Anonimização de Textos Clínicos Utilizando LLM

Arthur M. Pereira[1], Leonardo F. Martins[2], Laisa M.A. Sartes[1], Larissa F. de Almeida[1], Heder S. Bernardino[1], Jairo F. de Souza[1]

[1]Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brasil

[2]Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rio de Janeiro – RJ – Brasil

arthurmpereira2010@gmail.com,leonardomartins@puc-rio.br,
laisa.sartes@ufjf.br, larissafernanda40.lf@gmail.com,
heder.bernardino@ufjf.br,jairo.souza@ufjf.br

Abstract. *Data-driven model training is essential for healthcare advancements, enabling more personalized medicine. Clinical text anonymization protects patient privacy amid increasing digitalization. Traditional methods, while effective, may reduce data utility and fail in contextual anonymization. This study proposes a method based on large language models (LLMs), combining named entity recognition (NER) and text rephrasing to ensure coherence and anonymization. Tested on therapeutic transcripts, the method achieved high accuracy in removing sensitive information while preserving textual integrity, making it applicable across contexts.*

Resumo. *O uso de dados no treinamento de modelos é essencial para avanços na saúde, viabilizando um tratamento mais personalizado. A anonimização de textos terapêuticos protege a privacidade dos pacientes diante da digitalização crescente. Métodos tradicionais, embora eficazes, podem reduzir a utilidade dos dados e falhar na anonimização contextual. Este estudo propõe um método baseado em modelos de linguagem de grande porte (LLMs), combinando reconhecimento de entidades nomeadas (NER) e reformulação textual para garantir coerência e anonimização contextual. Testado em transcrições terapêuticas, o método demonstrou alta precisão na remoção de informações sensíveis sem comprometer a integridade textual,se tornando aplicável a diferentes contextos.*

1. Introdução

A privacidade nas interações terapêuticas é um elemento essencial para a relação de confiança entre pacientes e profissionais de saúde mental [Salles and Castelo 2023]. No entanto, a crescente digitalização dos atendimentos e o uso de inteligência artificial (IA) na análise de registros clínicos têm levantado preocupações sobre a confidencialidade dos dados sensíveis contidos em sessões de terapia [Isa 2024]. Diferentemente de prontuários médicos estruturados, as transcrições de consultas psicológicas contêm narrativas espontâneas, repletas de informações pessoais, eventos da vida do paciente e dados contextuais que podem facilitar sua reidentificação, mesmo após processos de anonimização convencionais [Allen et al. 2015, Britton et al. 2022].

Os desafios na anonimização de textos clínicos tornam-se ainda mais complexos quando aplicados a sessões terapêuticas, pois esses registros não apenas incluem identificadores diretos, como nomes e locais, mas também elementos subjetivos que podem

revelar a identidade do paciente de forma indireta. Expressões idiomáticas, relatos de eventos únicos e até o estilo narrativo da fala representam riscos adicionais à privacidade.

Além disso, a simples remoção de identificadores pode comprometer o significado original da fala, dificultando a utilização desses dados para pesquisa e análise clínica. Dessa forma, métodos mais avançados são necessários para equilibrar a proteção da privacidade com a preservação da integridade do conteúdo.

A importância desse problema se reflete no contexto da segurança de dados na saúde. Relatórios recentes indicam um crescimento alarmante no número de vazamentos de informações médicas, com 2023 registrando um recorde de 168 milhões de registros comprometidos [HIPAA Journal 2025]. Embora grande parte dessas violações envolva bancos de dados estruturados, a exposição de transcrições de sessões terapêuticas representa um risco ainda maior, pois pode resultar em danos psicológicos e sociais irreparáveis para os pacientes. Legislações como o Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia [União Europeia 2016] e a Lei de Portabilidade e Responsabilidade de Seguros de Saúde (HIPAA) nos Estados Unidos [U.S. Department of Health and Human Services 2003] estabelecem exigências para a proteção desses dados, reforçando a necessidade de técnicas de anonimização.

Diante desse cenário, este trabalho apresenta um método para a anonimização de transcrições de sessões terapêuticas, utilizando Modelos de Linguagem de Grande Escala (LLMs). Em contraste com abordagens convencionais que removem apenas identificadores explícitos. A estratégia proposta combina duas etapas: uma anonimização baseada em regras para ocultar elementos sensíveis, seguida por um refinamento realizado por um LLM, que reestrutura o texto para preservar seu significado e contexto. Esse método busca melhorar a proteção da privacidade sem comprometer a utilidade das transcrições para pesquisa e análise clínica.

