J. Health Inform. 2025 Janeiro-Dezembro; 17

Artigo Original

DOI: https://doi.org/10.59681/2175-4411.v17.2025.1227

JHI JOURNAL OF HEALTH INFORMATICS

Anonimização de textos médicos com processamento de linguagem natural

Anonymization of medical texts with natural language processing

Anonimización de textos médicos con procesamiento del lenguaje natural

Rildo Pinto da Silva[1], Antonio Pazin-Filho[2]

Descritores: anonimização de dados; prontuário médico; processamento de linguagem natural.

RESUMO

Objetivo: Apresentar e avaliar um método de anonimização para prontuários médicos em português, utilizando um modelo de reconhecimento de entidades nomeadas (NER) pré-treinado sem ajuste fino. Método: Aplicou-se o modelo Generalist and Lightweight Model for Named Entety Recognition (GLiNER) para identificar e mascarar informações potencialmente identificadoras (exemplo: nome, idade, organização e cidade) em 27.540 resumos de alta (12.163 pacientes) de um hospital terciário em São Paulo (2017-2023). Avaliou-se a perda de informação com ROUGE F1, BLEU-4, BERTscore e realizou-se análise humana de erros em amostra aleatória (N=400). Resultado: A análise humana mostrou falha de anonimização de dois casos (0,50%) permitindo a identificação do paciente ou do assistente. As métricas quantitativas indicaram preservação da utilidade textual (mediana BERTscore: 0,76) Conclusão: O método é eficiente, mas não perfeito, evidenciando a necessidade de uma abordagem híbrida de anonimização (automático e validação humana) para conformidade com a Lei Geral de Proteção de Dados Pessoais. Pode ser usado como um passo inicial para a criação de conjuntos de dados médicos necessários ao desenvolvimento do processamento de linguagem natural no Brasil.

Keywords: data anonymization; medical records; natural language processing.

ABSTRACT

Objective: To present and evaluate an anonymization method for medical records in Portuguese, using a pre-trained named entity recognition (NER) model without fine-tuning. Method: The GLiNER (Generalist and Lightweight Model for Named Entity Recognition) model was applied to identify and mask potentially identifying information (example: name, age, organization, and city) in 27,540 discharge summaries (12,163 patients) from a tertiary hospital in São Paulo (2017-2023). Information loss was evaluated with ROUGE F1, BLEU-4, BERTscore, and human analysis of errors was performed on a random sample (N=400). Result: Human analysis showed anonymization failure in two cases (0.50%), allowing the identification of the patient or the assistant. Quantitative metrics indicated preservation of textual utility (median BERTscore: 0.76). Conclusion: The model is efficient but not perfect, highlighting the need for hybrid anonymization approach (automatic and human validation) to comply with the General Law for the Protection of Personal Data. It can be used as a step for creationing necessary medical datasets for the development of natural language processing in Brazil.

Descriptores: anonimización de datos; historias clínicas; procesamiento del lenguaje natural.

RESUMEN

Objetivo: Presentar y evaluar un método de anonimización para historias clínicas en portugués, utilizando un modelo de reconocimiento de entidades nombradas (NER) pre-entrenado sin ajuste fino. Método: Se aplicó el modelo GLiNER (Generalist and Lightweight Model for Named Entity Recognition) para identificar y enmascarar Información Personal Identificable (IPI) (ej.: nombre, edad, org., ciudad) en 27.540 informes de alta (12.163 pacientes) de un hospital terciario en São Paulo (2017-2023). Se evaluó la pérdida de información con ROUGE F1, BLEU-4, BERTscore y análisis humano de errores en muestra aleatoria (N=400). . Resultado: El análisis humano reveló fallos de anonimización en dos casos (0,50%) permitiendo la identificación del paciente o del professional asistente. Las métricas cuantitativas indicaron preservación de la utilidad textual (mediana BERTscore: 0,76). Concluyendo: El modelo es eficiente pero no perfecto, evidenciando la necesidad de un enfoque híbrido de anonimización (automática y validación humana) para conformidad con la Ley General de Protección de Datos Personales. Puede utilizarse como un paso hacia la creación de conjuntos de datos médicos necesarios para el desarrollo del procesamiento del lenguaje natural en Brasil.

