UNIVERSIDADE FEDERAL DE MINAS GERAIS – UFMG

ESCOLA DE CIÊNCIA DA INFORMAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

GUILHERME FRANCIS DE NORONHA

TRATAMENTO DA INFORMAÇÃO DE SAÚDE PARA ATENDIMENTO À NECESSIDADE DE PRIVACIDADE: DESIDENTIFICAÇÃO TEXTUAL DE DOCUMENTOS CLÍNICOS NA LÍNGUA PORTUGUESA DO BRASIL

Belo Horizonte

2022

GUILHERME FRANCIS DE NORONHA

TRATAMENTO DA INFORMAÇÃO DE SAÚDE PARA ATENDIMENTO À NECESSIDADE DE PRIVACIDADE: DESIDENTIFICAÇÃO TEXTUAL DE DOCUMENTOS CLÍNICOS NA LÍNGUA PORTUGUESA DO BRASIL

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais para a obtenção do grau de Doutor, área de concentração Ciência da Informação.

Linha de Pesquisa: Gestão & Tecnologia da Informação e Comunicação

Orientador: Maurício Barcellos Almeida

Belo Horizonte

2022

N852T Noronha, Guilherme Francis.
Tratamento da informação de saúde para atendimento à necessidade de privacidade [recurso eletrônico] : desidentificação textual de documentos clínicos na língua portuguesa do Brasil / Guilherme Francis de Noronha . - 2022.
1 recurso online (f.173 : il., color.) : pdf.

Orientador: Maurício Barcellos Almeida.

Tese (doutorado)– Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 143-155.
Anexo: f. 159-171.
Exigência do sistema: Adobe Acrobat Reader.

1.Ciência da informação – Teses. 2. Registros médicos - Teses. 3. Proteção de dados - Teses. 4. Processamento de linguagem natural (Computação) - Teses. 5. Aprendizado do computador - Teses. I. Almeida, Maurício Barcellos. II. Universidade Federal de Minas Gerais. Escola de Ciência da Informação. III. Título.

CDU 004.056

Ficha catalográfica. Vanessa Marta de Jesus - CRB/6-2419

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPG-GOC

FOLHA DE APROVAÇÃO

TRATAMENTO DA INFORMAÇÃO DE SAÚDE PARA ATENDIMENTO À NECESSIDADE DE PRIVACIDADE: DESIDENTIFICAÇÃO TEXTUAL DE DOCUMENTOS CLÍNICOS NA LÍNGUA PORTUGUESA DO BRASIL

GUILHERME FRANCIS DE NORONHA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia.

Aprovada em 01 de junho de 2022, por videoconferência, pela banca constituída pelos membros:

Prof(a). Mauricio Barcellos Almeida (Orientador)
ECI/UFMG

Prof(a). Fernanda Farinelli
PUC-MG

Prof(a). Heliana Ribeiro de Mello
FALE/UFMG

Prof(a). Jeanne Louize Emygdio
PUC-MG

Prof(a). Zilma Silveira Nogueira Reis
UFMG

Prof(a). Eduardo Ribeiro Felipe
UNIFEI - Universidade Federal de Itajubá

Belo Horizonte, 01 de junho de 2022.

Documento assinado eletronicamente por **Mauricio Barcellos Almeida**, **Professor do Magistério Superior**, em 03/06/2022, às 15:54, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Ribeiro Felipe**, **Usuário Externo**, em 06/06/2022, às 15:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernanda Farinelli**, **Usuário Externo**, em 06/06/2022, às

15:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heliana Ribeiro de Mello**, **Professora do Magistério Superior**, em 06/06/2022, às 16:59, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jeanne Louize Emygdio**, **Usuária Externa**, em 08/06/2022, às 12:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Zilma Silveira Nogueira Reis**, **Professora do Magistério Superior**, em 24/08/2022, às 15:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1505503** e o código CRC **EFB6877E**.

Referência: Processo nº 23072.231453/2022-51 SEI nº 1505503

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPG-GOC

ATA DA DEFESA DE TESE DO ALUNO

GUILHERME FRANCIS DE NORONHA

Realizou-se, no dia 01 de junho de 2022, às 10:00 horas, por videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *TRATAMENTO DA INFORMAÇÃO DE SAÚDE PARA ATENDIMENTO À NECESSIDADE DE PRIVACIDADE: DESIDENTIFICAÇÃO TEXTUAL DE DOCUMENTOS CLÍNICOS NA LÍNGUA PORTUGUESA DO BRASIL*, apresentada por GUILHERME FRANCIS DE NORONHA, número de registro 2018666724, graduado no curso de CIÊNCIA DA COMPUTAÇÃO, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Mauricio Barcellos Almeida - ECI/UFMG (Orientador), Prof(a). Fernanda Farinelli - PUC-MG, Prof(a). Heliana Ribeiro de Mello - FALE/UFMG, Prof(a). Jeanne Louize Emygdio - PUC-MG, Prof(a). Zilma Silveira Nogueira Reis - UFMG, Prof(a). Eduardo Ribeiro Felipe - UNIFEI - Universidade Federal de Itajubá.

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 01 de junho de 2022.

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Eduardo Ribeiro Felipe**, **Usuário Externo**, em 06/06/2022, às 15:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernanda Farinelli**, **Usuário Externo**, em 06/06/2022, às 15:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mauricio Barcellos Almeida**, **Professor do Magistério Superior**, em 06/06/2022, às 16:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heliana Ribeiro de Mello**, **Professora do Magistério Superior**, em 06/06/2022, às 16:59, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jeanne Louize Emygdio**, **Usuária Externa**, em 08/06/2022, às 12:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Zilma Silveira Nogueira Reis**, **Professora do Magistério Superior**, em 24/08/2022, às 15:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1505467** e o código CRC **897C2F4C**.

Referência: Processo nº 23072.231453/2022-51 SEI nº 1505467

Todas minhas vitórias e conquistas são dedicadas aos meus pais. Aprendi com eles o que nenhuma universidade do mundo poderia me ensinar. Muito obrigado.

Agradecimentos

Primeiramente ao professor Maurício. Fora seis anos trabalhando juntos. O que preciso dizer é que tive a sorte e a honra de trabalhar com um profissional sério, honrado, dedicado e muito especial. Obrigado pela parceria e, principalmente, pela paciência.

À professora Heliana Mello pelo carinho com que me recebeu e por todas as dicas que me deu durante meu doutorado. Você foi uma grata surpresa na minha caminhada e espero que possamos nos encontrar noutras oportunidades.

À professora Zilma Reis pela solicitude com que me atendeu e por lutar comigo pelos dados do Hospital das Clínicas. Agradeço imensamente por todos os esforços. A batalha foi perdida, mas guardo comigo a gratidão pela ajuda cedida.

À colega de doutorado Amanda Damasceno por se dispor a me ajudar a conseguir os dados do Hospital Felício Rocho. Você foi um exemplo de altruísmo raro de ser ver no nosso cotidiano tão apressado. A batalha também foi perdida, mas espero poder retribuir sua generosidade um dia.

À CAPES por patrocinar este trabalho através de uma bolsa de estudos de demanda social. Sem esse apoio esse projeto não teria saído do papel.

Ao grande amor da minha vida Natanna Carvalho. Você tornou essa caminhada menos árdua. Não há palavras que descrevam o quão importante você é na minha vida. Na alegria e na tristeza, na saúde e na doença em todos os dias de nossas vidas. Te amo!

"I was the dreamweaver
But now I'm reborn
I was the walrus
But now I'm John
And so, dear friends
You'll just have to carry on
The dream is over"
(John Lennon)

Resumo

Introdução: A área de proteção à privacidade vem ganhando importância nos últimos anos. Iniciativas como a Lei Geral de Proteção de Dados, ou LGPD, surgem numa tentativa de proteger a privacidade individual e evitar mau uso de dados pessoais. A proteção se torna essencial no meio digital, em que vazamentos são impossíveis de serem revertidos. Na área de saúde, a adoção de prontuários eletrônicos de pacientes tornou possível a digitalização de dados sensíveis de milhões de pessoas. Uma forma de proteção é a desidentificação de dados sensíveis que garantem a privacidade individual. Além da proteção, na área da saúde a desidentificação permite que os documentos possam ser compartilhados para uso secundário da informação, permitindo que conhecimento seja adquirido por meio de pesquisa e análise de dados.

Problema: Documentos clínicos possuem uma série de campos de textos livres que podem conter informação sensível que precisa ser protegida. O processo de desidentificação manual de documentos clínicos é custoso devido à quantidade de dados produzidos diariamente nas unidades de saúde. Uma alternativa para esse problema é a desidentificação automática usando técnicas de processamento de linguagem natural e aprendizado de máquina. Esses algoritmos devem ser treinados com uma base de dados no idioma ao qual ele será executado. Uma pesquisa preliminar indicou que não existem trabalhos de desidentificação, para a língua portuguesa, publicados na literatura e terem seus dados disponibilizados para a comunidade científica. Logo percebeu-se a necessidade de pavimentar essa área de estudo, desenvolvendo técnicas de tratamento da informação de saúde para atendimento à necessidade de privacidade para a língua portuguesa do Brasil.

Metodologia: Para atacar o problema, o presente trabalho elaborou uma metodologia de desidentificação automática de documentos clínicos usando algoritmos de processamento de linguagem natural e aprendizado de máquina. Para isso, fez-se uma parceria com o Hospital das Clínicas da UFMG a fim de obter documentos clínicos. Esses documentos foram preprocessados e usados para o desenvolvimento de um algoritmo de desidentificação adaptado para textos na língua portuguesa.

Resultados: O algoritmo de desidentificação desenvolvido obteve um F-Score (micro) de 97,94% e um F-Score (macro) de 39,83% dos dados. Apenas 37,09% dos dados foram corretamente desidentificados, portanto não permitem uma generalização do problema. O trabalho, no entanto, apresenta, como contribuição, a metodologia para desidentificação de documentos clínicos, com aplicação em quaisquer áreas onde há a necessidade de proteção à privacidade. Os códigos desenvolvidos e o modelo de aprendizado gerado durante o desenvolvimento do trabalho foram compartilhados publicamente e podem ser reusados por qualquer pessoa.

Palavras-chave: documentos clínicos, desidentificação, privacidade de dados.

Abstract

Introduction: the privacy protection is becoming relevant nowadays. Initiatives, such as General Data Privacy Regulation, or GDPR, emerged worldwide in an attempt to protect individual privacy and avoid bad use of personal data. The data protection becomes essential within digital context, where data leaks cannot be reverted. In the health area, the adoption of electronic health records led to the digitalization of millions of people sensitive data. A way to protect the data is the de-identification which assures the individual privacy. Besides the data protection, the de-identification also allows the clinical documents to be shared, allowing knowledge acquisition through research and data analysis.

Problem: clinical documents have countless text fields that may have sensitive data to be protected. The manual de-identification in the health area is costly due to the amount of data created every day across several health facilities. An alternative to handle this situation is the automatic de-identification using techniques of machine learning and natural language processing. However, those algorithms should be trained using the local language where it will be validated. A preliminary research do not identified studies of de-identification for Brazilian Portuguese with available data. Therefore, was identified the opportunity to improve the field of study in de-identification for Brazilian Portuguese, developing research to privacy protection in clinical documents.

Methodology: to handle the problem, the present thesis built a methodology to automatic de-identification data from clinical documents using natural language processing and machine learning algorithms. To achieve this, a partnership was made with the Hospital das Clínicas de Minas Gerais to obtain the clinical documents. These documents were preprocessed and used to the development of the de-identification algorithm adapted to Brazilian Portuguese language.

Results: the deidentification algorithm obtained an F-Score (macro) of 97,94% and an F-Score (micro) of 39,83%. Only 37,09% of the data was correctly deidentified. Thus, the results were insufficient for a generalization. This thesis, however, presents as it contribution the methodology proposed to deidentify clinical documents. This methodology can be applied to any field, beyond the health, which has its needs on the privacy protection. Also, the source code developed during the methodology and the trained learning model is publicly available and can be used by everyone.

Keywords: clinical documents, de-identification, privacy protection.