A principal contribuição deste estudo está no desenvolvimento de uma abordagem de anonimização que vai além da simples substituição de identificadores, abordando os desafios específicos das interações verbais registradas em sessões terapêuticas. Ao considerar o contexto e a fluidez da fala, o método proposto permite que os dados sejam utilizados de maneira ética e segura, promovendo um equilíbrio entre inovação tecnológica e privacidade na saúde mental.

2. Fundamentação Teórica

Os dados clínicos desempenham um papel crucial na pesquisa médica, no desenvolvimento de modelos preditivos e na melhoria da assistência à saúde [Gates et al. 2024]. Esses dados, que incluem informações como histórico de pacientes, exames e diagnósticos, são essenciais para análises avançadas e personalização de tratamentos. No entanto, devido à sua sensibilidade, a proteção dessas informações tornou-se uma preocupação central, especialmente com o aumento do uso de inteligência artificial para sua análise.

O aprendizado de máquina tem sido amplamente empregado no processamento de texto clínico [Supriya and Deepa 2020], permitindo a extração automatizada de informações relevantes. Entre essas técnicas, o reconhecimento de entidades nomeadas (NER) se destaca como uma abordagem essencial para identificar informações sensíveis, como nomes de pacientes, datas e locais, viabilizando a anonimização desses dados

[Pettersson et al. 2024]. Modelos de aprendizado profundo, como redes neurais convolucionais e recorrentes, eliminaram a necessidade de engenharia manual de características ao utilizar representações vetoriais complexas que capturam significados contextuais dos textos, aumentando significativamente a precisão na identificação de entidades, mesmo em contextos ambíguos [Yadav and Bethard 2019, Fabregat et al. 2019].

Mais recentemente, os LLMs, como o GPT-4o mini, revolucionaram a anonimização automatizada. Baseados na arquitetura Transformer, esses modelos processam sequências textuais inteiras em paralelo, utilizando mecanismos de autoatenção para capturar relações complexas entre palavras e frases [Amazon Web Services 2025, IBM 2025]. Treinados em vastos corpora, como Common Crawl e Wikipedia, os LLMs podem realizar tarefas variadas, incluindo reconhecimento e substituição de entidades sensíveis. O GPT-4o mini, especificamente, combina uma arquitetura otimizada com bilhões de parâmetros, proporcionando alta precisão no NER e flexibilidade para se adaptar a diferentes contextos clínicos. Sua aplicação na anonimização de dados clínicos permite não apenas identificar entidades com precisão, mas também gerar substituições contextuais que preservam a coerência dos textos anonimizados, garantindo sua utilidade para pesquisas [Amazon Web Services 2025].

As técnicas tradicionais de anonimização, como supressão, substituição de caracteres, embaralhamento, adição de ruído, generalização, k-anonimidade e l-diversidade, são amplamente reconhecidas por sua importância na proteção de dados sensíveis. No entanto, essas abordagens apresentam limitações significativas quando aplicadas a grandes volumes de dados textuais, pois frequentemente resultam em perda de informações essenciais para análises subsequentes [Marques and Bernardino 2020, Mogre et al. 2012, Shamsinejad et al. 2024]. Métodos como k-anonimidade e l-diversidade exigem configurações precisas para evitar ataques de reidentificação, enquanto a supressão e a substituição podem comprometer a integridade dos dados [El Emam and Arbuckle 2013]. Esses desafios motivaram a busca por soluções mais inteligentes e automatizadas, como o uso de aprendizado de máquina e, mais recentemente, de grandes modelos de linguagem, que conseguem equilibrar privacidade e utilidade de maneira mais eficiente.

2.1. Projeto MATCH e sua Relevância

O **Projeto MATCH** (*Matching Alcoholism Treatments to Client Heterogeneity*) foi um estudo multicêntrico conduzido pelo *National Institute on Alcohol Abuse and Alcoholism* (NIAAA), com o objetivo de investigar a eficácia de diferentes abordagens terapêuticas no tratamento do alcoolismo [Kadden 1995]. Este ensaio clínico foi pioneiro na personalização de tratamentos para dependência de álcool, analisando como diferentes perfis de pacientes respondiam a distintas estratégias terapêuticas.

O projeto avaliou três modalidades de tratamento: (i) **Terapia Cognitivo-Comportamental (TCC)** – Enfatizando estratégias para modificar padrões de pensamento disfuncionais e desenvolver habilidades para lidar com situações de risco de recaída. (ii) **Entrevista Motivacional (MET - *Motivational Enhancement Therapy*)** – Baseada em entrevistas motivacionais para aumentar o comprometimento do paciente com a mudança comportamental; (iii) **Terapia de Doze Passos (TSF - *Twelve-Step Facilitation Therapy*)** – Inspirada nos princípios dos Alcoólicos Anônimos, incentivando a participação em grupos de apoio.