[1]Doutorando em clínica médica do Departamento de Clínica Médica da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. https://orcid.org/0000-0001-5718-2747
[2]Professor Titular do Departamento de Clínica Médica da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. https://orcid.org/0000-0001-5242-329X

Autor Correspondente: **Rildo Pinto da Silva**
e-mail: **rildo.silva@alumni.usp.br**

Artigo recebido: 21/08/2024
Aprovado: 07/05/2025

INTRODUÇÃO

O uso das informações dos prontuários médicos tem se ampliado em paralelo ao desenvolvimento dos modelos de processamento de linguagem natural (PLN). Esses dados têm sido utilizados nas tarefas de reconhecimento de entidades nomeadas[1], predição de risco, resumo de textos médicos, tradução[2], sistemas de perguntas e respostas e análise de sentimentos[3].

Para esse uso, é imprescindível a anonimização dos dados dos pacientes. Segundo a Lei Geral de Proteção de Dados Pessoais (LGPD), os dados de saúde são sensíveis[4]. A LGPD define anonimização como "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo".

Sweeney estabeleceu, em 2002, as bases teóricas de um processo de anonimização chamado k-anonymity[5]. A autora categorizou os dados em dois tipos: identificadores explícitos (como nome, endereço e telefone – chamados diretos na LGPD) e quase-identificadores (correspondentes aos indiretos na LGPD). Os quase-identificadores são atributos que, combinados, (por exemplo, endereço e data de nascimento) poderiam identificar pessoas em um conjunto de dados. O processo k-anonymity foi desenhado para dados estruturados[5].

Com o desenvolvimento do PLN e o acesso a dados não estruturados, surgiram modelos de anonimização baseados em aprendizado de máquina. A tentativa é automatizar o processo, devido ao alto custo da anonimização manual[6-8]. Esses modelos não são perfeitos por causa de sua característica probabilística[9]. Para contornar esses problemas, foram propostos métodos de criação de textos sintéticos baseados em textos reais[10].

A anonimização é um passo do processamento do texto de um prontuário para que esses dados sejam utilizados no treinamento, ajuste fino (fine-tuning) ou desenvolvimento de modelos de aprendizado de máquina. Também permite a criação de conjuntos de dados (dataset) públicos que podem ser utilizados no desenvolvimento de novos modelos. É importante que a utilidade do dado seja preservada[11] após a anonimização.

Um exemplo de conjunto de dados públicos resultante da anonimização é o Medical Information Mart for Intensive Care (MIMIC) atualmente na quarta versão[12]. Ele cobre uma década de dados clínicos não estruturados (2008-2019), com 73.181 internações e 50.920 pacientes atendidos no Beth Israel Deaconess Medical Center e abrange atendimentos na unidade de emergência ou no centro de terapia intensiva do hospital de pacientes maiores de 18 anos. Esse conjunto de dados é amplamente utilizado nos estudos de aprendizado de máquina e PLN[13-14].

O objetivo deste trabalho é apresentar um método de anonimização de informações de pacientes de prontuários médicos utilizando um modelo de reconhecimento de entidades nomeadas (named entity recognition – NER) sem ajuste fino, incluindo uma avaliação quantitativa da perda de informação pós-anonimização e uma avaliação do resultado por análise humana em uma amostra de internações.

Espera-se, com isso, contribuir para o desenvolvimento de conjuntos de dados de textos médicos para ensino e pesquisa em processamento de linguagem natural em português. A descrição do artigo seguiu o guia Reporting of Studies using Observational Routinely-collected health Data (RECORD). Este estudo foi aprovado por comitê de ética em pesquisa (CAAE: 75360523.7.0000.5440).

METODOLOGIA

Trata-se de um estudo observacional descritivo, utilizando dados secundários de pacientes atendidos em hospital de ensino terciário do estado de São Paulo. Foram selecionados resumos de alta (elemento_prontuário = epicrise) do período de 2017 a 2023. Os prontuários foram importados para um banco de dados relacional por meio de um script Python.

Tratamento dos dados

Variáveis selecionadas: cod_paciente (número de registro de paciente), seq_atendimento (sequência de atendimentos de um mesmo paciente), seq_item (ordena os registros de atendimento), dsc_id_informacao (categorias dos registros do prontuário) e ctu_informacao (campo texto não estruturado - corresponde ao texto do prontuário). Esse campo foi tratado, excluindo caracteres especiais, espaços duplos e marcadores especiais (exemplo: -> e ==). Foram mantidos os marcadores ponto final, ponto e vírgula e dois pontos. Foram excluídos os registros nulos (n = 33.669).