Lista de ilustrações

Lista de tabelas

Lista de códigos

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
ACID	Atomicidade, Consistência, Isolamento e Durabilidade
AIDS	Acquired Immunodeficiency Syndrome
AIH	Autorização de Internação Hospitalar
CADSUS	Sistema de Cadastramento de Usuários do Sistema Único de Saúde
CBO	Classificação Brasileira de Ocupações
CCHMC	Cincinnati Children's Hospital Medical Center
CEP	Comitê de Ética e Pesquisa
CEP	Código de Endereçamento Postal
CID	Classificação Internacional de Doenças
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CNS	Cartão Nacional de Saúde
CONEP	Comitê Nacional de Ética em Pesquisa
COEP	Comitê de Ética e Pesquisa
CPF	Cadastro de Pessoas Físicas
CPRD	Clinical Practice Research Datalink
CRF	Conditional Random Field
CTI	Centro de Tratamento e Terapia Intensiva
DER	Diagrama de Entidade e Relacionamento
DBMI	Department of Biomedical Informatics
EBSERH	Empresa Brasileira de Serviços Hospitalares
EMC	Erasmus Medical Center
EPR	Electronic Patient Records

GDPR	General Data Privacy Regulation
HC	Hospital das Clínicas
HIPAA	Health Insurance Portability and Accountability Act
HMM	Hidden Markov Model
HSCIC	the Health and Social Care Information Centre
i2b2	Informatics for Integrating Biology & the Bedside
IBGE	Instituto Brasileiro de Geografia e Estatística
IMEI	International Mobile Equipment Identity
INE	Identificação Nacional de Equipe
IOB	Inside-Outside-Beginning
IP	Internet Protocol
ISO	International Organization for Standardization
JCI	Joint Comission International
LGPD	Lei Geral de Proteção de Dados
LSTM	Long Short-Term Memory
MAMA	Model-Annotate Model-Annotate
MATTER	Model Annotate Train Test Evaluate Revise
MaxEnt	Máxima Entropia
MedLEE	Medical Language Extraction and Encoding System
MEMM	Maximum-Entropy Markov Model
MIMIC	Medical Information Mart for Intensive Care
MIT	Massachusetts Institute of Technology
NBR	Norma Brasileira
NEHTA	National e-Health Transition Authority
NIHR	National Institute for Health Research
PEP	Prontuário Eletrônico de Paciente

PHI	Protected Health Information
PLN	Processamento de Linguagem Natural
PoS	Part of Speech
RAM	Reação Adversa a Medicamento
REM	Reconhecimento de Entidade Mencionada
RG	Registro Geral
RNN	Recurrent Neural Network
SGBD	Sistema Gerenciador de Banco de Dados
SUS	Sistema Único de Saúde
SVM	Support Vector Machine
TCLE	Termo de Consentimento Livre e Esclarecido
THIN	The Health Improvement Network
THYME	Temporal History of Your Medical Events
TISS	Troca de Informação em Saúde Suplementar
UBS	Unidade Básica de Saúde
UFMG	Universidade Federal de Minas Gerais
UMLS	Unified Medical Language System
UPA	Unidade de Pronto Atendimento
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Sumário

1 Introdução

O estudo da medicina e o avanço nessa área de pesquisa está, em parte, condicionada ao acesso à informação. Pode-se citar como exemplos de acesso à informação: documentos clínicos emitidos em unidades de saúde; produção científica em forma de artigos, ensaios e outras formas de divulgação; estudos regionais de saúde realizados por órgãos governamentais e empresas, entre outros. Esses documentos oferecem aos pesquisadores fontes de informações que podem ser exploradas de diferentes maneiras para a aquisição de conhecimento. Dá-se o nome de uso secundário da informação quando documentos criados para fins de saúde, como prontuários de pacientes, por exemplo, são empregados para outros fins, tais como mineração e análise de dados.

A tecnologia vem-se mostrando uma aliada da ciência propiciando um maior acesso à informação digital. Na medicina, não é diferente. O movimento de informatização na área de saúde não é novo, como pode ser observado na pesquisa de Lopes e Araújo (2002). A pesquisa realizada por Lourenção e Junior (2016) identificou um crescente aumento no interesse das instituições de saúde brasileiras na adoção de documentos eletrônicos a partir de 2007. Embora o processo de adoção de sistemas de informação na área da saúde ainda esteja em andamento, como identificado por Santos, Pereira e Silveira (2017), percebe-se que este é um processo natural dos dias atuais. A busca por processos ágeis, segurança da informação, automatização de rotinas, escalabilidade de processos, entre outros, leva as organizações de saúde a informatizarem seus departamentos.

Uma das iniciativas que impulsionou a pesquisa em saúde foi a criação do prontuário eletrônico do paciente. Esse documento facilitou e modernizou o tratamento de saúde. A transformação de prontuários físicos em digitais trouxe uma série de vantagens para a sociedade. Pode-se listar algumas delas: redução do tempo que os profissionais de saúde levam para acessar os dados de pacientes, disponibilização simultânea e instantânea de prontuários para diferentes consultas, eliminação de espaço físico necessário para armazenar documentos, aumento de segurança, entre outros.

Além disso, os documentos clínicos possuem valor como informação. Os prontuários podem fornecer dados relevantes sobre uma série de situações envolvendo a saúde pública e os meios de tratamento praticados por profissionais de saúde. É possível inferir uma série de informações através de análise de grandes conjuntos de dados. Pode-se saber, por exemplo, sobre a ocorrência duma epidemia, quais os tratamentos mais eficazes para uma determinada doença, o perfil de pessoas que são mais afetadas por determinadas condições clínicas, etc. Esses documentos podem ser analisados sob o campo da mineração de dados em seu uso secundário. Essa área de estudo usa princípios da computação, linguística e

estatística para desenvolver técnicas de análise e extração de informação. A mineração de dados clínicos pode ser usada de diversas formas como, por exemplo:

- auxiliar na interoperabilidade semântica entre diferentes campos e instituições de saúde para identificar terminologias diferentes, para um mesmo assunto, em prontuários de pacientes (KOKKINAKIS, 2006);
- proporcionar a criação de um repositório de testes para sistemas de suporte à decisão clínica (SHIN; MARKEY, 2006; VELUPILLAI; KVIST, 2012);
- possibilitar a análise de dados clínicos para correlacionar diagnósticos e identificar doenças raras (MACLEOD et al., 2016; JENSEN et al., 2017);
- fomentar pesquisa em linguística computacional para identificação de acrônimos e abreviações (KVIST; VELUPILLAI, 2014), além de categorias de asserções em prontuários de pacientes, como especulações (VINCZE et al., 2008), negações (DALIANIS; SKEPPSTEDT, 2010) e níveis de certeza em afirmações (VELUPILLAI, 2011);
- gerar aplicações em Processamento de Linguagem Natural (PLN) para desambiguação de termos de saúde (MEYSTRE et al., 2008), diagnóstico de doenças (ZENG et al., 2006) e identificação de erros de escrita (NIZAMUDDIN; DALIANIS, 2014);
- realizar a predição de quadros clínicos baseados no histórico e sintomas apresentados pelo paciente (CHEN; ASCH, 2017).

Os prontuários de pacientes carregam um conteúdo informacional importante. Esses documentos possuem características peculiares e, se utilizados para fins de pesquisa, podem contribuir significativamente para o avanço da medicina. Pakhomov, Pedersen e Chute (2005) citam que os prontuários de pacientes possuem cerca de 30% de ruídos. Símbolos, abreviações, acrônimos, erros de digitação e uso incorreto da gramática são alguns exemplos de ruídos contidos em documentos de saúde. Embora as instituições de saúde possuam siglários para identificar abreviações, esses termos não são universais. A identificação de abreviações e acrônimos pode reduzir o problema de ambiguidades, facilitando a interpretação e recuperação de textos por pessoas e máquinas. Os prontuários de pacientes também possuem o problema de interoperabilidade semântica. Setores distintos de saúde empregam termos distintos para se referirem a mesma coisa. O desafio de comunicação entre diferentes vocabulários de saúde já é uma preocupação do Ministério da Saúde (2011) que, por meio da Política Nacional de Informação e Informática em Saúde, busca estruturar os dados de saúde do país.

Um dos documentos contidos no prontuário eletrônico do paciente (PEP) é o sumário de alta. Esse documento traz resumidamente o histórico do paciente durante sua

permanência em unidades de saúde. O sumário de alta também permite a assistência continuada de saúde ao paciente após o término do tratamento. Ele é um documento formal e estruturado (DEAN et al., 2016) e, por isso, tem a preferência dos pesquisadores como principal fonte de uso secundário da informação.

O acesso aos sumários de alta, no entanto, não é simples. A resolução n.º 1.638/2002 do Conselho Federal de Medicina (2002) define o prontuário do paciente como um documento único, sigiloso e científico cuja responsabilidade é do profissional de saúde que realizou o atendimento, da unidade de saúde onde ele atua e dos demais profissionais envolvidos no atendimento e em cargos de chefia. A violação desse prontuário fere a Constituição Federal que dá aos cidadãos o direito de intimidade (BRASIL, 1988). Essa intimidade está ligada também ao direito de privacidade das pessoas. Esse direito, por sua vez, é assegurado pela Lei Geral de Proteção de Dados, que garante titularidade da informação ao paciente, que seus dados não sejam usados em prejuízo próprio, dentre outras garantias (BRASIL, 2018).

Por esse motivo, o uso de dados pessoais de pacientes enfrenta restrições. O compartilhamento desses dados exige trâmites e autorizações em comitês de ética e pesquisa, que buscam proteger os dados de pacientes e seus direitos à privacidade como previsto pelo Conselho Nacional da Saúde (2012). Uma alternativa para esse impasse é a disponibilização pública de documentos clínicos. Para isso é preciso proteger a privacidade de indivíduos aos quais os sumários dizem respeito. Uma solução encontrada na legislação vigente e amparada por processos semelhantes ao redor do mundo é a **desidentificação** de documentos.

A desidentificação é responsável por remover qualquer informação sensível que possa ser usada para identificar um paciente, um profissional de saúde ou uma prestadora de serviços (CHEVRIER et al., 2019). Esse processo é importante, pois as consequências sociais e legais são graves caso haja revelação de dados sensíveis que possam prejudicar, tanto o paciente quanto os profissionais de saúde responsáveis (BRASIL, 2018). A desidentificação deve, se possível, ser automatizada em função do grande volume de dados produzidos nas unidades de saúde.

Uma das possibilidades do processo de desidentificação para fins de pequisa é a construção de um *corpus*, um conjunto de textos selecionados e agrupados que represente um contexto linguístico (DELEGER et al., 2014). No caso da pesquisa na área de saúde, esse *corpus* deve representar a linguagem e as terminologias da área sem expor ao risco as pessoas e entidades citadas nesses textos. Um *corpus* de documentos clínicos pode contemplar diversas áreas de saúde e oferecer uma rica amostra de heterogeneidade para estudos de integração de sistemas. Uma vez disponibilizado esse *corpus*, ele pode ser **usado para diferentes pesquisas de mineração, extração e análise de dados.**

Existem diversas iniciativas em criação de *corpora* de documentos clínicos ao redor

do mundo como para o inglês (VINCZE et al., 2008), alemão (SPAT et al., 2008), japonês (ARAMAKI et al., 2009), sueco (DALIANIS et al., 2012), espanhol (PEREZ et al., 2017), dentre outros. No entanto, não existe uma iniciativa similar para a língua portuguesa. Nesse sentido, é importante ressaltar que a pesquisa em linguística e mineração de dados varia conforme o idioma. Portanto o fato de haver diferentes *corpora*, em vários idiomas, não anula a necessidade de que haja um *corpus* em português. Pelo contrário. Segundo Névéol et al. (2018), a mineração de dados e PLN na área da saúde na língua portuguesa representa apenas 0,3% de toda pesquisa realizada no mundo. Dessa maneira, faz-se necessário estimular a pesquisa de dados na área de saúde para a língua portuguesa.