Os resultados do **Projeto MATCH** influenciaram significativamente as abordagens modernas para o tratamento da dependência de álcool, demonstrando que, embora as três abordagens tenham mostrado eficácia, diferentes perfis de pacientes se beneficiavam de maneira distinta de cada intervenção [Kadden 1995]. Isso reforçou a importância da personalização terapêutica, um conceito fundamental para os avanços recentes na saúde digital e no uso de inteligência artificial para análise de dados clínicos.

No contexto do presente estudo, os princípios da **Terapia Cognitivo-Comportamental (TCC)** descritos no **Projeto MATCH** e adaptados para o contexto brasileiro [Gumier 2019] são adotados como base para a análise de sessões psicoterapêuticas online, permitindo a extração e anonimização automatizada de dados sensíveis presentes nos diálogos.

3. Trabalhos Relacionados

A anonimização de textos clínicos tem sido amplamente estudada nos últimos anos [Ribeiro 2023, Pissarra et al. 2024], abrangendo desde abordagens tradicionais de Reconhecimento de Entidades Nomeadas (NER) até soluções mais avançadas baseadas em LLMs e técnicas de incorporação de palavras. Os trabalhos analisados exploram diferentes métodos para garantir a preservação da informação clínica, ao mesmo tempo em que reduzem o esforço manual necessário para anonimização. No entanto, a maioria das abordagens foca na substituição direta de identificadores, sem considerar a reescrita contextual do texto, o que pode comprometer a usabilidade dos dados anonimizados.

Ribeiro et al. [Ribeiro 2023] compararam métodos tradicionais de NER com técnicas baseadas em incorporações de palavras, como Word2Vec e GloVe. Embora essas técnicas apresentem potencial, os autores observaram que resultam em maior perda de informações quando comparadas a abordagens como Conditional Random Fields (CRF) e Microsoft Presidio. Para mitigar essas limitações, a plataforma INCOGNITUS [Ribeiro et al. 2023] integrou múltiplas técnicas de anonimização e alcançou um escore F1 de aproximadamente 95%, garantindo 100% de recall por meio da substituição de palavras por alternativas semanticamente equivalentes. No entanto, a generalização do modelo apresentou desafios ao ser testada em novos conjuntos de dados.

Técnicas baseadas em incorporações de palavras também foram exploradas por Hassan et al. [Hassan et al. 2019], que utilizaram essa abordagem para avaliar relações semânticas entre termos, melhorando a detecção de informações identificadoras de maneira independente do idioma. Apesar dos avanços, os autores destacaram dificuldades na captura de nuances linguísticas e na desambiguação de termos. Já Pissarra et al. [Pissarra et al. 2024] investigaram o uso de modelos generativos na anonimização de textos clínicos, especificamente o ClinicAlbert, demonstrando que esses modelos superam as técnicas tradicionais em termos de sensibilidade. No entanto, nenhum método analisado atingiu 100% de recall, e os autores mencionam desafios em relação à geração de textos que preservem integralmente a semântica original.

No contexto da língua portuguesa, Gonçalves et al. [Gonçalves 2023] analisaram cerca de 2.000 notas clínicas, identificando aproximadamente 4.000 blocos de texto contendo informações sensíveis. Os pesquisadores criaram um dicionário contendo 967 abreviações médicas e seus respectivos tipos semânticos, facilitando o processamento automatizado de textos clínicos em português.

Embora esses estudos tenham avançado na identificação e anonimização de dados clínicos, eles ainda enfrentam limitações no que diz respeito à reestruturação do texto para preservar o significado original sem comprometer a privacidade. Métodos tradicionais frequentemente focam na remoção ou substituição direta de palavras sensíveis, o que pode alterar o contexto e dificultar o uso posterior dos dados anonimizados.

Diante dessas limitações, este trabalho propõe uma abordagem baseada em LLMs para a anonimização contextual de textos clínicos com foco especial em transcrições de sessões terapêuticas. Diferentemente das abordagens convencionais, que apenas removem ou mascaram informações sensíveis, o método aqui apresentado gera uma nova versão do texto, preservando seu significado original por meio da reformulação semântica. Esse processo visa maximizar a anonimização sem comprometer a utilidade dos dados, permitindo que os registros anonimizados ainda sejam usados para pesquisa e análise clínica sem risco de reidentificação.

A contribuição deste trabalho está na introdução de uma estratégia que não apenas anonimiza os dados, mas também visa a coerência e a usabilidade dos textos reformulados, respondendo a uma das principais lacunas identificadas nas abordagens atuais.