Foram selecionadas apenas as categorias dos registros dos prontuários (dsc_id_informacao) que tivessem mais de 20.000 registros preenchidos, a fim de focar nas categorias mais representativas. As categorias selecionadas foram: história clínica, exame físico, exames e cirurgias realizadas, diagnósticos, planejamento terapêutico, tempo e local de internação e retornos agendados do paciente. No final do tratamento e seleção obteve-se 914.056 registros de 12.163 pacientes em 27.540 internações.

Modelo de anonimização

O texto do prontuário foi processado em quatro fases. O objetivo foi a exclusão dos identificadores e quase-identificadores, mantendo a utilidade do conjunto de dados.

Utilizou-se o conceito de reconhecimento de entidades nomeadas (named entity recognition – NER). Cada entidade correspondeu a um identificador ou quase-identificador e serviu como rótulo (label) do prompt do modelo. Nas outras três fases, informações

sobre diagnósticos, tempo de internação e idade foram processadas e reincorporadas ao prontuário anonimizado na fase 01. O modelo é detalhado a seguir.

Primeira fase

Para a identificação dos NER, foi utilizado o modelo Generalist and Lightweight Model for Named Entety Recognition (GLiNER). Trata-se de um modelo do tipo encoder pré-treinado baseado no modelo de linguagem bidirecional deBERTa, contendo dois componentes que representam: 1) os vetores (embeddings) de partes dos textos a serem identificados (span-based approach) e 2) os vetores correspondentes aos rótulos das entidades procuradas (label prompting). É um modelo pequeno que necessita de poucos recursos computacionais.

Como se trata de classificação de NER para anonimização, foi selecionado o modelo gliner-multi-pii-v1 específico para a identificação de dados pessoais, baixado do Hugging Face (https://huggingface.co/urchade/gliner_multi_pii-v1). O parâmetro threshold (limiar) controla a sensibilidade do modelo. Foram testados três limiares diferentes (0,30; 0,50 e 0,70). Optou-se pelo limiar mais baixo (0,30) pois este identificou maior número de entidades. Utilizaram-se rótulos em inglês para o processo de 'label prompting' do NER. Testes em português mostraram que o modelo não encontra todas as categorias dos NER listadas no prompt. O modelo tem limite de processamento de tokens, assim os textos foram particionados em no máximo 150 tokens, utilizando-se a biblioteca NLTK e o método word_tokenize.

Foram pesquisados 28 possíveis identificadores (entidades, como por exemplo: nome, idade, organização e cidade – a lista completa está no Apêndice) descritos em inglês.

Segunda Fase

O mesmo modelo serviu para identificação da Classificação Internacional de Doenças (CID) de alta, usando como prompt 'international classification of diseases'. Foram excluídos os registros sem CID. O resultado, um dataframe com CID de alta de cada internação, foi utilizado no cálculo do índice de comorbidade de Charlson. Para isso, foi empregada a biblioteca comorbidipy (https://github.com/vvcb/comorbidipy) desconsiderando a correção pela idade. Os cálculos foram validados manualmente pelo autor principal. Nesta fase também foi selecionado o CID mais importante. Para isso, criou-se um indicador de CID mais frequente de internação calculado pela divisão da quantidade de internações de determinado CID pelo total de internações do estado de São Paulo de 2019. Foram utilizados no cálculo somente CID com três dígitos. Os dados foram obtidos no Observatório de Política e Gestão Hospitalar da Fiocruz (https://observatoriohospitalar.fiocruz.br/). Esses dados foram cruzados com os CID identificados nesta fase, sendo escolhido aquele com maior indicador como o CID representante do diagnóstico daquela internação.

Terceira Fase

Foi extraída a informação de tempo de permanência para cada internação do formulário do resumo de alta. Foram excluídos os registros incorretos (descrições que deveriam ser incluídas em outros formulários).

Quarta Fase

Foram criadas faixas etárias de 10 anos, iniciando-se na faixa 0-9 anos até a faixa 70 e mais anos. O agrupamento ocorreu por paciente, assim aqueles que mudaram de faixa etária em diferentes internações foram alocados à menor faixa etária. Foi considerada a idade mais próxima do início do texto do prontuário, dado que o texto do tipo: "paciente [nome], [idade] anos." é um formato comum em prontuários.

Validação do modelo

Para a validação, foi selecionada uma amostra aleatória de 400 internações. Nesta análise, buscou-se por identificadores que deveriam ter sido mascarados, ou seja, não ocorreu a anonimização. Foi anotado, para aquela internação, qual identificador deveria ter sido substituído. Não foi analisado se a entidade que substituiu o identificador era correta ou não (exemplo 95 anos por [gênero]) por extrapolar o objetivo deste estudo.