Chega-se a **pergunta de pesquisa** que é: "Qual a possibilidade de disponibilizar dados de saúde publicamente num formato de corpus para uso secundário?". A **justificativa** desta pesquisa é para a criação de métodos automáticos de desidentificação de documentos clínicos na língua portuguesa. Esses métodos atendem à necessidade de privacidade ao proteger dados sensíveis em documentos clínicos. Entende-se por dado sensível qualquer tipo de informação que pode revelar a identidade de uma pessoa e/ou expô-la de alguma perante a sociedade. O risco de revelação de dados de pacientes é reduzido. Além disso, a desidentificação de dados propiciará a construção de *corpora* de documentos clínicos para uso secundário da informação. Segundo Dalianis (2018), documentos clínicos são difíceis de serem obtidos devido à sensibilidade da informação contida. Esses *corpora* podem trazer reprodutibilidade e estudos comparativos sobre um mesmo objeto de estudo. Também acredita-se que um *corpus* aumentará a quantidade de pesquisa feita em mineração, análise de dados e PLN na área de saúde, pois o acesso à informação seria facilitado por remover os entraves relacionados a segurança e privacidade de pacientes e outras entidades. Além disso, haverá a redução no número de pedidos para acesso aos documentos de saúde, reduzindo o esforço de análise de comitês e agilizando a realização desse tipo de pesquisa. Consequentemente mais cientistas terão acessos a uma base para pesquisa enquanto os comitês de ética terão uma carga menor de avaliações.

Este trabalho tem como **objeto de estudo** os dados de documentos clínicos do Hospital das Clínicas da UFMG[1] que foram devidamente requisitados ao Comitê de Ética e Pesquisa (COEP) (Ver Apêndice A). Para trabalhar sobre o objeto de estudo definido no presente projeto, foram demarcados o objetivo geral e os específicos. O **objetivo geral**:

- é construir um método automático de desidentificação de documentos clínicos na língua portuguesa.

Para alcançar o objetivo geral, os seguintes **objetivos específicos** foram definidos:

[1] Disponível em https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sudeste/hc-ufmg. Acessado em março de 2022.

- levantar as técnicas de desidentificação mais utilizadas na literatura;
- levantar as questões de privacidade que garantem a segurança de qualquer pessoa ou entidade envolvida em documentos de saúde;
- levantar as características de *corpora* de documentos clínicos disponibilizados para a comunidade científica;
- criar algoritmos que automatizem o processo de desidentificação usando as técnicas levantadas na literatura;
- criar um *corpus* de documentos clínicos, anotado, que será usado como treinamento de algoritmos de desidentificação.

A realização desses objetivos permite que métodos de desidentificação sejam aplicados em documentos clínicos. Para isso, o trabalho foi dividido da seguinte maneira: o Capítulo 2 destrincha as características dos *corpora* disponibilizados na internet. Isso, permitiu a elaboração de estratégias para montar um *corpus* para esta pesquisa. O Capítulo 3 analisa o sumário de alta, o principal candidato a documento a ser usado na pesquisa. Este capítulo mostra a importância do sumário de alta para a saúde e como documento de informação. O Capítulo 4 faz uma revisão de literatura sobre a desidentificação de documentos clínicos. Os tipos de dados sensíveis, os riscos do processo e as técnicas por trás da desidentificação e os principais trabalhos publicados na literatura deste século são analisados.

O Capítulo 5 descreve a metodologia elaborada para a execução deste projeto. O Capítulo 6 apresenta a execução da metodologia e os resultados obtidos. Por fim, o Capítulo 7 traz as contribuições do trabalho, ideias de trabalho futuro para melhorar os resultados e comentários finais sobre as adversidades encontradas durante o desenvolvimento.

2 *Corpora* de prontuários eletrônicos de pacientes ao redor do mundo

É difícil encontrar corpora de documentos clínicos disponíveis publicamente por conterem informações sensíveis sobre pacientes, profissionais de saúde e prestadores de serviços.

Para que um *corpus* seja disponibilizado publicamente, é necessário remover todas as informações sensíveis. Isso exige um processo de pesquisa, autorização de comitês de ética e uma rotina de desidentificação para que, só assim, sejam disponibilizados esses documentos. Um estudo realizado por Dalianis (2018) listou os principais *corpora* disponíveis no mundo. O presente capítulo aprofundou os detalhes dessa lista em busca de *corpora* de acesso público além de atualizá-la com outros *corpora* encontrados durante a pesquisa. A seguir, são listados os *corpora* clínicos que estão, de alguma maneira, acessíveis publicamente.

2.1 *Corpora* em inglês

Durante esta pesquisa, foram encontrados mais *corpora* em inglês do que nos demais idiomas. Uma hipótese para esse achado é o fato de o inglês ser a língua franca atual. Há também a tradição de institutos de linguística em países nativos da língua inglesa a citar a *Brown University*, *Lancaster University* e *Oxford University* como exemplos. Institutos esses apoiados com fundos nacionais para o desenvolvimento de pesquisas em linguística de *corpus*. As próximas subseções discutem esses *corpora* e suas características.

2.1.1 *Informatics for Integrating Biology & the Bedside*

Também conhecido como i2b2[1], trata-se de uma organização sem fins lucrativos com a missão de compartilhar, integrar, padronizar e analisar dados de saúde heterogêneos por uma comunidade de dados e códigos abertos.

O i2b2 foi pioneiro na criação de *corpora* de prontuários e na busca de soluções de PLN em dados clínicos. São ao todo dados de 4,6 milhões de pacientes e 1,2 bilhão de observações clínicas (MURPHY et al., 2010). Essas observações são divididas em documentos como prontuários de pacientes, documentos financeiros e contábeis, exames laboratoriais e sumários de alta. Os dados foram obtidos da Partners Healthcare[2], uma rede de hospitais sem fins lucrativos dos Estados Unidos.

[1] Disponível em https://www.i2b2.org/. Acessado em março de 2022.
[2] Disponível em https://www.partners.org/. Acessado em março de 2022.

A iniciativa se deu por desafios feitos à comunidade para a estimular a resolução de problemas relacionados ao PLN. Segue a lista de todos os desafios patrocinados pelo i2b2 até a data em que este capítulo estava sendo escrito:

- 2006: desidentificação de sumários de alta não estruturados (UZUNER; LUO; SZOLOVITS, 2007) e identificação de pacientes fumantes (UZUNER et al., 2008);
- 2008: identificação de pacientes obesos, suas doenças e relações com a obesidade;
- 2009: extração de medicamentos usados no tratamento de pacientes;
- 2010 (1): extração de diagnósticos de saúde, testes e tratamentos realizados;
- 2010 (2): classificação de asserções feitas por profissionais de saúde em avaliações de quadros clínicos;
- 2010 (3): reconhecimento de relações entre diagnósticos, testes e tratamentos;
- 2011: identificação de correferência[3] entre dados anotados;
- 2012: identificação de relações temporais;
- 2014: desidentificação de dados sensíveis e reconhecimento longitudinal de fatores para doenças no coração;
- 2016: extração de sintomas;
- 2018 (1): determinação, na triagem, se pacientes possuem critérios necessários para serem encaminhados ao atendimento;
- 2018 (2): identificação de nomes, dosagens, duração de uso e outros atributos relacionados a medicamentos;
- 2018 (3): reconhecimento de medicamentos e relações que causam reações adversas;
- 2019 (1): identificação de similaridades entre sentenças escritas;
- 2019 (2): extração de familiares, relacionamentos e assimilação de doenças hereditárias relacionadas aos pacientes;
- 2019 (3): normalização de conceitos de saúde;
- 2019 (4): estudos de novos usos secundários da informação;
- 2022 (1): reconhecimento de medicamentos mencionados;
- 2022 (2): extração de histórico;

[3] Correferência é a relação entre dois ou mais termos que se referem a uma mesma entidade.

- 2022 (3): predição de raciocínio entre problemas identificados por profissionais de saúde e o tratamento proposto.

Cada um desses desafios é realizado com um *corpus* previamente tratado para execução. Os dados para isso são criados em bases menores, montadas especificamente para a resolução de tais problemas. Os dados estão disponíveis ao público através de cadastro e solicitação de uso. A partir de 2014, os dados do i2b2 foram movidos para o departamento de informática biomédica (DBMI) da Universidade de Harvard[4] sob o nome de n2c2.

2.1.2 *Cincinnati Children's Hospital Medical Center Data*

O *Cincinnati Children's Hospital Medical Center Data*, ou CCHMC, possui sob a tutela de Pestian Lab Research[5] dois *corpora* clínicos. O primeiro contém 1954 documentos de radiologia renal e pulmonar. Todos os documentos foram desidentificados e anotados conforme a CID-9 (Classificação Internacional de Doenças) (PESTIAN et al., 2007).

O segundo *corpus* é composto de uma coleção de cartas de suicídio desidentificadas e anotadas. As anotações realizadas nesse *corpus* são para tarefas de análise de sentimento. Ao todo foram anotados 13 sentimentos diferentes como abuso, raiva, acusação, medo, culpa, desesperança, tristeza, perdão, felicidade/tranquilidade, confiança, amor, orgulho e gratidão. Também foram realizadas anotações sobre instruções deixadas pelos suicidas. O *corpus* contém cartas de 1319 pessoas coletadas entre os anos de 1950 até 2011 (PESTIAN et al., 2012).

Ambos os *corpora* estão disponíveis gratuitamente para fins acadêmicos e são liberados mediante requisição na página do Pestian Lab Research[6].

2.1.3 MIMIC *Critical Care Database*

MIMIC é o acrônimo para *Medical Information Mart for Intensive Care* e está na sua quarta versão denominada MIMIC-IV. Trata-se de um *corpus* com 448 972 documentos clínicos de diferentes hospitais. Dentre os tipos de documentos encontrados no MIMIC-IV estão admissões de emergência de pacientes adultos internados em Centros de Terapia Intensiva (CTIs), sumários de alta, dados demográficos, exames de imagem, anotações médicas etc. Os dados foram extraídos de unidades de saúde que usam o MetaVision[7] como sistema de informação popular em hospitais da Europa, Austrália, Canadá e dos Estados Unidos. Os dados do *corpus* foram coletados entre os anos 2008 e 2019 (JOHNSON et al., 2021).

[4] Disponível em https://portal.dbmi.hms.harvard.edu/. Acessado em março de 2022.
[5] Disponível em https://www.cincinnatichildrens.org/research/divisions/b/bmi/labs/pestian. Acessado em março de 2022.
[6] Disponível em https://bmi.cchmc.org/help/project-requests. Acessado em março de 2022.
[7] Disponível em: https://www.imd-soft.com/. Acessado em março de 2022

O MIMIC-IV possui informação longitudinal de pacientes por mais de uma década. Segundo os autores, esse é o único *corpus* de acesso gratuito com essa característica. Os dados são mantidos e disponibilizados pelo *MIT Lab for Computational Physiology* do Instituto de Tecnologia de Massachusetts (MIT). O *corpus* é gratuito e de uso restrito, demandando que sejam aceites os termos de uso.

O *corpus* tem seu uso em diferentes aplicações, desde a área de ensino até a mineração e análise de dados. Os autores sugerem que os dados sejam usados para aprendizado de máquina, predição e análise semântica.

Para ter acesso ao MIMIC-IV, é preciso que o usuário complete o curso *Human Subjects Research, Data or Specimens Only Training*. Esse curso é mandatório, pois capacita o usuário a conduzir pesquisa envolvendo dados coletados de humanos. Uma vez autorizado é preciso cadastrar-se no site PhysioNet[8] e fazer uma requisição de acesso ao MIMIC-IV.

2.1.4 BioScope *corpus*

O BioScope *corpus* foi criado e é mantido pelo Instituto de Informática da Universidade de Szeged[9], na Hungria. O conteúdo do *corpus* é dividido em três partes: (1) resumos de documentos clínicos; (2) artigos completos de medicina; (3) textos livres de documentos de radiologia. Esse conteúdo foi anotado por uma equipe de linguistas para estudar níveis das asserções feitas por profissionais de saúde. O *corpus* possui mais de 20 000 sentenças e os autores concluíram que o uso de negação e incerteza para descrever quadros clínicos corresponde a mais de 10% das sentenças (VINCZE et al., 2008).

Os textos de artigos completos e resumos podem ser baixados separadamente sem necessidade de cadastro. Eles estão sob supervisão do grupo de PLN da Universidade de Szeged[10]. Já os textos de saúde estão sob supervisão do Pestian Lab sendo preciso fazer uma requisição para ter acesso aos dados (PESTIAN et al., 2007).