4. Materiais e Métodos

Nesta seção, descrevemos os dados utilizados no estudo e os métodos empregados para a anonimização das transcrições terapêuticas. Primeiramente, apresentamos a origem e a estrutura das transcrições analisadas, destacando sua relevância para a pesquisa. Em seguida, detalhamos a metodologia adotada para anonimização dos textos, utilizando um pipeline baseado em técnicas de *Reconhecimento de Entidades Nomeadas* (NER) e reformulação textual assistida por modelos de linguagem de grande porte (*Large Language Models* – LLMs). O objetivo dessa abordagem é garantir a proteção das informações sensíveis sem comprometer a integridade e a utilidade dos dados para análises subsequentes.

4.1. Dados Utilizados

Para a aplicação do método, foram analisadas transcrições de sessões terapêuticas de quatro pacientes, seguindo a estrutura da Terapia Cognitivo-Comportamental (TCC) proposta no Projeto MATCH. Especificamente, foram utilizadas as transcrições das sessões 1 a 7 de cada paciente, totalizando 28 sessões analisadas.

No contexto deste estudo, seguimos a estrutura baseada na Terapia Cognitivo-Comportamental (TCC), conforme descrita por Kadden et al. [Kadden 1995], complementada por estratégias motivacionais e princípios do Modelo de Prevenção de Recaída.

As transcrições das sessões terapêuticas foram utilizadas para avaliar o desempenho de técnicas de anonimização baseadas em aprendizado de máquina e LLMs, garantindo a remoção segura de informações sensíveis sem comprometer a integridade do conteúdo para futuras análises qualitativas e quantitativas. As sessões analisadas cobrem aspectos essenciais da terapia, incluindo psicoeducação sobre alcoolismo, identificação de situações de risco, manejo da fissura, estratégias de prevenção de recaída e desenvolvimento de habilidades de enfrentamento. Essas transcrições foram utilizadas como base para o desenvolvimento e avaliação da metodologia de anonimização proposta.

Os dados utilizados nesta pesquisa foram cedidos por um projeto da Universidade Federal de Juiz de Fora (UFJF), que visa explorar a eficácia de diferentes abordagens

terapêuticas no tratamento de dependências, com foco na personalização de tratamentos. Este projeto é parte de um esforço maior para integrar inteligência artificial na análise de dados clínicos, garantindo a proteção de informações sensíveis.

4.2. Métodos Utilizados

A anonimização de textos clínicos tem sido testada com diversas estratégias, que vão desde técnicas baseadas em *Reconhecimento de Entidades Nomeadas* (*NER*, *Named Entity Recognition*) até abordagens mais avançadas utilizando Modelos de Linguagem de Grande Escala (*LLMs*, *Large Language Models*). Essas técnicas buscam equilibrar a remoção de informações sensíveis com a preservação do significado do texto, garantindo que os dados anonimizados possam ser utilizados para pesquisa e análise sem comprometer a privacidade dos indivíduos.

Tradicionalmente, abordagens baseadas em NER são amplamente utilizadas para identificar e classificar informações sensíveis, como nomes, endereços e identificadores numéricos [Shamsinejad et al. 2024]. No entanto, essas técnicas frequentemente exigem grandes quantidades de dados anotados manualmente, além de dependerem da capacidade do modelo de reconhecer variações linguísticas e contextuais.

Com o avanço dos modelos de aprendizado profundo, abordagens baseadas em redes neurais recorrentes (RNNs) e transformadores, como BERT e GPT, passaram a ser exploradas na anonimização de textos médicos [Liu et al. 2023, Vakili et al. 2024]. Essas abordagens apresentam vantagens na identificação de padrões linguísticos complexos e na generalização para diferentes domínios clínicos. Além disso, estudos recentes têm demonstrado a eficácia da reformulação textual como estratégia complementar à anonimização tradicional, utilizando modelos generativos para reescrever trechos sensíveis sem comprometer o conteúdo informativo [Pissarra et al. 2024].

No contexto específico de sessões terapêuticas, ao contrário de prontuários médicos, a forma livre e narrativa dos textos apresenta desafios adicionais para anonimização. Métodos que utilizam reformulação textual têm sido propostos para lidar com esse tipo de dado, garantindo que informações confidenciais sejam protegidas sem perda significativa do contexto clínico [Larbi et al. 2023].

A seguir, é apresentada a proposta deste trabalho, que utiliza um pipeline de anonimização baseado na combinação de NER e reformulação contextual para garantir a máxima proteção dos dados sensíveis em transcrições de sessões terapêuticas. O modelo GPT foi utilizado tanto na fase de NER quanto na reformulação textual, aproveitando sua capacidade de identificar e substituir entidades sensíveis de forma contextualizada.