Para avaliar a perda de informação, utilizaram-se métricas quantitativas comparando o texto processado com o original. Foi utilizado o Recall-Oriented Understudy for Gisting Evaluation (Rouge[15]), técnica que avalia a sensibilidade considerando sobreposição de palavras únicas (unigramas, Rouge-1 – palavras individuais), duplas (bigramas, Rouge-2) ou do texto mais longo (Rouge-L). Para cada uma destas variantes, foi calculado o F1-score. Outra métrica usada foi o BLEU (Bilingual Evaluation Understudy[16]). Trata-se de um método modificado de cálculo de precisão de n-gramas em sentenças, utilizado originalmente para avaliar qualidade de tradução automática. Foi utilizado o Bleu-4 (4 palavras). Essas duas métricas não avaliam a equivalência semântica entre os textos, para isso foi empregado o BERTscore[17]. As métricas foram calculadas para cada internação sendo apresentadas as medianas e os intervalos interquartis (IIQ).

RESULTADOS

O Quadro 01 mostra um exemplo do resultado da anonimização. Observa-se que a maioria dos identificadores foi corretamente substituída pelas suas respectivas entidades, como, por exemplo, nome, idade e datas. Contudo, um identificador (ue – unidade de emergência) não foi identificado (falha de anonimização) e outros foram identificados, mas substituídos por entidades incorretas (nome de cidade por address e não city; resultado de exame por date ou social security number).

Quadro 01 – Exemplo de um resumo de alta de paciente com os identificadores substituídos pelas entidades indicadas entre colchetes.

sequencia n 1 **[person]** , **[age]** , procedente de **[address]** adm ue : **[date]** adm uco : **[date]** ap : has ex tbg (menos de 10 anos maço , parou há **[time duration]**) medicações de uso prévio : enalapril 20mg 1 0 1 exames de relevancia na internacao : cate **[date]** : cd sem lesao obstrutiva dp sem lesao obstrutiva tce sem lesao obstrutiva da sem lesao obstrutiva diag sem lesao obstrutiva cx nao dominante sem lesao obstrutiva 1 marginal de moderada importancia apresenta lesao obstrutiva severa 80 % extensa desde o ostio ate o segmento distal (71 a 99 %) tipo b2 . eco **[date]** : sem laudo no sistema até o momento sequencia n 3 caso : iamssst timi 4/ grace 98 > dor precordial com irradiacao para mse na noite do dia **[date]** > tropo **[date]** : 0,51 > ckmb : **[social_security_number]** resumo da internacao : **[person]** admitido no dia **[date]** em demanda espontanea relato de ter apresentado dor de intensidade **[date]** com irradiacao para mse e mandibula durante a realizacao de exercicio fisico negava outras queixas espontaneas realizadas medidas para sca resultado de troponina positiva no dia **[date]** queixa de cefaleia e nova dor precordial e sensacao de mal estar levada a sec para monitorizacao clinica porem aprsentou melhora do quadro suspensa estatina por cpk de 1465. optamos por retomar uso devido a sca com beneficio com uso de estatina realizado cateterismo anotado em exames de relevancia abaixo . nao realizada angioplastia sequencia n 4 exame fisico da alta : bom estado geral , corada , hidratada , acianotica , afebril , anicterica , consciente e orientada no tempo e espaço ssvv : pa : 90x60 (70) mmhg fc : 68 bpm sat : 96 % em aa ecg : 15 ar : mv presente , simetrico , sem ra acv : rcr 2t , bnf , sem sopros abdome : semi globoso , flacido , sem sinais de irritacao peritoneal . sem massas ou vmg extremidades : quentes , bem perfundidas , sem sinais de tvp curativo em msd , boa perfusao em mao curativo neuro : pupilas isocoricas e fotoreagentes sequencia n 5 tropo **[date]** : 0,51 > ckmb : 31 168 167 100 56 sequencia n 6 cate **[date]** eco **[date]** sequencia n 7 seguimento em **[organization]** sequencia n 8 oriento habitos de vida saudaveis seguimento com cardiologista na **[organization]** sequencia n 9 programação : tratamento clínico otimizado forneço relatório médico e receitas oriento habitos de vida saudável seguimento : como **[person]** nao realizou angioplastia , oriento encaminhamento via **[organization]** para controle de fatores de risco cardiovascular levar este relatório em **[organization]** para encaminhamento & comorbidity_score: 1.0 & nome_cid: I21.9 Infarto agudo do miocárdio não especificado & estadia: 02 d & df_faixa_etaria: 60-69

Fonte: Elaborado pelos próprios autores. Em negrito – os identificadores foram substituídos pelas entidades corretas, em vermelho sublinhado foram substituídos por entidades incorretas e em amarelo não houve a identificação dos identificadores e, portanto, houve falha de anonimização. & comorbidity_score – escore de risco de Charlson, & nome_cid – CID mais importante indicado no resumo de alta e & df_faixa_etaria – faixa etária do paciente.