2.1.5 THYME

THYME é acrônimo para *Temporal History of Your Medical Events*, um *corpus* especializado em anotações temporais (IV et al., 2014). O THYME implementa relações usando o TimeML[11], uma linguagem de marcação especializada em etiquetas temporais. Segundo os autores, as relações temporais são importantes no campo da medicina, pois estão intrinsecamente ligadas a doenças, sinais, sintomas e tratamentos. Conhecer essas informações é, portanto, um aspecto chave para a generalização de comportamentos em estudos com quantidades massivas de dados.

[8] https://physionet.org/. Acessado em março de 2022.
[9] Disponível em http://www.inf.u-szeged.hu/en. Acessado em março de 2022.
[10] Disponível em https://rgai.inf.u-szeged.hu/node/105. Acessado em março de 2022.
[11] Disponível em http://timeml.org/site. Acessado em março de 2022.

As anotações no THYME foram feitas manualmente por dois estudantes de linguística e supervisionadas por um especialista no domínio de saúde. Foram anotados documentos clínicos de pacientes com câncer no cólon da clínica Mayo[12]. Os dados podem ser obtidos por requisição ao grupo de semântica computacional, da Universidade de Colorado[13], que supervisiona o THYME.

2.1.6 *Clinical Practice Research Datalink*

Também conhecido como CPRD[14], esse *corpus* é uma iniciativa ousada do governo britânico de estabelecer uma base completa com dados de pacientes do Reino Unido. Através de uma parceria com a Agência Regulatória de Medicina e Produtos Médicos e o Instituto Nacional de Pesquisas Médicas (NIHR) do Reino Unido, o CPRD coleta dados desidentificados de pacientes através de documentos fornecidos por clínicos gerais em todo o território britânico. Os dados são relacionados com outros documentos de saúde e fornecem uma visão representativa e longitudinal dos cuidados de saúde desses pacientes. Ao todo são mais de 60 milhões de pacientes abrangidos pelo *corpus* (HERRETT et al., 2015).

O CPRD é um *corpus*-monitor alimentado diariamente através de diferentes sistemas de saúde onde os pacientes têm seus dados cadastrados. Os dados foram divididos em *subcorpora*. O primeiro contém dados de pacientes atendidos por clínicos gerais. Esse *corpus* é dividido em outros 2 *subcorpora*. Essa última divisão foi motivada pela diferença de estruturas entre os documentos, pois são extraídos de sistemas de informação diferentes. Já o segundo *subcorpus* possui documentos de saúde conectados longitudinalmente.

Esses dados, no entanto, não são disponibilizados gratuitamente ao público. Além de ter que submeter um projeto rigoroso com detalhes da pesquisa e do uso de dados, os pesquisadores precisarão desembolsar quantias que variam de 15 000 a 75 000 libras dependendo de qual *subcorpus* for utilizado.

2.1.7 *The Health Improvement Network*

Também conhecido como THIN, presta um serviço similar ao oferecido pelo CPRD. O THIN é um *corpus*-monitor que coleta dados mensalmente dos habitantes do Reino Unido. Ao todo, 6% da população é representada longitudinalmente de 1994 até os dias atuais. Os dados coletados são referentes a tratamentos de saúde realizados em território britânico (BLAK et al., 2011).

Os dados coletados são anonimizados conforme o regulamento nacional e fornecidos para uso acadêmico e comercial. Para ter acesso aos dados, é necessário preencher um

[12] Disponível em https://www.mayoclinic.org/portugues. Acessado em março de 2022.
[13] Disponível e http://clear.colorado.edu/compsem/index.php. Acessado em março de 2022.
[14] Disponível em: https://www.cprd.com/. Acessado em março de 2022.

formulário de requisição[15]. A empresa que disponibiliza o THIN não informa se é necessário pagar para usar o *corpus* com fins acadêmicos.

2.2 *Corpora* em outros idiomas além do inglês

Além do inglês, há *corpora* de diferentes idiomas espalhados pelo mundo. Este estudo dá uma ideia da importância da criação de *corpora* de documentos clínicos em idiomas diferentes, pois cada um apresenta suas particularidades linguísticas que não são transmitidas de um idioma para outro. As próximas subseções detalham esses *corpora*.

2.2.1 *Stockholm EPR corpus*

O *Stockholm Electronic Patient Records* (EPR) *corpus* foi criado e é mantido pelo departamento de computação da Universidade de Estocolmo[16], na Suécia. Esse *corpus* faz parte de um conjunto maior de dados denominado Health Bank, formado por prontuários eletrônicos de pacientes, ou PEPs. O Stockholm EPR *corpus*, especificamente, possui dados de mais de 512 unidades de tratamento que compõem o hospital universitário de Karolinska. Todos os registros são desidentificados, sendo compostos de dados estruturados e desestruturados.

Os documentos com dados estruturados contêm um número serial para cada paciente. Esse número substitui o nome, idade, gênero, identificações de doenças baseadas na CID-10, dados laboratoriais, coletas de sangue, nomes e códigos de medicamentos, datas de admissão e de alta. Já os dados desestruturados contêm textos livres de diferentes documentos. São dados de mais de 2 milhões de pacientes coletados entre 2006 e 2014 com mais de 500 milhões de *tokens* (DALIANIS; HASSEL; VELUPILLAI, 2009). Um *token* é a menor parte de um *corpus* e, geralmente, é dividido em duas categorias: palavras e não palavras. *corpus*

Além disso, o Stockholm EPR *corpus* pode ser dividido em diferentes *subcorpora* com as seguintes características:

- desidentificado para atendimento à privacidade;
- anonimizado para aplicação de desidentificação;
- anotado com incertezas e negações;
- diagnosticado em níveis de asserção através de categorias que vão de totalmente positivas até totalmente negativas;

[15] Disponível em: https://www.the-health-improvement-network.com/access-to-thin-data. Acessado em março de 2022

[16] Disponível em https://www.su.se/english/. Acessado em março de 2022.

- anotado em categorias como doenças, diagnósticos, suspeitas, partes do corpo, medicamentos etc.;
- identificado se documentos contêm informações de tratamento de saúde ou não;
- relacionado em níveis temporais, de asserção e outros fatores relacionados a Reações Adversas a Medicamentos (RAMs).

Para ter acesso aos dados é preciso fazer parte do grupo de Mineração de Dados Clínicos ou fazer uma requisição para o coordenador do grupo, Hercules Dalianis[17].

2.2.2 *Erasmus Medical Center Dutch Clinical corpus*

O *EMC Dutch Clinical corpus* é um *corpus* em holandês anotado com relações temporais, propriedades de negação e análise das experiências dos usuários. O *corpus* é, ao todo, composto por quatro categorias de documentos (AFZAL et al., 2014):

1. Prontuários longitudinais de pacientes que se consultaram com clínicos gerais. Contêm dados de mais de 1,5 milhão de pacientes holandeses.
2. Receituários escritos por médicos especialistas. Esses receituários emitidos são dos mesmos pacientes que se consultaram com clínicos gerais e depois foram encaminhados para um especialista. São documentos escaneados e podem possuir erros gramaticais.
3. Laudos de radiologia com descrições e conclusões de exames de imagem realizados no hospital. São relatos de diferentes médicos e radiologistas, geralmente gerados por reconhecimento automático. Por esse motivo, possuem gramática e estrutura condizentes com o vocabulário usado no departamento de radiologia.
4. Sumários de alta contendo dados de todos os pacientes já atendidos pelo hospital. Documentos bem redigidos devido à sua natureza de continuidade de cuidados de saúde, eles são usados em comparativos com sumários de admissão do paciente para identificar os resultados e os problemas que persistem após alta.

O *corpus* foi anotado por duas pessoas diferentes. Essas anotações foram integradas por um terceiro especialista no domínio. Para ter acesso ao EMC Dutch *corpus*, é preciso fazer uma requisição ao seu criador, Zubair Azfal[18].

[17] E-mail para contato: hercules@dsv.su.se
[18] E-mail para contato: m.afzal@erasmusmc.nl

2.2.3 *Turku Clinical TreeBank and PropBank*

Turku Clinical Treebank and Propbank é um *corpus* produzido pela Universidade de Turku, na Finlândia. Esse *corpus* possui documentações de enfermagem de oito pacientes de um hospital finlandês não especificado (HAVERINEN et al., 2009). São ao todo 2800 sentenças e mais de 17 000 *tokens*. Embora as sentenças sejam curtas e ocasionalmente consistam em fragmentos repetidos, o número de sentenças únicas nesse *corpus* é de cerca de 2000 (HAVERINEN et al., 2010).

Todos os dados sensíveis foram retirados dos documentos e devidamente anonimizados. Nomes de pacientes e informações relacionadas a eles foram trocados. A anonimização foi realizada manualmente por dois pesquisadores diferentes.

Turku Clinical TreeBank and PropBank foi disponibilizado sob a licença da Creative Commons Attribution-ShareAlike. Isso quer dizer que qualquer um é livre para compartilhar e adaptar o *corpus*, desde que faça referência ao conteúdo original. O *corpus* pode ser adquirido diretamente do site Turku BioNLP Group[19] e não precisa de permissões.

2.2.4 *Synthethic Clinical Text corpus*

O *Synthethic Clinical Text corpus* é um pequeno *corpus* composto de documentos nefrológicos escritos em alemão. São ao todo 55 documentos de notas clínicas e sumários de alta que somam mais de 13 000 *tokens* (KARA et al., 2018). O *corpus* foi anotado com classes gramaticais, também chamadas de PoS (do inglês *part of speech*), e anotações de dependência.

A anotação sintática dos documentos foi realizada por estudantes de pós-graduação em linguística. Já a anotação de PoS foi feita usando um etiquetador para o alemão chamado STTS-tagset (SCHILLER et al., 1999; SEIFFE, 2018).

Os dados foram devidamente desidentificados e estão disponíveis gratuitamente na internet[20].

2.2.5 *Chinese Electronic Medical corpus*

O *Chinese Electronic Medical corpus* foi desenvolvido por Gao et al. (2019) para aplicação de técnicas de reconhecimento de entidade mencionada (REM). O *corpus* consiste em 255 prontuários de admissão colhidos de um hospital na província de Hunan.

Os dados foram manualmente anotados para reconhecimento de entidades como medicamentos, tratamentos, doenças, partes do corpo, entre outras. As anotações foram feitas manualmente por três anotadores independentes e depois foram submetidas a uma

[19] Disponível em: http://bionlp.utu.fi/clinical*corpus*.html. Acessado em março de 2022.
[20] Disponível em http://macss.dfki.de/index.html. Acessado em março de 2022.

rede neural para reconhecimento automático de entidades. O *corpus* e o código desenvolvido no artigo podem ser encontrados gratuitamente na internet[21].

2.2.6 MEDDOCAN

O MEDDOCAN é um *corpus* em espanhol constituído de mil documentos selecionados manualmente. Esses documentos foram preenchidos com frases com alta densidade de dados sensíveis retirados de sumários de alta e registros genéticos de pacientes. São ao todo 33 000 frases e 495 000 *tokens* (MARIMON et al., 2019).

O *corpus* é dividido em *subcorpora* de treinamento, desenvolvimento e teste, o primeiro composto de 500 documentos e os demais com 250 cada. O *corpus* foi primariamente desenvolvido para a execução de duas tarefas: (1) classificação e REM; (2) detecção de dados sensíveis. O MEDDOCAN pode ser encontrado gratuitamente na internet[22].

2.3 *Corpora* em português

Dalianis (2018) não citou em seu trabalho, *corpora* em português. Isso nos leva a duas hipóteses: (1) o autor não tem a língua portuguesa como idioma conhecido, impossibilitando-o a realizar buscas nesse sentido, ou; (2) não existem *corpora* de pacientes em português.

Para tentar responder essa hipótese fez-se uma pesquisa nas bases de dados Pubmed, Scopus, Web of Science e na Biblioteca Digital Brasileira de Teses e Dissertações por ocorrências de trabalhos que usaram *corpora* de pacientes. A pesquisa retornou seis resultados, sendo que apenas quatro usam *corpora* de pacientes.

Bacic (2007) utilizou um *corpus* de 2800 laudos médicos de Raio-X do Instituto do Coração[23] como base de estudo. Os laudos foram divididos em dois *subcorpora*, um para testes e o outro para treinamento. A aplicação desenvolvida pela autora é a extração automática de termos de saúde baseados no vocabulário controlado UMLS (*Unified Medical Language System*). Esse trabalho sugeriu a possibilidade da disponibilização da base usada, porém nada foi encontrado sugerindo que a pesquisa foi descontinuada.