5. Proposta

A abordagem proposta para a anonimização de transcrições de sessões terapêuticas foi estruturada em um pipeline composto por múltiplas fases, conforme ilustrado na Figura 1. O objetivo é garantir a proteção de informações sensíveis, preservando a coerência e o significado original dos textos, permitindo seu uso seguro para análise e pesquisa.

5.1. Pipeline de Anonimização

O processo de anonimização proposto é composto pelas seguintes etapas:

1. **Identificação de Entidades Sensíveis**: O primeiro passo consiste na extração e categorização das entidades presentes nos textos. Essas entidades são organizadas em dois grupos:
    - **Grupo de Entidades de Segurança**: inclui informações que devem ser anonimizadas, como nomes, endereços e identificadores numéricos.
    - **Grupo de Entidades de Informação**: inclui informações que devem ser preservadas para manter o contexto e a utilidade dos textos.

   A extração das entidades é realizada por meio de um modelo de *Named Entity Recognition* (NER), treinado para reconhecer padrões específicos em diálogos terapêuticos.
2. **Anonimização Inicial**: As entidades classificadas no grupo de segurança são substituídas por etiquetas genéricas representativas de seu tipo (por exemplo, nomes próprios são substituídos por "[NOME]"). As entidades do grupo de informação são preservadas e destacadas no texto.
3. **Reformulação Textual**: Um modelo de linguagem de grande escala (*LLM*) é utilizado para reescrever os trechos anonimizados de forma contextualizada, substituindo elementos específicos por sinônimos e generalizações, garantindo que o significado original seja mantido.

A Figura 1 ilustra o fluxo completo do processo, desde a identificação inicial das entidades até a reformulação do texto anonimizado.

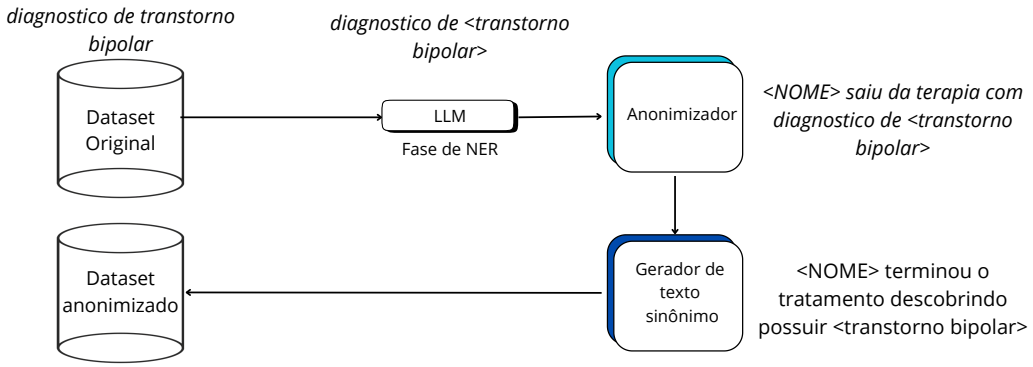

Figura 1. Pipeline do processo de anonimização.

5.2. Métricas de Avaliação

A avaliação dos resultados foi conduzida a partir de duas perspectivas principais: a eficácia do processo de anonimização e a preservação das informações essenciais do texto.

Para medir a eficácia da anonimização, utilizamos uma métrica baseada na contagem de ocorrências das entidades do grupo de segurança nas versões anonimizadas do texto (*Recall*) [Hassan et al. 2018]. Essa análise foi realizada tanto após a primeira fase de anonimização tradicional quanto após a segunda fase de reformulação textual. A comparação entre as ocorrências no texto original e nas versões anonimizadas permitiu avaliar o grau de remoção das informações sensíveis. Como essa verificação exige identificar manualmente as entidades anonimizadas, foi necessário um processo de anotação manual dos textos, o que pode introduzir margem de erro devido à subjetividade humana.

Para avaliar a preservação das informações essenciais durante o processo de anonimização, propomos uma abordagem estruturada baseada no protocolo do projeto

MATCH. O objetivo é garantir que os principais conteúdos abordados nas sessões 1 a 7 da terapia sejam mantidos na versão anonimizada do texto. Para isso, utilizamos um modelo de linguagem (*GPT-4o Mini*) para comparar a retenção dessas informações entre o texto original e sua versão anonimizada, verificando a equivalência dos conteúdos.

