to legal requirements and the importance of implementing effective measures for protecting personal data. This study highlights that implementing the LGPD requires a comprehensive technical and organizational approach. Multiples tools, such as Data Management Systems (DMS), anonymization, pseudonymization, encryption, and auditing, are highlighted as essential for compliance. Techniques like Big Data and Machine Learning can improve compliance, while privacy by design and eDiscovery are emphasized as crucial for protecting data.

GR, Shinu, and Agarwal (2023) [5] presented a hybrid model that combines RegEx and NER for resume analysis and matching. This work illustrates the applicability of NER techniques along with regular expressions to efficiently extract and process information in structured documents like resumes. The proposed methodology is relevant to our study as it demonstrates the effectiveness of integrating NER and RegEx in data identification and pseudonymization tasks.

Luz de Araujo et al. (2018) [2] developed LeNER-BR, a dataset specifically for Named Entity Recognition in Brazilian legal texts. Their results provide an essential foundation for validating NER models in the Brazilian legal context. The dataset they proposed serves as a valuable resource for training and testing new methodologies, including the approach of our study.

Oliveira et al. (2024) [10] explored the combination of prompt-based language models and weak supervision to label the task of Named Entity Recognition (NER) in legal documents. They highlighted the effectiveness of their techniques in improving the accuracy and automation of the NER process, which is crucial for data pseudonymization in compliance with privacy regulations.

4 Methodology

4.1 Personal Data Used

We categorized terms according to some types of Personal Data (PD). It is important to note that the spectrum of terms is much broader than those discussed in this study, however, we highlight some popular terms. Note that the CPF is a Brazilian ID number. The terms used to generate synthetic data are presented in Table 1 below.

Table 1. Selected Terms for Identification of Personal Data (PD)

Term	**Example**
NAME	*João da Silva*
DATE	*12 de janeiro de 2013*
ADDRESS	*Rua do Alecrim, 0*
CPF	*123.456.789-00*
PHONE	*(11) 99999-9999*
EMAIL	*example@example.com*
MONEY	*R$ 1,000.00*
ZIP CODE	*12345-678*

4.2 Dataset

For this study, we used the LenerBR dataset derived from the study by Luz de Araujo and Pedro Henrique [2] and available at [8]. This dataset consists exclusively of legal documents. It includes labels for people, places, temporal entities, and organizations, as well as specific tags for legal entities and judicial processes.

We used the functions of the Faker library [9] to generate synthetic data. The *Faker* library is a Python tool that generates fake data as per the function used. It is useful for creating synthetic data for testing and prototyping. We created a Python function that generates random data for one to two items per term using the *Faker* library. Figure 2 below show examples of data generated by our function.

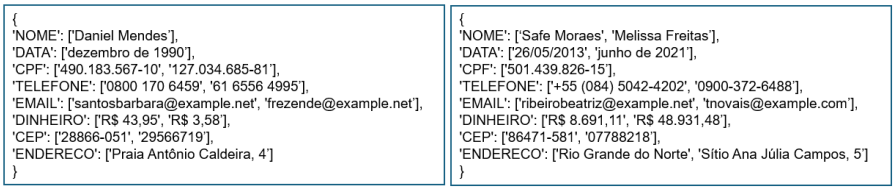

Fig. 2. Example of data generated by running our function twice.

The entities in portuguese and its respective translation to english, hidden the entities that has the same translation, are: "*NOME*" (NAME), "*DATA*" (DATE), "*ENDERECO*" (ADDRESS), "*TELEFONE*" (PHONE), "*DINHEIRO*" (MONEY), "*CEP*" (ZIP CODE).

As seen in the previous images, we generated synthetic data for the term CPF. For this term, being a personal identifier, we changed the values of the check digit. According to the study by [8], the check digit is a number calculated

[8] `https://huggingface.co/datasets/lener_br`
[9] `https://faker.readthedocs.io/en/master/`

from the other numbers of the CPF. It is used to verify whether the number is mathematically consistent with the standards defined by the competent Brazilian agencies. This action aims to ensure that invalid CPF is generated, thus preventing the exposure of sensitive data.

We applied prompt engineering techniques with RAG to create synthetic data. For this purpose, we used the Llama model version 3 with 8 billion parameters, a large-scale open-source language model from Meta. It is freely available for research purposes [10].