Pacheco (2009) utilizou como *corpus* uma base de dados do Hospital das Clínicas de Porto Alegre[24] obtida por meio do CONEP, Comitê Nacional de Ética em Pesquisa. O *corpus* foi formado por mais de 8000 documentos sendo todos sumários de alta de

[21] Disponívem em https://github.com/taotao033/Chinese_electronic_medical_record_NER. Acessado em março de 2022.
[22] Disponível em https://temu.bsc.es/meddocan/index.php/datasets/. Acessado em março de 2022.
[23] Disponívem em: https://www.incor.usp.br/sites/incor2013/. Acessado em março de 2022.
[24] Disponível em: https://www.hcpa.edu.br/. Acessado em março de 2022

cardiologia. O trabalho usa o *corpus* para aplicar técnicas de PLN na tentativa de converter códigos em narrativas da área de saúde.

Oliveira et al. (2013) utilizaram um *corpus* de 5617 sumários de alta também do Hospital das Clínicas de Porto Alegre. O trabalho usou esses dados para identificar processos de tratamento continuado nesses sumários. Para isso foram aplicadas técnicas de PLN baseadas em regras para identificar trechos de texto que sugerem que o paciente recebe tratamento continuado.

Oliveira et al. (2019) criaram e treinaram um *corpus* para identificação de doenças urinárias. Esse *corpus*, chamado GROUP-ALL, contém 745 000 documentos de três tipos: sumários de alta, documentos de enfermagem e relatórios ambulatoriais. Esse *corpus* foi classificado pelos autores como sendo de baixa granularidade com dados genéricos não especificamente relacionados a pacientes com problemas urinários. O objetivo da pesquisa é avaliar se *corpora* de granularidades baixas são úteis em tarefas de PLN na medicina.

Dentre os casos citados, nenhum *corpus* foi encontrado disponível.

2.4 Uma análise dos *corpora* disponíveis

A presente seção consolida os *corpora* encontrados nas seções anteriores para encontrar pontos de interseção entre os diferentes estudos ao redor do mundo. Para isso, foi elaborado a Tabela 1 com as características principais dos *corpora* analisados. Os resultados dessa análise são discutidos nessa seção.

Tabela 1 – Características dos *corpora* de pacientes

Nome	**Idioma**	**Quantidade de documentos**	**Tipos de documentos**	**Aplicações**	**De acesso gratuito?**	**É anotado?**	**Está desidentificado?**
i2b2	Inglês	1,2 bilhão	Prontuários de pacientes internados, documentos financeiros, exames laboratoriais e sumários de alta	Desidentificação, extração classificação, identificação e normalização de dados diversos	Sim, mediante autorização	Possui ambos dados anotados e não anotados	Sim
CCHMC	Inglês	3273	Radiologia renal e pulmonar e cartas de suicídio	Identificação de doenças de acordo com a CID-9 e análise de sentimentos	Sim para uso acadêmico e mediante requisição	Sim	Sim
MIMIC-IV	Inglês	448 972	Admissões de pacientes em CTI e sumários de alta, dados demográficos, exames de imagem, anotações médicas, etc	Diversas	Sim, mediante a autorização e exigir que se faça um curso	Não	Sim
BioScope	Inglês	3236	Resumos de documentos de saúde, artigos completos de medicina e textos livres de radiologia	Estudar níveis de asserção	Sim	Sim	Sim
THYME	Inglês	1254	Registros clínicos de pacientes com câncer no cólon	Identificar relações temporais	Sim, mediante requisição	Sim	Sim
CPRD	Inglês	11 milhões	Prontuários de pacientes	Diversas	Não	Não	Sim
THIN	Inglês	$\cong$ 3,9 milhões	Prontuários longitudinais de pacientes	Diversas	N/A	Não	Sim
Stockholm EPR *corpus*	Sueco	2 milhões	Prontuários de pacientes	desidentificação, anonimização, identificação de incertezas, negações e asserções, anotações clínicas, classificação de documentos e relações temporais	Sim, mediante requisição	Sim	Sim
EMC Dutch Clinical corpus	Holandês	7500	Prontuários longitudinais de pacientes, receituários, laudos de radiologia e sumários de alta	Relações temporais, negação e análise de experiências dos usuários	Sim, mediante requisição	Sim	Sim
Turku Clinical Treebank and Propbank	Finlandês	8	Documentos de enfermagem	Análise de dependência de estruturas linguísticas	Sim	Sim	Sim
Synthethic Clinical Text corpus	Alemão	55	Anotações clínicas e sumários de alta	REM	Sim	Sim	Sim
Chinese Electronic Medical corpus	Chinês	255	Prontuários de admissão	REM	Sim	Sim	Sim
MEDDOCAN	Espanhol	1000	Casos clínicos	REM e detecção de dados sensíveis	Sim	Sim	Sim

Analisando os *corpora* encontrados e suas características, foi observado a predominância de conteúdo em inglês. Uma hipótese para esse fato é a consolidação do inglês como lingua franca. Quando se trata de *corpus*, no entanto, a questão está mais intimamente ligada a língua materna dos pesquisadores e da localização de centros de pesquisa onde trabalham. Nesse sentido, observa-se que todos os *corpora* em inglês são originários de projetos nos Estados Unidos e na Grã-Bretanha, dois países falantes nativos do inglês que também figuram entre os cinco maiores produtores científicos do mundo (National Science Board, 2018).

No entanto, por se tratar de uma questão linguística, não é possível concentrar todos os estudos em inglês, ainda que os resultados sejam majoritariamente publicados na língua inglesa. Cada língua possui suas particularidades, vocabulários e formas diferentes de se expressar. Isso exige, naturalmente, abordagens diferentes em idiomas diferentes. Destaca-se, portanto, a importância da representatividade dos *corpora* em outros idiomas encontrados disponíveis para uso.

A desidentificação foi uma característica encontrada em todos os *corpora*. Pelo fato de o objeto de estudo ser dados de pacientes, é necessário remover os dados sensíveis. Além de proteger os pacientes, esse procedimento também desvincula os *corpora* de qualquer entidade mencionada, protegendo, assim, os pesquisadores. Outra vantagem desse procedimento é a possibilidade de disponibilização desse conteúdo sem precisar de autorização de pacientes (BRASIL, 2018). Apesar disso, são poucos os *corpora* que disponibilizam esses dados sem um processo de requisição ou autorização. Acredita-se que esse procedimento é uma cautela, extra, tomada pelos pesquisadores que disponibilizarão os dados. É uma maneira de assegurar que os dados serão usados de forma ética, com fins de pesquisa e sem o intuito de prejudicar a instituição ou os pacientes que forneceram esses dados primariamente.

Apenas dois *corpora* não foram anotados. Coincidentemente são *corpora*-monitores de aplicações diversas. Os demais *corpora* anotados tiveram como objetivo o desenvolvimento de aplicações específicas. Conclui-se que *corpora* não anotados fornecem aos pesquisadores um texto bruto que pode ser trabalhado para pesquisas de diversas finalidades. A contrapartida é que os trabalhos de pesquisa precisarão de tempo extra para desenvolverem métodos e técnicas de anotação.

As aplicações desenvolvidas com base nos *corpora* apresentados são diversas. É importante citar o i2b2 como exemplo dessa diversidade, pois a organização ocasionalmente promove diferentes desafios para PLN em dados clínicos. Isso só é possível graças a uma base grande de dados e a um comprometimento da comunidade com a evolução da pesquisa. É possível identificar que cada uma das aplicações possuem importância e relevância no mundo científico e comercial. Seja para identificação de doenças, classificação automática, relações entre sintomas e medicamentos, REM para desidentificação, entre outros.

As categorias de documentos também são diversas, com predominância de prontuários, pois são documentos completos sobre a história do paciente. Outro documento comum encontrado em diversos *corpora* é o **sumário de alta**. Esse documento costuma ser preterido para o processamento de texto por se tratar de um documento formal, relevante e escrito corretamente sem uso de gírias, abreviações e outras situações corriqueiras usadas para diminuir o texto. Os sumários de alta contêm um resumo da história do paciente que vai desde a sua admissão até a alta, contando o histórico de tratamento do paciente.

Já o tamanho desses *corpora* são discrepantes indo de oito documentos até mais de um bilhão. Isso indica ser possível fazer *corpora* relevantes com baixo número de documentos. De outra maneira, os *corpora* gigantes fornecem mais material para estudos e aplicações diversas. Não é possível concluir, nessa análise, qual seria o tamanho ideal para a criação de um *corpus* de pacientes. No entanto, é sugerido que a aquisição de dados seja a maior possível, devido à representatividade desses dados e também a possibilidade da elaboração de múltiplos trabalhos derivados de uma base mais robusta.

2.5 Estratégias para construção de *corpora* de documentos clínicos

A revisão de literatura do presente capítulo permitiu a elaboração etapas com estratégias de criação de *corpora* de documentos clínicos para domínio público. São ao todo sete etapas que orientam a tomada de decisão para a construção desses *corpora* sendo discutidas nas subseções seguintes.

2.5.1 Definir o escopo do *corpus*

A primeira etapa de construção define o escopo do *corpus*. A projeção de uso do *corpus* é fundamental para a orientação das próximas etapas e também para evitar possíveis retrabalhos. Entre os tipos disponíveis de escopos para a construção de *corpora* podem estar: **escopo único**, **multiescopo** ou **sem escopo**.

Corpora com apenas um escopo são desenvolvidos visando uma única aplicação. Exemplos de aplicações podem ser encontrados na Tabela 1 como REM, desidentificação, identificação de relações temporais, entre outros. Da mesma maneira, *corpora* com múltiplos escopos terão dois ou mais escopos definidos no início de sua construção. Por fim, um *corpus* sem escopo definido é projetado para atuar de forma genérica, podendo ser adaptado para diferentes aplicações. Com o escopo do *corpus* definido, dá-se início a segunda etapa.

2.5.2 Definir o tamanho do *corpus*

A segunda etapa define o tamanho dos *corpora*, podendo ser medido pela quantidade de *tokens* ou pela quantidade de documentos. Observou-se na Tabela 1 que *corpora*

com apenas um escopo, possuem baixa quantidade de documentos, indo de 8 até 3273. Quando observamos *corpora* multiescopos ou sem escopo definido a quantidade de documentos aumenta consideravelmente. A hipótese para esse comportamento dá-se tanto pela quantidade de aplicações distintas que esses *corpora* comportam, tanto pela necessidade de construção de *corpora* genéricos para atender diferentes aplicações. Nesses casos, a quantidade de documentos encontrados variam de 7500 até 1,2 bilhão de documentos.

Dessa maneira, recomendam-se pequenos volumes de documentos para *corpora* de escopos únicos. Nesse caso, a importância para a construção de *corpora* é resolver uma aplicação específica definida pelo escopo. Uma estratégia para definir a quantidade de dados a ser utilizada num *corpus* é fazer a incrementação periódica com testes. Primeiro avalia-se o desempenho da aplicação perante a quantidade de dados coletados previamente. Se os resultados se mostrarem insatisfatórios ou inconclusivos adiciona-se mais dados aos *corpora*. Esse processo pode ser repetido até que a quantidade coletada de dados seja suficiente para a realização da aplicação proposta. Portanto, para uma primeira coleta, recomenda-se cerca de 3254 documentos. Esse número foi obtido tirando-se a mediana de documentos encontrados nos *corpora* de escopos únicos na Tabela 1.

Para *corpora* multiescopos recomendam-se grandes volumes de documentos. Em primeira instância, pode-se adotar a estratégia de 3254 documentos para cada escopo incluído no projeto, desde que o desempenho das aplicações propostas sejam satisfatórios. Para *corpora* sem escopo a recomendação é que a quantidade de documentos seja a maior possível, visando abranger maior capacidade de adaptação em aplicações futuras. A Tabela 2 sintetiza as sugestões dadas nessa subseção. É importante ressaltar que esses números são valores de referência e podem mudar conforme a necessidade de ajuste de cada *corpus*. A maneira como esses *corpora* são construídos depende da escolha ou não de escopos. Esse processo é detalhado na terceira etapa.

Tabela 2 – Sugestão de tamanho de *corpora* em relação à quantidade de escopos.