A análise humana da amostra de 400 internações mostrou que 103 (25,75%) apresentaram falha de anonimização, ou seja, algum identificador não foi substituído pela respectiva entidade. A maioria dos casos corresponde a números de exame (n=42 – 10,50%) ou datas (n=47 – 11,75%). Em sete (1,75%) internações, o nome parcial do paciente permaneceu identificado e, em duas (0,50%), o registro do paciente no hospital. No caso dos profissionais assistentes, também houve falha na anonimização (seis - 1,50% - nomes parciais de assistentes e três – 0.75% - registros). Os resultados por entidade são apresentados na Tabela 01.

Tabela 01 – Distribuição das internações com falha de anonimização, por tipo de entidade: Número (N), Percentual (%) e Observações - Análise humana de amostra de 400 prontuários, Hospital de Ensino Terciário, São Paulo, Brasil

Entidade	Internações com falha de anonimização	%	Observação
registro paciente	2	0,50%	Nos dois casos o registro completo não foi desidentificado
número telefone	2	0,50%	Todos os casos correspondem a registros completos
registro assistente	3	0,75%	Dois casos são registros COREN e 01 provável COREN (1)
nome assistente	6	1,50%	Todos os nomes não desidentificados são parciais
nome paciente	7	1,75%	Todos os nomes não desidentificados são parciais
cidade	9	2,25%	Todos os nomes não desidentificados são completos
organização	10	2,50%	Não foram desidentificados nomes de especialidades, unidade de emergência e o nome do hospital no campus
número exames	42	10,50%	Todos os casos correspondem a registros completos (2)
datas	47	11,75%	Não foram desidentificadas datas parciais (dd/mm ou aa) e completas (dd/mm/aaaa)
demais entidades	0	0,00%	

Fonte: Elaborado pelos próprios autores. (1) inferiu-se tratar de registro porque é um número que segue o nome do profissional assistente. (2) são números não sequenciais que ocorrem após a expressão número do exame. % calculado sobre N=400. Uma internação pode ter mais de uma falha de anonimização. Entidade corresponde a tradução para o português dos rótulos NER utilizados no prompting do modelo.

Os textos com falhas de anonimização são mais longos. A mediana do número de palavras dos textos com falha foi de 553 palavras (IIQ 299-863), enquanto a mediana dos textos sem falha foi de 164 palavras (IIQ 78-409).

Em uma (0,25%) internação, é possível identificação direta do paciente (o modelo não anonimizou parte do nome, data de nascimento e o telefone de recados). Em outra, é possível identificação direta do assistente (não anonimizou parte do nome, número de registro profissional e número de exame). O Quadro 02 mostra o caso do paciente passível de identificação – o registro é parcial e foi modificado manualmente para não permitir a identificação.

Quadro 02 – Internação com erro de anonimização que permite identificar o paciente - prontuários médicos de hospital de ensino terciário – São Paulo – Brasil.

sequencia n 3 relatório de transferencia para **[organization]** nome : **[person]** idade : **[age]** nascimento : dd/mm/aaaa nome da mãe : **[person]** cartão sus : **[social security number]** endereço do paciente : **[address]** . **[city]** telefone : **[phone number]** / celular **[mobile phone number]** / recados nn nnnnnnnnn procedência : **[city]** internação pqu ue : **[date]** admissao epib : **[date]** transferencia **[organization]** : **[date]** identificação : nome parcial do paciente , **[age]** , natural e procedente de **[city]** , aposentado há 1,5 ano por {diagnóstico omitido}, {local trabalho omitido}. cursou ensino médio até o 2º colegial. **[religion]** , frequenta a **[religion]** . divorciado há [time duration] , tem **[person]** (**[age]** e **[age]**) . **[person]** : **[person]** contratado pela família para acompanhá lo na **[organization]** durante internação {nome parcial do cuidador omitido}