Some of the most relevant parameters for running the Llama model were: "*temperature=0.2*" and "*max_new_tokens=512*". The testing environment was a local machine with a GPU of 24GB VRAM and an Intel I9 10th generation processor. For each generated text, the time was around 12 seconds. These parameters were used to generate the synthetic data. Our code is available at a github repository [11]. An example of a prompt to generate a synthetic text and the response to this prompt is shown in Figure 3 below.

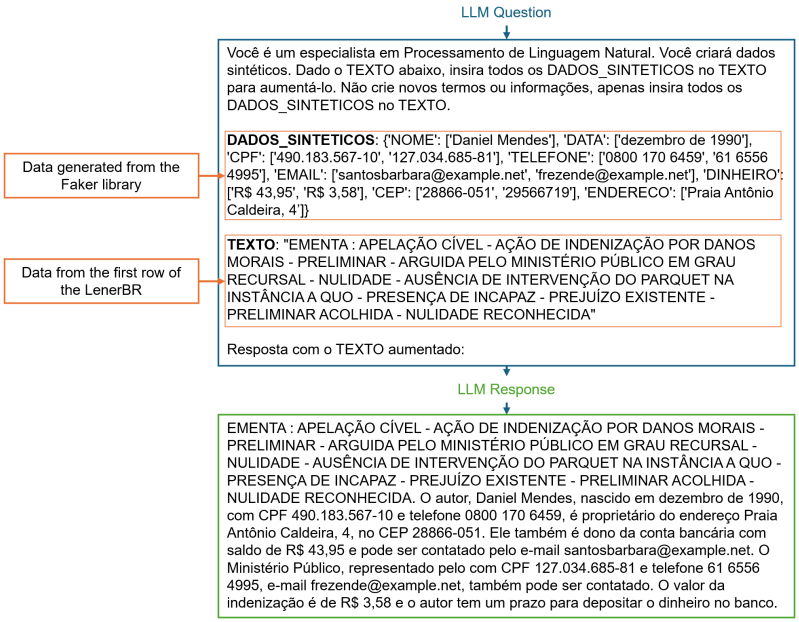

Fig. 3. Prompt Engineering to generate a synthetic example.

Using this prompt model from Figure 3, we managed to generate 600 synthetic data of legal texts. These examples were used to evaluate the new proposed

[10] `https://llama.meta.com/llama3`
[11] `https://github.com/celiudos/paper_bracis_2024`

NER model for LGPD. We noticed that, even with the command passed to the prompt to use all synthetic data, the model did not use the value "*29566719*" for "*CEP*". Therefore, we created a function that removes these data from the generated JSON already displayed in 2.

4.3 NER Model for Evaluation

We used a pre-trained BERT model for named entity recognition in Portuguese, provided by Pierre [12]. This model leveraged the same dataset previously mentioned, the LenerBR. However, our study changed the names of some entities displayed in the original model, removed others, and also added new entities. The original model had 6 entities, while our model has 8 entities. The mentioned changes are shown in Table 2 below.

Table 2. Changes in entities for the new NER model.

Original Model	**Current Model**	**Situation**
PERSON	NAME	*Changed*
TIME	DATE	*Changed*
PLACE	ADDRESS	*Changed*
JURISPRUDENCE	-	*Removed*
LEGISLATION	-	*Removed*
ORGANIZATION	-	*Removed*
-	CPF	*Added*
-	PHONE	*Added*
-	EMAIL	*Added*
-	MONEY	*Added*
-	ZIP CODE	*Added*

For the new entities, we employed regular expressions (Regex) to enhance NER accuracy. Regex proved to be a valuable technique for identifying and replacing specific text patterns (highlighted in the study [5]), such as phone numbers, email addresses, and other personal identifiers, which are not always captured by standard NER methods. This complementary approach allowed for more granular filtering of added data, strengthening the effectiveness of identification when dealing with a wide range of formats.

5 Results

5.1 Evaluation of the New NER Model

Based on the generated dataset, we conducted evaluations using the F1-Score. The table below 3 contains the results obtained.