Escopo do *corpus*	**Quantidade de documentos**
Escopo único	3254
Múltiplos escopos	3254 × (Quantidade de escopos)
Sem escopo definido	$\approx +\infty$

2.5.3 Definir a composição do *corpus*

A terceira etapa ocorre simultaneamente com a segunda, durante a definição do escopo. O tipo de escopo definirá a característica do *corpus* que será construído. *corpora* de escopos únicos terão estruturas simples com uma coleção de documentos focada na resolução das aplicações. Sugere-se construir um ***corpus* de estudo** para a resolução de uma aplicação específica. Segundo Viana e Tagnin (2011), *corpus* de estudos é a

terminologia usada para referir-se a uma coletânea de documentos para pesquisa, nesse caso o desenvolvimento de aplicações.

Por exemplo, um *corpus* com a finalidade de executar uma aplicação para reconhecer nomes de medicamentos deve ser primariamente construído com documentos onde essas entidades são mencionadas como receituários, sumários de alta e evolução médica. Documentos financeiros de unidade de saúde, por exemplo, não são interessantes para a resolução da aplicação e deverão ser desconsiderados.

Para *corpora* multiescopos recomenda-se a construção de um *corpus* composto de múltiplos *corpora* de estudo. Esse tipo de *corpus* representa o contexto linguístico de saúde e ajuda no entendimento de diferentes escopos. Essa representação linguística dependerá da quantidade de escopos definidos para o projeto do *corpus*. Ela pode representar desde uma especialidade da medicina, a própria unidade de saúde ou até mesmo o vocabulário da área de saúde em sua totalidade. Um *corpus* sem escopo definido pode ser representado tanto como um ***corpus* de referência** tanto como um ***corpus* monitor**. Um *corpus* de referência serve de termo de comparação para *corpora* de estudos. Geralmente possui cinco vezes o tamanho de um *corpus* de estudo com uma variação linguística maior. Já o *corpus* monitor é composto de uma coleção de documentos constantemente incrementada para acompanhar a evolução da língua (VIANA; TAGNIN, 2011). Esse modelo é recomendado caso haja a disponibilidade de tempo e recursos para a sua criação, dado que possui a maior concentração de documentos. A Tabela 3 sintetiza a sugestão de característica de construção do *corpus*. Uma vez definidos escopo e estrutura do *corpus*, a próxima etapa é a coleta de dados.

Tabela 3 – Características de *corpora* para composição.

Escolha de escopo	**Tipo de *corpus***
Escopo único	*corpus* de estudo
Múltiplos escopos	*corpus* composto de múltiplos *corpora* de estudo.
Sem escopo definido	*corpus* de referência ou *corpus* monitor

2.5.4 Coletar os documentos

A quarta etapa define quais documentos devem ser coletados. A escolha de documentos está intimamente ligada com o escopo do *corpus* e sua estrutura. Os documentos mais comuns são os prontuários de pacientes e seus subprodutos como sumários de admissão, alta, entre outros. Esses documentos são preferidos pelos pesquisadores devido a sua formalidade, baixa taxa de erros de digitação, legibilidade, etc. Em geral é o escopo que decidirá quais as categorias de documentos mais relevantes para pesquisa. Por exemplo, para análise de sentimento de pacientes depressivos, os documentos mais relevantes seriam cartas de suicídio ou transcrições de sessões de terapia. No entanto, quando os escopos são

múltiplos, são necessários diferentes tipos de documentos para garantir a diversidade e representatividade linguística do *corpus*. O mesmo raciocínio se aplica em *corpora* sem escopos definidos, visto que eles precisam atender ao máximo de aplicações possíveis.

2.5.5 Desidentificar os documentos

A quinta etapa, de desidentificação, é necessária para a proteção à privacidade de pacientes. Ainda que o escopo do *corpus* seja a desidentificação, essa etapa precisa ser executada previamente substituindo-se todos os dados sensíveis por pseudônimos. Essa etapa deve ser executada, preferencialmente, manualmente, pois é preciso assegurar que os dados sensíveis sejam removidos completamente antes que as aplicações sejam desenvolvidas. Essa etapa garante a segurança de profissionais responsáveis pela construção do *corpus*, bem como possibilita a realização da sexta etapa discutida a seguir.

2.5.6 Anotar os documentos

A sexta etapa é a anotação de dados para realização das aplicações propostas na subseção 2.5.1. Essa etapa se aplica apenas aos *corpora* que possuem um ou mais escopos definidos. *corpora* sem escopos não precisam ser anotados, visto que a anotação faz parte da preparação para execução de uma aplicação. Embora um *corpus* anotado ofereça mais significado, alguns autores argumentam que o estudo de *corpora* em sua forma pura evita as predileções e possíveis erros de anotadores (LEECH, 2004). Por isso, o modelo de anotação a ser executado durante a construção do *corpus* dependerá do escopo proposto na previamente. Por exemplo, um *corpus* que se propuser a desenvolver aplicações para identificar relações temporais entre a prescrição de remédios e os sintomas de pacientes, provavelmente realizará a anotação usando o TimeML[25] ou outra linguagem de marcação similar.

2.5.7 Disponibilizar o *corpus*

Por fim, a sétima etapa é a disponibilização do *corpus* para o público. O *corpus* pode ser disponibilizado na internet através de páginas de institutos responsáveis ligados diretamente aos documentos ou em plataformas gratuitas como GitHub[26]. Nessa etapa são discutidos os aspectos para a disponibilização de dados. Diferentes *corpora* mostrados na Tabela 1 são disponibilizados mediante autorização. Isso implica numa política de acesso como preencher formulários de solicitação, fazer cursos e respeitar eventuais licenças. Para cada política definida acima, deve-se elaborar o material correspondente. Um *corpus* pode estar sob uma licença proprietária ou de código aberto. A maioria dos trabalhos exige

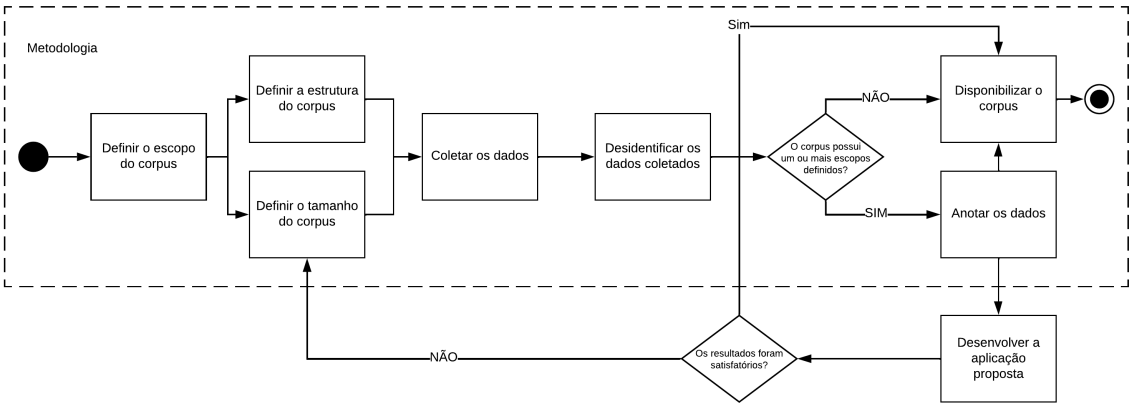
[25] Disponível em https://www.cs.brandeis.edu/ cs112/cs112-2004/annPS/TimeML12wp.htm. Acessado em março de 2022.

[26] Disponível em https://github.com/. Acessado em março de 2022.

uma documentação para requisição. As licenças exigem a citação do trabalho original na divulgação de resultados. Alguns poucos *corpora* são disponibilizados sem nenhuma exigência.

Ao término dessas sete etapas, tem-se um produto final, um *corpus* de natureza da saúde contendo dados desidentificados e disponibilizados para uso secundário da informação. A Figura 1 traz um fluxograma com o resumo das etapas apresentadas nesta seção. Note que o processo pode ser cíclico e algumas etapas podem ser revisitadas durante o processo de construção de um *corpus*. Se a aplicação desenvolvida para um *corpus* não atingir resultados satisfatórios, pode-se optar por coletar mais dados para tentar capturar uma variedade linguística maior e, consequentemente, novos padrões da língua. A etapa de desenvolvimento da aplicação, apesar de importante, não é discutida neste capítulo. Entende-se que esta etapa não faz parte da construção do *corpus*.

Figura 1 – Fluxograma de execução da metodologia.



Fonte: elaborado pelo próprio autor

As implicações das contribuições que esses *corpora* oferecem para estudos, tanto na área de medicina quanto na área de PLN, são discutidas na próxima seção.

2.6 Discussões

O fato de não serem encontrados *corpora* de documentos clínicos disponíveis em português é um fator preocupante. Uma pesquisa feita por Névéol et al. (2018) realizou um levantamento da produção científica de PLN em textos clínicos em outros idiomas além do inglês e evidenciou que os trabalhos realizados na língua portuguesa ocupam o décimo lugar no ranking com apenas 0,3% de representatividade entre as pesquisas realizadas no mundo. Esse fato reflete um pouco os dados levantados pelo National Science Board (2018), onde o Brasil ocupa a 11.ª posição entre os países que mais produzem pesquisa científica do mundo, sendo o país lusófono mais bem colocado. Dos quatorze trabalhos desenvolvidos para a língua portuguesa, encontrados no trabalho de Névéol et al. (2018),

apenas a pesquisa de Neves, Jimeno-Yepes e Névéol (2016) tem *corpus* como objeto de estudo. Trata-se, no entanto, de publicações científicas biomédicas e não de documentos clínicos, o objetivo desse estudo.

Acredita-se que a construção e disponibilização de um *corpus* de documentos clínicos possa impulsionar a pesquisa de PLN na língua portuguesa, na área de saúde. Do ponto de vista da ciência, um *corpus* propicia a reprodutibilidade e comparação de pesquisas sobre o mesmo objeto estudo. Para tal, parte-se do pressuposto que todo *corpus* de documentos clínicos precisa ser, primeiramente, desidentificado. Removendo-se os dados sensíveis, o material fica apto a ser utilizado por outras pesquisas. Quando dados sensíveis são eliminados, elimina-se também a necessidade de requisição ao COEP ou CONEP para acesso aos dados. Como visto na Tabela 1, a maioria dos *corpora* está sob tutela de pesquisadores sendo disponibilizados gratuitamente sem a necessidade de processo de requisição a um órgão controlador. Isso reduz o tempo de acesso aos dados enquanto aumenta a produtividade de pesquisa.

O PLN aplicado à área de saúde permite o desenvolvimento de diferentes aplicações. Há os sistemas de suporte à decisão clínica (DEMNER-FUSHMAN; CHAPMAN; MCDONALD, 2009), predição de casos clínicos baseados em sintomas (YANG et al., 2009), identificação de doenças raras (MACLEOD et al., 2016), para citar alguns exemplos dessas aplicações. Acredita-se que essas aplicações precisam ser desenvolvidas no contexto da língua portuguesa para haver mais oportunidades de melhorias do sistema de saúde.

Para isso, é importante levantar a questão da construção de um *corpus* de documentos clínicos para a língua portuguesa e defender a bandeira do acesso público e gratuito desses dados para fomentar as pesquisas no país. Se possível, esse *corpus* deve ser representativo. Ou seja, deve conter uma quantidade de dados suficiente para englobar todas as áreas envolvendo atendimento de saúde. Nesse sentido, o *corpus* deve ter as dimensões e propósitos similares ao CPRD e ao THIN. O que é, entretanto, uma meta ambiciosa, mas que deve ser objetivada. Em ambos os casos, há uma sincronia entre os órgãos de saúde britânicos e as entidades prestadoras de serviço para fornecer os dados a serem tratados. Para que o mesmo aconteça no Brasil será necessária uma série de adoções de políticas públicas e convênios com empresas e universidades para desenvolverem um projeto de mesmo porte. Esse *corpus* não precisa ser anotado, visto que sua disponibilização seria para uso geral. Sendo assim, a anotação seria desenvolvida em aplicações distintas derivadas desse *corpus* geral, originando *subcorpora* de estudo.