Fonte: Elaborado pelos próprios autores. Em negrito – os identificadores foram substituídos pelas entidades corretas, em vermelho sublinhado foram substituídos por entidades incorretas e em amarelo não houve a identificação dos identificadores e, portanto, houve falha de anonimização. & comorbidity_score – escore de risco de Charlson, & nome_cid – CID mais importante indicado no resumo de alta e & df_faixa_etaria – faixa etária do paciente

Os resultados das métricas de avaliação de desempenho do modelo de linguagem são apresentados na Tabela 02. Para o F1-score do Rouge-1, o valor mínimo foi de 0,15, enquanto a mediana foi de 0,70, com um IIQ de 0,54 a 0,80 e um valor máximo de 0,90. Similarmente, para o F1-score do Rouge-2, os valores variaram de um mínimo de 0,09 a um máximo de 0,87, com uma mediana de 0,65 (IIQ de 0,49 a 0,76). O F1-score do Rouge-L apresentou um mínimo de 0,10, uma mediana de 0,67 (IIQ de 0,48 a 0,78) e um máximo de 0,89. Os resultados do BLEU-4 foram piores comparados às outras métricas (mediana de 0,55 e IIQ 0,31 a 0,70). Já o BERTscore mostrou perda moderada de informação (mediana de 0,76 e IIQ 0,73 a 0,79). A análise humana mostrou que os textos acima do escore mediano do BERTscore permitiram o entendimento adequado do prontuário.

Tabela 02 – Estatísticas descritivas (Mínimo, Mediana, IIQ, Máximo) das métricas de avaliação de desempenho da anonimização (ROUGE F1, BLEU-4, BERTscore) prontuários médicos de hospital de ensino terciário – São Paulo – Brasil.

Métricas	Mínimo	Mediana	IIQ	Máximo
Rouge-1	0,1485	0,6989	0,5444-0,7986	0,9027
Rouge-2	0,0916	0,6463	0,4864-0,7574	0,8740
Rouge L	0,1039	0,6698	0,4774-0,7812	0,8947
BLEU-4	0,0004	0,5481	0,3066-0,7008	0,8546
BERTscore	0,5985	0,7606	0,7277-0,7859	0,9497

Fonte: Elaborado pelos próprios autores. Valores para ROUGE referem-se aos F1-scores das variantes ROUGE-1, ROUGE-2 e ROUGE-L

DISCUSSÃO

O modelo foi capaz de anonimizar dados dos pacientes, substituindo-os por entidades correspondentes, sem perda significativa da informação do prontuário médico. Contudo, ele não foi perfeito. Foi possível identificar dois casos – um paciente e um profissional assistente – na análise humana.

O desempenho do modelo foi pior na anonimização dos números de exames e datas. Esse problema pode ser facilmente corrigido usando expressões regulares, sem perda de informação. Os demais identificadores apresentaram erros menores de anonimização, mas tal afirmação esconde um problema importante: eles falharam na anonimização completa.

É comum o uso de métricas do tipo revocação (recall), precisão, F1-score e acurácia para avaliar resultados de modelos de anonimização utilizando textos previamente anotados[6-8]. Lee et al., por exemplo, criaram um conjunto de dados anotando 1.700 resumos de alta, sobre os quais aplicaram diferentes modelos de anonimização. Nenhum deles foi perfeito para anonimizar o nome do paciente[18]. Já Vakili et al., apresentaram resultados mistos – anonimização perfeita para o primeiro nome e de 98% para o último nome[8]. O que se busca é o estado da arte dos modelos, cujas métricas de avaliação de resultados são automáticas. Este estudo, usando um modelo simples, foi muito bem na desidentificação, mas não foi perfeito, como observado na literatura.

Essa imperfeição, com potenciais problemas éticos, é reconhecida nos diferentes estudos de uso dos modelos[19-20] sem, contudo, afetar a busca pelo modelo estado-da-arte. A natureza inerentemente probabilística dos grandes modelos de linguagem[9], implica que não se pode garantir a ausência total de erros. Essa limitação levanta, portanto, questionamentos sobre a viabilidade de se alcançar a robustez exigida pela LGPD em processos de anonimização que dependam unicamente de modelos estatísticos, dado o risco de falhas residuais como as identificadas neste trabalho.

Um método híbrido – método simples de anonimização mais validação humana deve ser o melhor processo de tratamento de informações de saúde. Haverá ganhos no tempo da análise humana, redução do erro do analista e maior segurança para o uso dos dados. Em contrapartida, haverá um custo da análise humana, que não pode ser minimizado, pois os dados não estruturados dos prontuários dificultam essa análise.