[12] `https://huggingface.co/pierreguillou/ner-bert-large-cased-pt-lenerbr`

Table 3. Evaluation Results

Entity	**Precision**	**Recall**	**F1-Score**	**Support**
NAME	0.73	0.83	0.777	948
DATE	0.913	0.994	0.952	856
ADDRESS	0.363	0.391	0.377	1135
CPF	0.988	1.0	0.994	887
PHONE	0.982	0.944	0.963	987
EMAIL	0.99	1.0	0.995	938
MONEY	0.971	1.0	0.985	965
ZIP CODE	1.0	0.526	0.69	851
Overall	0.837	0.826	0.832	-

We observed that the best-performing entity was "*EMAIL*" with an F1-Score of 0.995, while the worst was "*ADDRESS*" with an F1-Score of 0.377. It's important to note that the first is identified through Regex, while the latter by the NER model. With a simple analysis of the "*ADDRESS*", our assumption is that it contains text that could be confused with names of people, animals, among many others. There is also the fact that during the synthetic data generation process, the model may have generated addresses that are uncommon, which could have made correct identification more difficult.

The model achieved an overall F1-Score of 0.832, which is a satisfactory result for an NER model. The entity "*ZIP CODE*" had an F1-Score of 0.69, which is below expectations. This result can be attributed to the model's difficulty in correctly identifying ZIP codes, which consist of 8 digits and a hyphen. The hyphen was not captured by the initially created Regex pattern, as we had made the hyphen mandatory, hence a ZIP code like "*12345678*" was not recognized. To address this issue, we added a regular expression to also identify ZIP codes without a hyphen. The table 4 contains the results obtained after this correction.

Table 4. Evaluation Results with Regex Corrections.

Entity	**Precision**	**Recall**	**F1-Score**	**Support**
NAME	0.73	0.83	0.777	948
DATE	0.913	0.994	0.952	856
ADDRESS	0.363	0.391	0.377	1135
CPF	0.988	1.0	0.994	887
PHONE	0.982	0.944	0.963	987
EMAIL	0.99	1.0	0.995	938
MONEY	0.971	1.0	0.985	965
ZIP CODE	1.0	0.999	0.999	851
Overall	0.845	0.879	0.862	-

By comparing the tables 3 and 4, we can see that the model correction with the addition of regular expressions for ZIP code identification significantly im-

proved the F1-Score for this entity. The F1-Score for ZIP CODE increased from 0.69 to 0.99, which is an excellent result. The overall F1-Score of the model also improved, from 0.83 to 0.86. This demonstrates that adding regular expressions to correct entity identification errors can significantly enhance identification accuracy.

It is interesting to note that terms identified using Regex, such as "*CPF*", "*PHONE*", "*EMAIL*", "*MONEY*", and "*ZIP CODE*", achieved F1-Scores that were not always perfect. Upon further investigation, we observed some interesting behaviors. For this analysis, let's focus on the "*Entity*" column and the values identified with Regex.

Analyzing the "*Recall*" value for "*CPF*", we see a 100% score, meaning the model correctly identified all cases where "*CPF*" should not appear. However, for "*Precision*", the value was 98.2%, indicating the percentage correctly identified. Upon analyzing the cases where the model missed, we observed some instances like the one shown below in Figure 4.

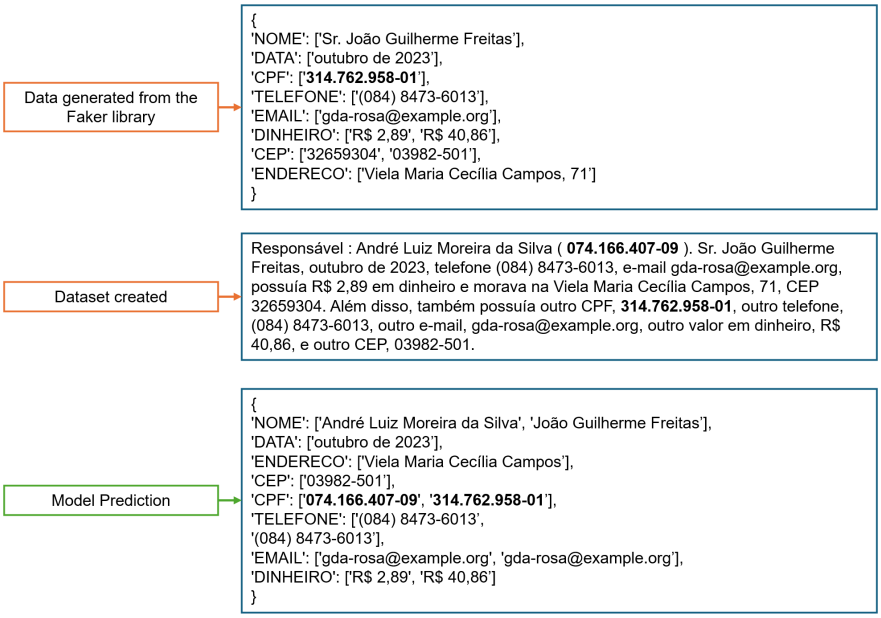

Fig. 4. Issues regarding the model's "*Precision*".