5.2.1. Verificação Estruturada de Retenção de Informações

A ficha de verificação estruturada foi desenvolvida para avaliar a integridade das informações terapêuticas na versão anonimizada do texto. Esta ficha segue um modelo predefinido contendo dois componentes principais: a verificação de vazamento de dados críticos e a verificação de aderência ao conteúdo esperado.

Na verificação de vazamento de dados críticos, garantimos que informações sensíveis, como nome do paciente, data de nascimento, sexo, idade, telefone, e-mail, endereço, cidade, estado, país, CEP, CPF, RG, estado civil, profissão e nome do terapeuta, estejam completamente removidas da versão anonimizada final do texto.

Para a verificação de aderência ao conteúdo esperado, cada sessão analisada é verificada quanto à presença dos principais tópicos terapêuticos conforme o protocolo do projeto MATCH. Os tópicos avaliados para cada sessão são: (i) **Sessão 1 - Introdução e Acolhimento**: Rapport estabelecido, compreensão dos objetivos da terapia, psicoeducação realizada, metas terapêuticas definidas e automonitoramento solicitado; (ii) **Sessão 2 - Identificação de Riscos**: Identificação de situações de risco, discussão de estratégias preventivas e apresentação de técnicas de manejo de fissura; (iii) **Sessão 3 - Pensamentos Perigosos**: Compreensão de estados de risco, reafirmação do compromisso com a recuperação, discussão de pensamentos justificadores e conclusão de uma lista de prós e contras; (iv) **Sessão 4 - Resolução de Problemas**: Identificação de problemas pós-abstinência, aplicação de estratégias de resolução e prática com role-play; (v) **Sessão 5 - Recusa de Bebidas**: Discussão de estratégias de recusa, identificação de gatilhos e compreensão da pressão social e emocional; (vi) **Sessão 6 - Prevenção de Recaídas**: Desenvolvimento de um plano de prevenção e discussão de estratégias para reagir a recaídas; e (vii) **Sessão 7 - Decisões e Racionalizações**: Discussão de riscos de decisões aparentemente irrelevantes, definição de estratégias para comportamentos de risco e identificação de racionalizações.

Cada item é avaliado através de respostas binárias (sim/não), permitindo a verificação de vazamento de dados sensíveis e a confirmação da manutenção do conteúdo terapêutico essencial. A comparação entre as respostas extraídas dos textos anonimizado e original permite verificar quantas respostas de "sim"ou "não"são equivalentes, garantindo que o conteúdo essencial das sessões seja mantido, mesmo após as modificações introduzidas pelo processo de anonimização. Essa abordagem objetiva avalia a eficácia da metodologia proposta, garantindo que a anonimização dos dados seja realizada sem comprometer a integridade das informações relevantes para a análise terapêutica.

6. Resultados

Nesta seção, são apresentados os resultados obtidos após a aplicação das técnicas de anonimização desenvolvidas. Os dados analisados contemplam o desempenho nas duas fases do processo de anonimização, assim como a avaliação baseada no uso de LLMs para a geração das fichas de verificação.

A implementação do pipeline de anonimização foi realizada em Python, utilizando a API da OpenAI para acesso ao modelo *GPT-4o Mini*. Os dados utilizados nesta pesquisa foram cedidos por um projeto da Universidade Federal de Juiz de Fora (UFJF) e não são públicos, sendo tratados de acordo com protocolos éticos para garantir a proteção das informações sensíveis.Além disso o conjunto foi limitado a quatro pacientes pela disponibilidade e integridade de todos os dados dos mesmos, mantendo um padrão mesmo que com uma base reduzida.

6.1. Desempenho nas Fases de Anonimização

A Tabela 1 apresenta uma comparação entre a primeira e a segunda fase do processo de anonimização (anonimização tradicional e geração de texto sinônimo, respectivamente), destacando o número de ocorrências identificadas e anonimizadas com sucesso, bem como a taxa de acerto percentual para cada tipo de entidade do grupo de segurança.

Tabela 1. Comparação entre a Primeira Fase e Segunda Fase para anonimização. A linha "Total" representa a soma das ocorrências e a taxa global de sucesso.

Tipo	Ocorrências	Únicas	Primeira Fase		Segunda Fase	
			Sucesso	% Sucesso	Sucesso	% Sucesso
Pessoa	790	241	783	99,11%	785	99,37%
Endereço	20	17	18	90,00%	19	95,00%
Data	78	62	61	78,21%	65	83,33%
E-mail	2	1	2	100,00%	2	100,00%
Total	**890**	**321**	**864**	**97,19%**	**871**	**97,86%**

Os resultados evidenciam um bom desempenho do sistema em ambas as fases, com taxas de sucesso na casa de 97%. Observa-se uma melhoria na segunda fase subindo de 97,19% para 97,86%, indicando que o refinamento do processo contribuiu para a obtenção de resultados ainda mais precisos, especialmente na anonimização de endereços e datas, que apresentaram um aumento nos índices de acerto.