Enquanto isso não acontece, é necessário estimular a criação de pequenos *corpora* de estudo. O primeiro e mais importante *corpus* é para a aplicação de REM para fins de desidentificação. Qualquer *corpus* de paciente precisa ser desidentificado e o desenvolvimento dessas técnicas para a língua portuguesa poupará esforços futuros. É importante ressaltar que esses *corpora* sejam disponibilizados de alguma forma ao público geral. Isso

facilita o acesso e incentiva novos pesquisadores a trabalharem usando dados previamente coletados e tratados. Recomenda-se trabalhar com **sumários de alta** e **prontuários de admissão** nesses pequenos *corpora*. Por serem documentos mais formais, eles são uma forma preliminar de obter dados ricos em conteúdo e de fácil manuseio. Quanto ao tamanho desses *corpora*, dependerá do planejamento de escopos. Da mesma maneira, o tipo de anotação que conterão esses *corpora* também dependerá do escopo definido. Espera-se que este trabalho jogue uma luz para o desenvolvimento desta importante área carente de referências no assunto. Como contribuição, este capítulo apresentou uma metodologia para orientar os trabalhos nesta área.

3 Sumário de alta no Brasil

Uma das discussões levantadas no Capítulo 2 foi sobre qual é o documento mais adequado na construção de um *corpus* de documentos clínicos. Foi advogado que esse *corpus* possua representatividade quanto à diversidade de assuntos produzidos na área de saúde. Do ponto de vista teórico da linguística, o cenário ideal seria compor um *corpus* monitor que represente a variedade linguística da área de saúde na sua totalidade.

Para aplicações de nicho, em que *corpora* de estudos são usados, o sumário de alta é um candidato a documento a ser estudado. O sumário representa um documento formal, de fácil leitura e estruturado, o que permite aos pesquisadores o desenvolvimento de ferramentas mais eficientes para a mineração e análise de dados desse tipo de documento. Além disso, o sumário é conhecido por ser um resumo de toda a história do paciente durante sua permanência em instituições de saúde. A contrapartida é que tal documento apresenta apenas informações resumidas sobre o histórico do paciente e não contempla os informalismos que estão presentes noutros documentos. Isso, de certa forma, faz com que a variedade linguística capturada pelo sumário de alta esteja restrita apenas ao formalismo que ele emprega.

A escolha, no entanto, é justificada para o escopo deste trabalho. Por tempo e aplicabilidade, o sumário de alta é um bom ponto de partida para a tarefa de proteção de dados sensíveis em documentos da área de saúde. Conforme comentado no Capítulo 2, a primeira etapa no processo de proteção de dados sensíveis na área de saúde é o desenvolvimento de aplicações de desidentificação. Para esse propósito, a escolha do sumário de alta como documento para compor o *corpus* de estudo é justificada pela sua formalidade e relevância entre outros documentos da área de saúde. O sumário de alta desidentificado fornecerá conteúdo relevante para pesquisa no uso secundário de saúde.

É necessário, portanto, conhecer o sumário de alta em suas particularidades. É preciso ter uma base referencial sólida que suporte o sumário de alta como objeto de estudo deste trabalho e justifique a metodologia deste projeto. Para atingir tal objetivo, este capítulo esmiúça o sumário de alta. A seção 3.1 descreve sua importância e seu potencial para uso secundário da informação. A seção 3.2 apresenta as influências que moldaram o sumário de alta brasileiro. A seção 3.3 descreve o modelo de informação usado no Brasil para o sumário de alta. Por fim, a seção 3.4 sintetiza o capítulo com as discussões finais.

3.1 Importância do sumário de alta como documento de informação

O sumário de alta é um documento composto por registros que montam resumidamente a história do paciente durante o período em que foi atendido numa unidade de saúde. Evolução clínica, procedimentos assistenciais, intervenções, diagnósticos e tratamentos adotados são exemplos de registros que podem ser encontrados num sumário de alta (Ministério da Saúde, 2017). Um dos objetivos desse documento é fornecer meios para estabelecer a continuidade de tratamento do paciente em necessidades futuras Dean et al. (2016).

Para entender a importância do sumário de alta como documento de informação, é preciso entender o contexto ao qual ele pertence. Desde 1990, os brasileiros possuem acesso ao SUS, sistema que garante assistência universal de cuidados de saúde a todos (BRASIL, 1990). Uma das diretrizes da lei de 1990 é a descentralização político-administrativa dos cuidados de saúde. Na prática, cada unidade de saúde, respeitando as hierarquias do SUS, é responsável pela informação gerada em atendimentos. Outra diretriz garante que todos os cidadãos tenham integralidade de assistência em diferentes níveis de complexidade. Isso traduz-se na necessidade de comunicação e interoperabilidade de diferentes unidades de saúde para haver uma atuação contínua, humana e ágil na assistência ao cidadão (REIS et al., 2015). Ainda segundo Reis et al. (2015), a continuidade de assistência do cidadão pressupõe a longitudinalidade, prática que estende as relações entre médico e usuário para além de episódios específicos de tratamento.

Nesse contexto, o sumário de alta mostra-se um documento importante para a interoperabilidade de unidades de saúde. Com a informatização de processos de saúde, o sumário de alta fornece as informações necessárias para a troca de informações entre essas unidades. O sumário de alta também é usado para conectar os níveis primário, onde atuam as Unidades Básicas de Saúde UBS, secundário, em que atuam as Unidades de Pronto Atendimento UPA e terciário de saúde, em que atuam os hospitais de grande porte.

De acordo com o Ministério da Saúde (2017), o sumário de alta é um documento essencial para:

- apoiar eficientemente a comunicação das informações entre os diferentes níveis de atenção de saúde;
- contribuir para uma atenção coordenada entre os cuidadores do indivíduo;
- apoiar a continuidade dos cuidados de saúde;
- fornecer dados que melhorem a qualidade de atenção em saúde;
- contribuir para reduzir o número de reinternações;

- garantir a padronização das informações e interoperabilidade entre sistemas;
- facilitar a análise de dados, tomada de decisão e produção de conhecimento;
- reduzir o custo de manutenção de sistemas de saúde;
- reduzir o tempo de implantação de sistemas de informação de saúde;
- promover simplificação e padronização da comunicação de eventos de saúde;
- alimentar o registro pessoal de saúde do indivíduo;
- fornecer subsídios ao uso secundário da informação.

Os argumentos citados até então são suficientes para compreender a dimensionalidade e importância do sumário de alta na área de saúde. Para complementar esses argumentos, é necessário também debruçar-se sobre a estrutura do sumário de alta, suas influências e as informações que contêm. Esses assuntos são discutidos nas próximas seções.

3.2 Influências do sumário de alta brasileiro

Por se tratar de um documento formal, espera-se que o sumário de alta seja um documento que sirva tanto para a comunicação entre profissionais de saúde, quanto para leigos como pacientes, familiares e seus representantes legais. Dessa forma, há uma necessidade de que o documento seja conciso, preciso, objetivo, de fácil compreensão profissional e apresente informações estruturadas e organizadas. Este é, inclusive, um dos objetivos do Ministério da Saúde (2017) por meio de consulta pública para adotar modelos de informação referentes ao sumário de alta.

Uma das influências do sumário de alta brasileiro é a *Joint Comission International* (JCI), empresa sem fins lucrativos que tem como objetivo a melhoria mundial de processos de saúde. A JCI fornece uma série de normativas e boas práticas, dentre elas questões relativas ao sumário de alta. Uma dessas normativas se refere aos componentes que nele devem estar presentes. São eles (KIND; SMITH, 2008):

- motivo de internação;
- descobertas significativas;
- procedimentos e tratamentos realizados;
- condição do paciente no momento da alta;
- instruções de assistência continuada para o paciente e a família;

- assinatura do médico.

Ainda a JCI estabelece que cabe aos hospitais decidirem a forma onde os sumários de alta serão construídos e disponibilizados. Ou seja, embora haja um consenso de quais dados compõem o sumário de alta, a sua disposição no documento é realizada livremente pelas instituições de saúde (Joint Commission International, 2020).

Esse cenário também é refletido no Brasil. Segundo a ABNT/CEE-78 (2014), não há um consenso sobre quais informações deve conter um sumário de alta. As regulações vigentes mais próximas são a Autorização de Internação Hospitalar (AIH) para a rede pública (Ministério da Saúde, 2007) e o Resumo de Internação Hospitalar como parte do TISS (Documento usado para a Troca de Informação em Saúde Suplementar). Tanto a AIH quanto o TISS são de natureza comercial, usados para prestação de serviços de saúde por seguradoras (BRASIL, 2016). Eles não fornecem a informação necessária para que a assistência continuada ao paciente seja efetuada.

A proposta de um sumário de alta padronizado no Brasil foi feita por ABNT/CEE-78 (2014) para entrar em conformidade com os padrões internacionais. O documento tem como base as padronizações feitas pelo já mencionado Joint Commission International (2020), National e-Health Transition Authority (2010) e The Health and Social Care Information Centre (2013). O documento de National e-Health Transition Authority (2010) refere-se às padronizações realizadas pelo governo australiano por meio do NEHTA, agência reguladora de saúde do país. Já o documento de The Health and Social Care Information Centre (2013) refere-se ao HSCIC, órgão de regulação britânica que padroniza os documentos de saúde, incluindo o sumário de alta.

Além disso, o documento da ABNT/CEE-78 (2014) indica que um estudo nacional analisou sumários de alta de diferentes hospitais, públicos e privados, que possibilitou a comparação de campos existentes nos sumários de alta analisados. Como resultado, foram listadas sete categorias principais:

- identificação do paciente;
- diagnóstico(s);
- resumo e evolução de quadro clínico;
- exames ou procedimentos realizados;
- informações de internação (data e hora);
- identificação do profissional;
- recomendações pós-alta.

Observa-se que o resultado produzido pela ABNT/CEE-78 (2014) possui muitas semelhanças com os padrões internacionais citados pela Joint Commission International (2020). Na Tabela 4 é possível ver essas semelhanças. Cada linha representa uma característica observada nos dois modelos. A linha onde as duas colunas estão preenchidas representa o mesmo conceito observado em ambas as normativas. A diferença ficou com a identificação do paciente que está somente na normativa da ABNT.

Embora os conceitos das normativas apresentadas na Tabela 4 possam variar na prática, é importante ressaltar que a ideia principal de qual informação deve estar contida no sumário de alta está presente em ambas as diretrizes. Isso também valida o sumário de alta brasileiro como um documento alinhado com as perspectivas internacionais.

Tabela 4 – Comparativo entre as normativas da JCI e ABNT

Normativas da Joint Commission International (2020)	**Normativas da ABNT/CEE-78 (2014)**
Motivo de internação	Informações de internação (data e hora)
Descobertas significativas	Diagnóstico(s)
Procedimentos e tratamentos realizados	Exames ou procedimentos realizados
Condição do paciente no momento da alta	Resumo e evolução de quadro clínico
Instruções de assistência continuada para o paciente e família	Recomendações pós-alta
Assinatura do médico	Identificação do profissional
	Identificação do paciente

Com base nessas categorias, foi possível elaborar um modelo de informação do sumário de alta. Esse modelo, que será discutido na próxima seção, é importante para entender quais tipos de informação possuem potenciais dados sensíveis a serem protegidos por um tratamento de informação.

3.3 Modelo de informação de sumário de alta

Pode-se entender o modelo de informação como a diretriz básica de implementação de um documento. O modelo identifica a forma de representar conceitos, relacionamentos, restrições, regras e especificações de dados (VERYARD, 1992).

No caso de sumário de alta, o modelo de informação é responsável por definir as diretrizes de implementação desse documento. Ou seja, quais serão os dados que o sumário de alta conterá, o tipo de informação, a ocorrência obrigatória ou opcional deles e a formatação de dados. Em sistemas de informação digital, o modelo de informação pode ser convertido num Diagrama de Entidade e Relacionamento (DER), um modelo visual de dados usado para implementação de SGBDs.

Os bancos de dados constituem o cerne principal para armazenamento de dados em Sistemas de Informação digitais. Através de propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade), os SGBDs conseguem persistir os dados clínicos enquanto oferecem segurança e disponibilidade para profissionais de saúde.