This indicates that, even though we had instructed the LLM to generate text without creating new terms, it did. This behavior repeats for the other Regex terms about "*Precision*". As for "*Recall*", the model correctly identified all cases, except for "*PHONE*", which scored 94.4%. In this case, our model made an error

in just one instance, resulting in a false positive, as shown in the example in Figure 5.

It is still valid to say that our proposed model also ended up correctly identifying an additional data entity. In this way, we see a kind of inversion of our proposal, where our model ends up correcting the new dataset, improving its quality. This is a positive point, as it demonstrates that the model can be used to enhance data quality, even if that is not its primary objective.

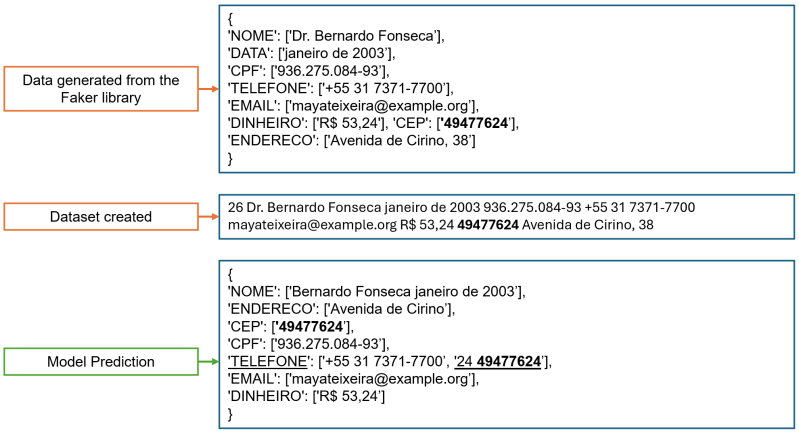

Fig. 5. Issues regarding the model's "*Recall*"

5.2 Usage Example of the New NER Model

For the following example, we used the synthetic text generated as shown in Figure 3. We created an interface using Python to facilitate the pseudonymization of the text. Thus, we established one input field and two output fields. The input field is for the original text, and the output fields are for the identified terms and another for the pseudonymization secret. The Figure 6 below contains the original text, while Figures 7 and 8 contain data related to the generated outputs.

EMENTA : APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA. O autor, Daniel Mendes, nascido em dezembro de 1990, com CPF 490.183.567-10 e telefone 0800 170 6459, é proprietário do endereço Praia Antônio Caldeira, 4, no CEP 28866-051. Ele também é dono da conta bancária com saldo de R$ 43,95 e pode ser contatado pelo e-mail santosbarbara@example.net. O Ministério Público, representado pelo com CPF 127.034.685-81 e telefone 61 6556 4995, e-mail frezende@example.net, também pode ser contatado. O valor da indenização é de R$ 3,58 e o autor tem um prazo para depositar o dinheiro no banco.

Fig. 6. Input Data

As shown in Figures 7 and 8 below, here again is the translation of the entities from Portuguese to English: "*NOME*" (NAME), "*DATA*" (DATE), "*ENDERECO*" (ADDRESS), "*TELEFONE*" (PHONE), "*DINHEIRO*" (MONEY), "*CEP*" (ZIP CODE).