6.2. Avaliação da anonimização por meio de LLM e ficha de informação

A Tabela 2 apresenta os resultados obtidos a partir da abordagem baseada em LLMs, utilizada para a anonimização de fichas de pacientes. Os dados analisados incluem a quantidade de entidades anonimizadas corretamente e o volume de informações preservadas após o processo de anonimização.

Os resultados obtidos apresentam boas taxas de acerto na preservação de informação, com resultados em uma media de 98,86%,e apesar de poucas ocorrencias, uma taxa de acerto de 100% na retirada de informações sensiveis. Além disso, os dados apresentados reforçam a eficácia dos LLMs, como o GPT-4o mini, na anonimização automatizada, destacando seu potencial para lidar com volumes significativos de dados e contextos complexos. Apesar disso temos ressalvas a fazer na conclusão.

A anotação manual foi realizada para gerar as métricas de sucesso, onde avaliadores identificaram manualmente as entidades anonimizadas, o que pode introduzir margem de erro devido à subjetividade humana.

Tabela 2. Análise de anonimização e informações por ficha de paciente. A linha "Total" apresenta a soma dos dados encontrados e a porcentagem global de acerto.

Ficha	Anonimização			Informação		
	Original	**Anon.**	**% Acerto**	**Original**	**Anon.**	**% Acerto**
Paciente 1	3	3	100%	22	22	100%
Paciente 2	5	5	100%	22	22	100%
Paciente 3	3	3	100%	22	22	100%
Paciente 4	4	4	100%	22	21	95,45%
Total	15	15	100%	88	87	98,86%

7. Conclusões e Trabalhos Futuros

A anonimização de dados em textos de linguagem natural é um desafio crucial, especialmente em contextos que lidam com informações sensíveis, como na área clínica. A preservação da privacidade sem comprometer a utilidade dos dados é uma demanda crescente, tornando essencial o desenvolvimento de abordagens eficazes. Neste estudo, foi proposto um processo de anonimização baseado em LLMs, demonstrando sua viabilidade e adaptabilidade para diferentes cenários.

Os resultados obtidos indicam que o processo proposto possui potencial de aplicação, podendo ser explorado em diversos tipos de dados para avaliar sua capacidade de generalização. No entanto, desafios foram identificados ao longo do estudo, principalmente no que se refere à avaliação dos resultados. A inexistência de métricas padronizadas para esse tipo de abordagem ressalta a necessidade de pesquisas futuras voltadas ao desenvolvimento de indicadores mais precisos e consistentes para a aferição da qualidade da anonimização realizada por LLMs.

Outra limitação é a subjetividade envolvida na avaliação da eficácia da anonimização. O processo de anotação manual das entidades pode introduzir margem de erro, uma vez que depende da interpretação individual do avaliador. Para tornar essa avaliação mais precisa, estudos futuros podem incorporar múltiplos avaliadores independentes.

Além disso, a execução do estudo foi impactada por limitações financeiras, principalmente devido ao custo elevado de acesso e uso de LLMs de última geração. Esse fator restringiu a possibilidade de testar múltiplos modelos e explorar diferentes variações de *prompts*, limitando a análise comparativa que poderia oferecer uma visão mais abrangente sobre o desempenho da proposta. Assim, a continuidade dessa linha de pesquisa dependerá de investimentos que viabilizem experimentações mais amplas, bem como do desenvolvimento de modelos mais acessíveis e eficientes e de preferencia que possam ser usados de maneira local aumentando o controle e privacidade dos dados usados.

Vale destacar que o procedimento proposto demonstrou potencial para aprimorar a anonimização de textos sensíveis. Com investimentos adequados em pesquisa, desenvolvimento de métricas de avaliação e exploração de bases de dados diversificadas, espera-se que essa abordagem contribua não apenas para a área clínica, mas também para outros domínios que demandam a proteção de informações sensíveis sem comprometer a qualidade dos dados anonimizados.

Agradecimentos

Os autores agradecem o apoio financeiro da CAPES, CNPq, FAPEMIG e UFJF. Também, agradecemos pela Bolsa Jovem Cientista do Nosso Estado da FAPERJ e Bolsa de Incentivo à Produtividade da PUC-Rio.

Referências

Allen, C. O., Carrier, S. R., Harold Moss, I., and Woods, E. (2015). Anonymizing sensitive identifying information based on relational context across a group. US Patent 9,047,488.

Amazon Web Services (2025). What is a large language model? Acesso em: 15 fev. 2025.