Segundo Gadotti et al.[11], os modelos tendem a ter melhor desempenho em dados vistos e, portanto, aprendidos. O modelo usado neste estudo foi treinado em várias línguas, inclusive português. A análise humana das internações revelou que três nomes não anonimizados são incomuns. Chamou atenção ainda que, um deles, é

um nome de origem indígena, provavelmente não foi identificado, porque o modelo não foi treinado para reconhecê-lo como nome.

Essa situação, reforça o cuidado no uso ético e legal da informação, mas expõe também a necessidade premente de tropicalização de modelos de PLN, através de treinamento ou ajuste fino, o que somente será possível com a disponibilização de conjuntos de dados em português brasileiro adequadamente anotados.

Outro aspecto importante é a troca entre a segurança da desidentificação e o custo de revisão dos dados. Das 400 internações manualmente analisadas, sete nomes parciais permitiram a identificação de um paciente e um assistente. Essa análise foi trabalhosa – levou-se aproximadamente 40 horas buscando apenas erros de anonimização. Os textos não eram extensos e a exclusão prévia dos identificadores facilitou a análise. Não por acaso, são raros conjuntos de dados anotados. No Brasil, até onde sabemos, existe apenas um conjunto baseado em prontuários médicos[21] que é, em parte, afeito ao estudo de reconhecimento de entidades nomeadas.

Provavelmente cientes dessa troca, alguns pesquisadores acreditam que a anonimização perfeita não deve restringir o acesso à informação – o custo da barreira ao progresso do desenvolvimento dos métodos de aprendizado de máquina é maior que o risco de reidentificação de dados de saúde públicos[22] e assim os dados deveriam ser liberados sem que houvesse anonimização perfeita. Neste estudo, observou-se que a maioria dos casos considerados identificáveis corresponde a quase-identificadores que, mesmo em conjunto, não devem identificar um paciente (exemplo: as datas de exames e os números de exames).

Para resolver o problema, surgiram técnicas sofisticadas de inteligência artificial que podem gerar dados artificiais baseados em dados reais[10]. O problema deriva para outro: a questão da garantia de que os dados artificiais representem os reais. As métricas são novamente automáticas.

Em última instância, quer-se evitar a análise humana. Um exemplo de um conjunto de dados amplamente usado pode mostrar que isso não é fácil. O processamento do MIMIC-IV tem uma série de mecanismos para anonimização. São aplicados dois algoritmos diferentes e o resultado é revisado por humanos. O acesso aos dados é controlado e liberado após treinamento. Uma breve pesquisa no PubMed mostra 231 artigos usando o MIMIC-IV. O aparente sucesso de seu uso pode indicar que a solução conservadora, ainda que custosa, pode ser a melhor.

O modelo apresentou resultados adequados quanto à perda da informação. Contudo, observou-se variação nas diferentes métricas, provavelmente a combinação de uma nota de corte baixa com textos curtos, leva ao mascaramento de muitas palavras com consequente perda de significado do texto. O grau de variação nas métricas que se baseiam em n-gramas reforça a hipótese. Esse dado pode indicar que a construção de um conjunto de dados precisará de textos mais longos para evitar a perda de informação dos registros mais curtos.

Este estudo apresenta algumas limitações. Ele não é generalizável, já que foi utilizado um prontuário de um único hospital. Contudo, o modelo é simples e adaptável. Outra limitação é que a análise humana foi feita sobre um conjunto aleatório pequeno de internações por um único pesquisador, que pode não ter identificado todos os erros do modelo. Contudo, foi demonstrado que o erro existe, que a liberação desse conjunto de dados seria antiética e que se advoga por um modelo híbrido de anonimização. As métricas utilizadas para avaliar a perda de informação foram automáticas. São métricas amplamente utilizadas e o uso do BERTscore melhora a avaliação da qualidade semântica do texto. Contudo há limitação desses resultados por não ter havido avaliação humana ainda que o BERTscore tenha apresentado alto índice de similaridade.

Uma vantagem do modelo é que ele é simples, baseado na aplicação de um modelo sem nenhum tipo de ajuste fino, utilizando diretamente os prompts definidos. Outro diferencial foi a análise humana. Seu resultado valoriza a discussão da acurácia de modelos ao expor os riscos da anonimização automática, que, embora pequenos, são reais. Mostrou também o tempo de análise das 400 internações e o percentual de erros de anonimização contribuindo para avaliar possíveis custos de análise de um conjunto de dados.