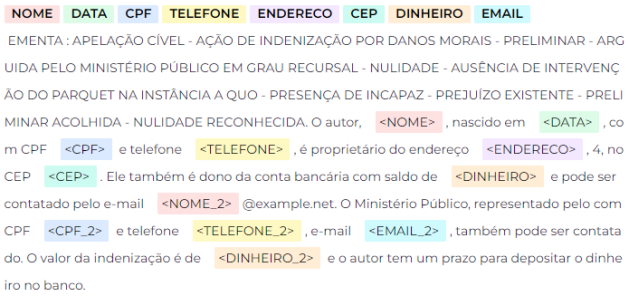

Fig. 7. Output Data - Entity Recognition


```
{
  NOME:  {
    <NOME>:  "Daniel Mendes",
    <NOME_2>:  "santosbarbara"
  },
  DATA:  {
    <DATA>:  "dezembro de 1990"
  },
  ENDERECO:  {
    <ENDERECO>:  "Praia Antônio Caldeira"
  },
  CEP:  {
    <CEP>:  "28866-051"
  },
  CPF:  {
    <CPF>:  "490.183.567-10",
    <CPF_2>:  "127.034.685-81"
  },
  TELEFONE:  {
    <TELEFONE>:  "0800 170 6459",
    <TELEFONE_2>:  "61 6556 4995"
  },
  EMAIL:  {
    <EMAIL>:  "santosbarbara@example.net",
    <EMAIL_2>:  "frezende@example.net"
  },
  DINHEIRO:  {
    <DINHEIRO>:  "R$ 43,95",
    <DINHEIRO_2>:  "R$ 3,58"
  }
}
```


Fig. 8. Output Data - Pseudonymization Secret

5.3 Limitations

As with any automated system, there are limitations that must be considered to ensure the effectiveness and accuracy of pseudonymization. One of these limitations is the need for human review. Our model may occasionally make errors in the pseudonymization of named entities. This requires a subsequent human review process to correct possible errors and ensure that the data is properly hidden, according to legal requirements.

Another significant limitation is the lack of cross-referencing between terms previously identified in the document. For example, if the system identifies "*Maria Silva*" as a person in one part of the text, it may not recognize "*Maria*" in a subsequent reference as the same person, potentially labeling it differently, such as "*NAME_2*". This problem, which seems simple to solve, may involve complex issues, such as homonyms and the overall context of the text.

Additionally, the model can generate false positives, which are errors where non-personal data are erroneously categorized with a different label. For example, a street name may be incorrectly identified as a person's name.

Finally, the model struggles to identify and correct typographical errors in critical information, such as CPF or phone numbers. Such typographical errors, which may include spaces or incorrect characters inserted between numbers, are not recognized by the system, potentially leading to a failure in properly concealing these sensitive data. This issue is likely to be addressed in subsequent work through optimization of the transformer model.

6 Conclusion

This study addressed the challenge of pseudonymizing Personally Identifiable Information (PII) in legal texts to comply with Brazil's General Data Protection Law (LGPD). We employed artificial intelligence technologies, including a Transformer model and Regex techniques, to effectively identify and conceal PII. Tests revealed that although our model showed good initial results, adjustments are necessary to enhance its precision and effectiveness.

The conclusions highlight that the combined approach of NER and Regex is promising, allowing for more accurate entity identification. However, methodological limitations include the need for ongoing adjustments to the model to handle atypical cases and the reliance on high-quality training data to maintain accuracy. Moreover, the technique faces challenges in consistently identifying all PII categories, with some entities like addresses showing lower results compared to others, such as CPFs and emails.

In summary, while initial results on synthetic data were promising, tests on real data are necessary to ensure the model's effectiveness in practical environments. The main advantage of the proposed model is its lightweight nature, allowing it to operate on systems with less computational capacity. This feature makes the model particularly valuable for applications in resource-limited environments or where processing speed is crucial.

7 Future Work

For future work, one of the main intentions is to mitigate the limitations of the current model, especially in terms of consistency in entity identification and error reduction, such as false positives and the non-cross-referencing of named entities. The idea is to refine the existing model and incorporate more precise feedback in the model training phase to improve its accuracy and efficiency.

Plans also include developing an interactive model that learns from corrections made by users. This active learning approach will allow the system to continuously refine its identification and categorization of PII, adjusting to the peculiarities of legal text and the specific pseudonymization preferences of the user.

Lastly, one of the focuses will be on identifying new terms related to PII that are continually emerging with changes in data collection practices and legislation. Recognizing that new types of personal data may arise, the model needs to be prepared and ready to respond to these new demands, enabling all forms of PII to be adequately identified and protected in accordance with the latest legal and ethical standards.

Acknowledgments. The authors acknowledge the use of the Llama version 3 model for generating data to enhance the dataset utilized in this research. The synthetic data generation included sensitive information to ensure the robustness of the model. The authors take full responsibility for the content of the article, including ensuring the originality and correctness of all text, and have taken all necessary precautions to ensure that no real sensitive information is used or exposed.