Britton, F. C., Dowling, S., and Frain, M. (2022). A contribution towards the regulation of anonymised datasets within the framework of gdpr. In *2022 Cyber Research Conference-Ireland (Cyber-RCI)*, pages 1–6. IEEE.

El Emam, K. and Arbuckle, L. (2013). *Anonymizing health data: case studies and methods to get you started*. "O'Reilly Media, Inc.".

Fabregat, H., Duque, A., Martinez-Romo, J., and Araujo, L. (2019). De-identification through named entity recognition for medical document anonymization. In *IberLEF@ SEPLN*, pages 663–670.

Gates, J. D., Yulianti, Y., and Pangilinan, G. A. (2024). Big data analytics for predictive insights in healthcare. *Intl. Transactions on Artificial Intelligence*, 3(1):54–63.

Gonçalves, A. C. M. (2023). Text mining de relatórios clínicos. Master's thesis, ISCTE Lisboa.

Gumier, A. B. (2019). *Terapia cognitivo-comportamental por internet para dependentes de álcool: viabilidade e estudo piloto de um ensaio clínico randomizado*. PhD thesis, Universidade Federal de Juiz de Fora.

Hassan, F., Domingo-Ferrer, J., and Soria-Comas, J. (2018). Anonymization of unstructured data via named-entity recognition. In *Proc. of the Intl. Conf. on Modeling Decisions for Artificial Intelligence (MDAI)*, pages 296–305. Springer.

Hassan, F., Sánchez, D., Soria-Comas, J., and Domingo-Ferrer, J. (2019). Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In *Proc. of the IEEE Intl. Conf. On Trust, Security And Privacy In Computing And Communications / IEEE Intl. Conf. On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365. IEEE.

HIPAA Journal (2025). Healthcare data breach statistics. Acesso em: 21 fev. 2025.

IBM (2025). What are large language models (llms)? Acesso em: 15 fev. 2025.

Isa, A. K. (2024). Exploring digital therapeutics for mental health: Ai-driven innovations in personalized treatment approaches. *World J. of Advanced Research and Reviews*.

Kadden, R. (1995). *Cognitive-behavioral coping skills therapy manual: A clinical research guide for therapists treating individuals with alcohol abuse and dependence*. Number 94. US Department of Health and Human Services, Public Health Service.

Larbi, I. B. C., Burchardt, A., and Roller, R. (2023). Clinical text anonymization, its influence on downstream nlp tasks and the risk of re-identification. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111.

Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P., et al. (2023). Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Marques, J. F. and Bernardino, J. (2020). Analysis of data anonymization techniques. In *KEOD*, pages 235–241.

Mogre, N. V., Agarwal, G., and Patil, P. (2012). A review on data anonymization technique for data publishing. *International Journal of Engineering Research & Technology (IJERT)*, 1(10):2278–0181.

Pettersson, E., Borin, L., and Lenas, E. (2024). Swener-1800: A corpus for named entity recognition in 19th century swedish. In *Digital Humanities in the Nordic and Baltic Countries*, volume 6.

Pissarra, D., Curioso, I., Alveira, J., Pereira, D., Ribeiro, B., Souper, T., Gomes, V., Carreiro, A. V., and Rolla, V. (2024). Unlocking the potential of large language models for clinical text anonymization: A comparative study. *arXiv preprint arXiv:2406.00062*.

Ribeiro, B., Rolla, V., and Santos, R. (2023). Incognitus: A toolbox for automated clinical notes anonymization. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 187–194.

Ribeiro, R. A. P. (2023). *Anonimização Automática de Texto Clínico: um estudo sobre técnicas emergentes e métodos de avaliação*. PhD thesis, "ISEP - Instituto Superior de Engenharia do Porto".

Salles, A. A. and Castelo, L. (2023). Privacy and confidentiality in therapeutic process: contributions from bioethics. *Revista Bioética*, 31:e3340PT.

Shamsinejad, E., Banirostam, T., Pedram, M. M., and Rahmani, A. M. (2024). A review of anonymization algorithms and methods in big data. *Annals of Data Science*, pages 1–27.

Supriya, M. and Deepa, A. (2020). Machine learning approach on healthcare big data: a review. *Big data and information analytics*, 5(1):58–75.

União Europeia (2016). Regulamento geral sobre a proteção de dados (gdpr). Acesso em: 31 ago. 2024.

U.S. Department of Health and Human Services (2003). Health insurance portability and accountability act of 1996 (hipaa). Acesso em: 31 ago. 2024.

Vakili, T., Henriksson, A., and Dalianis, H. (2024). 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.

Yadav, V. and Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.