Alguns estudos adicionais são necessários. Avaliar os pontos de corte e identificar se eles podem facilitar a anotação humana. Definir qual tamanho de texto é adequado para desidentificação com manutenção de informações adequadas ao uso. Este estudo foi restrito a algumas características comuns da internação para garantir que não houvesse perda de informação, há espaço para inclusão de outras que vão enriquecer o conjunto de dados. Pode ser interessante avaliar o mesmo método para identificar entidades nomeadas (NER) para outras aplicações.

CONCLUSÃO

Demonstrou-se que o modelo híbrido de anonimização – uma fase automática seguida da análise manual – é necessário no contexto atual da LGPD para o uso ético e legal de informações sensíveis de saúde. O modelo simples utilizado, sem ajuste fino, não foi perfeito, resultado este semelhante ao que se observa com modelos mais complexos. Tal afirmação só foi possível através da análise humana. Dado que a análise humana é necessária, é melhor o uso de um modelo simples de fácil aplicabilidade.

Espera-se que o artigo contribua para que grupos de pesquisa se engajem na aplicação do modelo híbrido para produzir os necessários conjuntos de dados em português brasileiro. Também parece claro que o desenvolvimento de conjuntos de dados precisará de textos longos, que apresentam menor risco de perda de informação com a desidentificação.

REFERÊNCIAS

1. Landolsi MY, Hlaoua L, Ben Romdhane L. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst 2023; 65: 463–516.
2. Hossain E, Rana R, Higgins N, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Comput Biol Med 2023; 155: 106649.
3. Luo X, Deng Z, Yang B, et al. Pre-trained language models in medicine: A survey. Artif Intell Med 2024; 154: 102904.
4. Brasil, Lei no. 13.709, de 14 de Agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD): LGPD, 2018.
5. Sweeney L. k-Anonymity: A model for protecting privacy. Int. J. Unc. Fuzz. Knowl. Based Syst. 2002; 10: 557–570.
6. Liu J, Gupta S, Chen A, et al. OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study. J Med Internet Res 2023; 25: e48145.
7. Johnson AEW, Bulgarelli L, Pollard TJ. Deidentification of free-text medical records using pre-trained bidirectional transformers. Proc ACM Conf Health Inference Learn (2020) 2020; 2020: 214–221.
8. Vakili T, Henriksson A, Dalianis H. End-to-end pseudonymization of fine-tuned clinical BERT models Privacy preservation with maintained data utility. BMC Med Inform Decis Mak 2024; 24: 162.
9. Minaee S, Mikolov T, Nikzad N, et al. Large Language Models: A Survey, 2024.
10. Yoon J, Drumright LN, van der Schaar M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE J. Biomed. Health Inform. 2020; 24: 2378–2388.
11. Gadotti A, Rocher L, Houssiau F, et al. Anonymization: The imperfect science of using data while preserving privacy. Sci Adv 2024; 10: eadn7053.
12. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023; 10: 1.
13. Nigo M, Rasmy L, Mao B, et al. Deep learning model for personalized prediction of positive MRSA culture using time-series electronic health records. Nat Commun 2024; 15: 2036.
14. Falter M, Godderis D, Scherrenberg M, et al. Using natural language processing for automated classification of disease and to identify misclassified ICD codes in cardiac disease. Eur Heart J Digit Health 2024; 5: 229–234.
15. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out, pp. 74–81. Barcelona, Spain: Association for Computational Linguistics.
16. Papineni K, Roukos S, Ward T, et al. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. (ed Isabelle P, Charniak E and Lin D), pp. 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
17. Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT, 2019.
18. Lee Y-Q, Chen C-T, Chen C-C, et al. Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study. J Med Internet Res 2024; 26: e48443.
19. Preiksaitis C, Ashenburg N, Bunney G, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Med Inform 2024; 12: e53787.
20. Park Y-J, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak 2024; 24: 72.
21. Oliveira LESE, Peters AC, Da Silva AMP, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics 2022; 13: 13.
22. Seastedt KP, Schwab P, O'Brien Z, et al. Global healthcare fairness: We should be sharing more, not less, data. PLOS Digit Health 2022; 1: e0000102.

APÊNDICE

Lista de identificadores pesquisados no modelo de desidentificação

labels = ['person', 'age', 'organization', 'city', 'address', 'postal code', 'date', 'gender', 'religion', 'phone number', 'email', 'username', 'social security number', 'time duration', 'mobile phone number', 'cpf', 'identity card number', 'national id number', 'registration number', 'student id number', 'digital signature', 'social media handle', 'license plate number', 'cnpj', 'serial number', 'fax number', 'identity document number', 'social_security_number']