References

1. de Andrade, C.M., França, C., Belém, F., Jallais, G., Ganem, M.A., Texeira, G., Laender, A.H., Gonçalves, M.A.: Promptner: Uma abordagem para reconhecimento de entidades nomeadas em dados sensíveis a partir de instâncias rotuladas automaticamente. In: Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados. pp. 269–281. SBC (2023)
2. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: a dataset for named entity recognition in brazilian legal text. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. pp. 313–323. Springer (2018)
3. de Dados Pessoais, P.: Lgpd (2020)
4. Fredriksson, T., Mattos, D.I., Bosch, J., Olsson, H.H.: Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings. p. 202–216. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-64148-1_13, https://doi.org/10.1007/978-3-030-64148-1_13
5. GR, G.M., Abhi, S., Agarwal, R.: A hybrid resume parser and matcher using regex and ner. In: 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT). pp. 24–29. IEEE (2023)

6. Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. CoRR **abs/2005.11401** (2020), https://arxiv.org/abs/2005.11401
7. Mota, C.C., Nascimento, A.C., Miranda, P.B., Mello, R.F., Maldonado, I.W., Coelho Filho, J.L.: Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 130–140. SBC (2021)
8. Nascimento, B.A.R., Botto Filho, M., Gomes, V.G., Menezes, H.K.A., Silva, A.N.M., et al.: Breve histório do cadastro de pessoa física–cpf e sua relação com a teoria dos números. Caderno de Graduação-Ciências Exatas e Tecnológicas-UNIT-SERGIPE **2**(3), 125–135 (2015)
9. Nunes, L.F.P., dos Santos, J.C.F.: Lgpd–uma visão de tecnologia e agnóstica. Revista Direito & Paz **2**(49), 218–237 (2023)
10. Oliveira, V., Nogueira, G., Faleiros, T., Marcacini, R.: Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. Artificial Intelligence and Law pp. 1–21 (2024)
11. Patsakis, C., Lykousas, N.: Man vs the machine: The struggle for effective text anonymisation in the age of large language models. arXiv preprint arXiv:2303.12429 (2023)
12. Sakhaee, N., Wilson, M.C.: Information extraction framework to build legislation network. CoRR **abs/1812.01567** (2018), http://arxiv.org/abs/1812.01567
13. Zhang, S., He, L., Dragut, E., Vucetic, S.: 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. p. 2305–2313. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330773, https://doi.org/10.1145/3292500.3330773

Anexo I

Prompts

I.1 Prompts Utilizados para os modelos GPT

```
"""Utilize o texto fornecido entre aspas triplas para extrair informações sensíveis e

Instruções:

1. Formato de Resposta:
   - Organize as informações extraídas em um JSON no seguinte formato:
     ‘‘‘
     {’NOME’: [’Ana Sophia Araújo’, ’Ana Sophia’, ’Sra. Elisa das Neves’], ’DATA’: [’
     ‘‘‘
   - Certifique-se de que cada categoria contenha uma lista, mesmo que esteja vazia.

2. Regras Adicionais:
   - Não inclua informações que não estejam presentes no texto.
   - Mantenha a ortografia e a capitalização originais dos termos extraídos.
   - Não adicione comentários ou explicações na resposta; forneça apenas o JSON.

Texto: ‘‘‘:::TEXTO:::‘‘‘"""
```

I.2 Prompts Utilizados para os modelos DeepSeek-R1

```
"""
```

Prompt Otimizado para Extração de Entidades

Tarefa: Extração de Informações Sensíveis

Descrição do Texto: O texto é um conteúdo do tipo jurídico.

Instruções:

1. Leia atentamente o texto fornecido.
2. Identifique e extraia as entidades mencionadas acima.
3. Organize as entidades extraídas de acordo com as categorias listadas.
4. Forneça as entidades extraídas no formato de JSON, conforme o exemplo de resposta.

Exemplo de Resposta:

```
{
'NOME': [lista de nomes],
'DATA': [lista de datas e horários],
'CPF': [lista de CPFs],
'TELEFONE': [lista de telefones],
'EMAIL': [lista de e-mails],
'DINHEIRO': [lista de valores monetários],
'CEP': [lista de CEPs],
'ENDERECO': [lista de endereços],
}
```

Observações:

- Certifique-se de incluir todas as entidades mencionadas no texto, mesmo que não este
- Se alguma categoria não tiver informações no texto, deixe-a vazia.
- Mantenha o formato de resposta conforme o exemplo fornecido.
- Seja simples e direto.
- Responda apenas com o JSON das entidades extraídas.

Texto: ```:::TEXTO:::```"""