O Ministério da Saúde (2017) propôs quais são os componentes do modelo de informação de sumário de alta. Eles são mostrados na Tabela 5. Ao analisar essa tabela é importante considerar o método usado:

- A primeira coluna descreve o nível de importância do elemento no modelo de informação. Esse nível varia de 1 a 5 sendo 1 o mais importante;
- A segunda coluna descreve a ocorrência que o elemento tem no sumário, podendo ter as seguintes representações:
  - [0..1]: elemento opcional que pode ocorrer, no máximo, uma vez;
  - [1..1]: elemento obrigatório que deve ocorrer apenas uma vez;
  - [0..N]: elemento opcional que pode ocorrer zero ou mais vezes;
  - [1..N]: elemento obrigatório que pode ocorrer uma ou mais vezes;
- A terceira coluna apresenta o tipo de informação que deve ser preenchida;
- A quarta coluna apresenta a formatação esperada da informação;

Tabela 5 – Modelo de informação do sumário de alta. Adaptado de Ministério da Saúde (2017).

Nível	**Ocorrência**	**Tipo de informação**	**Formatação**
1	[1..1]	Caracterização do atendimento	
1	[1..1]	Motivo da admissão, diagnósticos relevantes e patologias associadas desenvolvidas na internação	
1	[0..1]	Restrições funcionais e incapacidades em saúde	
1	[1..1]	Procedimentos realizados ou solicitados	
1	[1..1]	Resumo da evolução clínica do indivíduo durante a internação	
1	[0..1]	Alergias e/ou reações adversas na internação	
1	[1..1]	Prescrição da alta	

Continua na próxima página

Tabela 5 – Continuação da página anterior

Nível	Ocorrência	Tipo de informação	Formatação
1	[0..N]	Instruções, orientações e recomendações da alta	
1	[1..1]	Informações da alta	
1	[0..N]	Anexos com os resultados de exames	
1	[0..1]	Informações adicionais/complementares	
2	[0..1]	Identificação pelo Cartão Nacional de Saúde	
2	[0..1]	Identificação por dados demográficos	
2	[0..1]	Estabelecimento de Saúde	Número do Cadastro Nacional de Estabelecimentos de Saúde (CNES) válido
2	[0..1]	Local de atendimento	Texto codificado: Domicílio, Instituição/Abrigo, Unidade prisional ou congêneres, Unidade socioeducativa, Outros.
2	[1..1]	Procedência	Texto codificado: Ordem Judicial; Retorno; Demanda espontânea; Demanda referenciada.
2	[0..1]	Identificação da equipe de saúde	Número do Identificador Nacional de Equipe (INE)
2	[1..1]	Caráter da Informação	Texto codificado: Eletiva; Urgência
2	[1..1]	Data e hora da internação	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
2	[1..1]	Modalidade assistencial	Texto codificado: Atenção Domiciliar;
2	[1..1]	Terminologia que descreve o diagnóstico	Identificador único do objeto
2	[1..1]	Nome e versão da terminologia que descreve o diagnóstico	Texto livre
2	[1..N]	Diagnósticos	
2	[1..1]	Terminologia que descreve a restrição funcional ou incapacidade	Identificador único do objeto

Continua na próxima página

Tabela 5 – Continuação da página anterior

Nível	Ocorrência	Tipo de informação	Formatação
2	[1..1]	Nome e versão da terminologia que descreve a restrição funcional ou incapacidade	Texto livre
2	[0..N]	Restrições funcionais ou incapacidades	
2	[1..1]	Terminologia que descreve procedimento	Identificador único do objeto
2	[1..1]	Nome e versão da terminologia que descreve o procedimento	Texto livre
2	[1..N]	Procedimentos	
2	[1..1]	Descrição da evolução clínica do indivíduo durante a internação	Texto livre
2	[1..N]	Alergia e/ou reação adversa	
2	[0..1]	Medicamentos prescritos na alta (não estruturado)	
2	[0..1]	Lista de medicamentos da alta (estruturada)	
2	[1..1]	Descrição da instrução, orientação ou recomendação	
2	[1..1]	Data e hora da saída da internação	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
2	[1..1]	Condição do indivíduo na alta	Texto codificado: Bom estado geral; Falecido; Inalterado; Melhorado; Piorado
2	[1..1]	Desfecho da internação	Texto codificado: Alta clínica; Alta voluntária; Encaminhamento; Evasão; Óbito; Ordem Judicial; Permanência; Retorno; Transferência
2	[0..1]	Declaração de óbito	Número da Declaração de Óbito
2	[0..1]	Encaminhamento pós-alta	
2	[0..1]	Profissional responsável pela alta	
2	[1..1]	Descrição dos anexos	Texto livre
2	[1..1]	Descrição das informações	Texto livre

Continua na próxima página

Tabela 5 – Continuação da página anterior

Nível	Ocorrência	Tipo de informação	Formatação
3	[1..1]	Cartão Nacional de Saúde (CNS)	Número do CNS
3	[1..1]	Nome completo	Texto livre
3	[0..1]	Nome social	Texto livre
3	[1..1]	Nome completo da mãe	Texto livre
3	[1..1]	Data de nascimento	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
3	[1..1]	Sexo	Texto codificado: Masculino; Feminino; Ignorado
3	[1..1]	Raça/Cor	Texto Codificado conforme o Instituto Brasileiro de Geografia e Estatística IBGE: Branca; Preta; Parda; Amarela; Indígena; Sem Informação.
3	[0..1]	País de Nascimento	Texto codificado conforme Sistema de Cadastramento de Usuários do Sistema Único de Saúde (CADSUS).
3	[0..1]	Município de Nascimento	Texto codificado conforme IBGE
3	[0..1]	País de Residência	Texto codificado conforme CADSUS
3	[0..1]	Município de Residência	Texto codificado conforme IBGE
3	[0..1]	CEP de Residência	Texto codificado conforme Correios
3	[1..1]	Diagnóstico	Texto codificado por terminologia externa
3	[1..1]	Categoria do diagnóstico	Texto Codificado: Principal; Secundário
3	[1..1]	Indicador de presença na admissão	Texto Codificado: Sim; Não; Desconhecido
3	[1..1]	Categoria de atividade	Texto Codificado: Ativo; Inativo

Continua na próxima página

Tabela 5 – Continuação da página anterior

Nível	Ocorrência	Tipo de informação	Formatação
3	[1..1]	Estado de resolução	Texto Codificado: Resolvido; Resolvendo; Não resolvido; Indeterminado
3	[1..1]	Restrição funcional ou incapacidade	Texto codificado por terminologia externa
3	[1..1]	Status da restrição funcional ou incapacidade	Texto codificado: Ativo; Inativo
3	[1..1]	Procedimento	Texto codificado por terminologia externa
3	[1..1]	Tipo de procedimento	Texto codificado: Cirúrgico; Diagnóstico; Terapêutico
3	[1..1]	Status do procedimento	Texto codificado: Concluído; Suspenso; Solicitado
3	[0..1]	Resultado ou observações do procedimento	Texto livre
3	[1..1]	Categoria do agente causador da alergia ou reação adversa	Texto codificado: Alimento; Medicamento; Outros
3	[1..1]	Agente/substância específica	Texto livre
3	[0..1]	Manifestação	Texto livre
3	[0..1]	Grau de certeza	Texto codificado: Confirmado; Resolvido; Refutado; Suspeito
3	[0..1]	Criticidade	Texto codificado: Alta; Baixa; Indeterminada
3	[0..1]	Data/hora da instalação da reação adversa	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
3	[0..1]	Evolução da alergia/reação adversa	Texto livre
3	[1..1]	Descrição da prescrição	Texto livre
3	[1..1]	Terminologia que descreve o medicamento	Identificador único do objeto
3	[1..1]	Nome e versão da terminologia de medicamentos	Texto livre
3	[1..1]	Nome do profissional responsável pela instrução, orientação ou recomendação	Texto livre

Continua na próxima página

Tabela 5 – Continuação da página anterior

Nível	Ocorrência	Tipo de informação	Formatação
3	[1..1]	CNS do profissional responsável pela instrução, orientação ou recomendação	Número do CNS
3	[1..1]	Ocupação do profissional responsável pela instrução, orientação ou recomendação	Número de Classificação Brasileira de Ocupações (CBO) válido
3	[0..1]	Tipo de estabelecimento de saúde	Texto Codificado por terminologia externa (CNES)
3	[0..1]	Descrição do serviço ou especialidade	Texto livre
3	[1..1]	Nome do profissional	Texto livre
3	[1..1]	CNS do profissional	Número do CNS
3	[1..1]	Ocupação do profissional responsável pela alta	Número de CBO válido
4	[1..N]	Medicamentos	
5	[1..N]	Medicamento (nome do princípio ativo, concentração, unidade de medida e forma farmacêutica)	Texto codificado por terminologia externa
5	[1..1]	Quantidade da unidade farmacêutica	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
5	[1..1]	Unidade farmacêutica	Texto codificado por terminologia externa
5	[1..1]	Frequência de uso do medicamento	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
5	[1..1]	Via de administração	Texto codificado por terminologia externa
5	[1..1]	Duração de uso do medicamento	Texto padronizado conforme capítulo 6 de Ministério da Saúde (2017)
5	[1..1]	Estado do medicamento	Ativo, Descontinuado, Nunca ativo, Tratamento completo, Substituído
5	[0..1]	Orientação sobre o uso do medicamento	Texto livre

A Tabela 5 apresenta um modelo de informação extenso com um total de 92 tipos diferentes de informação. Alguns tipos são claramente compostos por informação sensível como nome, data de nascimento e CEP de residência. Em outros tipos, essa informação é

subjetiva. Durante a descrição dos anexos, por exemplo, um profissional de saúde pode se referir ao paciente pelo nome ou mencionar colegas de profissão. Não há como saber exatamente quantos tipos de informação podem ter dados sensíveis. Os tipos que possuem formatação preestabelecida com texto codificado dão um indicativo se contêm informação sensível ou não. Ao analisar o tipo de informação "estado de resolução", por exemplo, observa-se que os únicos valores possíveis são resolvido, não resolvido e indeterminado, um tipo de informação sem dados sensíveis. Os demais tipos de informação, principalmente os formatados como texto livre, são candidatos potenciais para conter dados sensíveis.

A implementação desse modelo de informação seria dada pelo modelo computacional, uma proposta que evidenciaria as estruturas da interface computacional do sumário de alta de forma que estabelecesse padrões para a interoperabilidade entre sistemas de saúde brasileiros (ABNT/CEE-78, 2014). Um modelo computacional de sumário de alta não foi elaborado até o momento em que esse trabalho foi escrito, embora os padrões de regulamentação já estejam aprovados pela Portaria N.º 2.073 do Ministério da Saúde (2011).

3.4 Discussões

No decorrer desse capítulo, ficou clara a importância do sumário de alta como um documento de saúde e de informação. O benefício aos pacientes é possibilitar a assistência continuada; aos profissionais de saúde, fornecer dados que melhorem a qualidade de atenção em saúde; aos profissionais de informação, fornecer insumos para uso secundário da informação.

Um dos pontos importantes do sumário de alta discutidos neste capítulo é seu uso secundário, ou seja, para outros fins além da saúde. Tarefas como análise, mineração de dados, tomada de decisão e produção de conhecimento são algumas das possibilidades por esse documento oferecidas. Esse ponto reforça a escolha desse documento como objeto de estudo do presente trabalho.

A criação de um *corpus* de saúde baseado em sumários de alta pode facilitar o compartilhamento e reúso da informação. Por ser um documento padronizado, a mineração e análise podem ser alcançadas sem muitos custos tecnológicos para extração e transformação de dados.

Uma análise do modelo do sumário de alta identificou um documento rico em informação a ser explorado em seu uso secundário. Embora o sumário apresente informações resumidas, há informação suficiente que o torna relevante para pesquisa e outras aplicações.

São ao todo 92 tipos distintos de informação que podem estar representados no sumário de alta. Eles possuem diferentes graus de importância e estruturas. Enquanto

algumas informações devem seguir terminologias e formatações preestabelecidas, outras estão em formato de texto livre, sendo de responsabilidade do profissional de saúde a escolha da forma de preenchimento.

Isso dá ao sumário de alta uma característica desafiadora. Embora seja um documento formal e padronizado, há também algumas brechas sobre como ele é estruturado e preenchido. A ausência do modelo computacional dá ao sumário de alta certa liberdade de implementação. Cabe então às instituições de saúde entender como o modelo computacional deve ser implementado. Os 92 tipos de informação listados na Tabela 5 podem ser implementados literalmente ou não. Ou seja, há a possibilidade de condensar alguns tipos de informação num único tipo ou até mesmo desmembrar um tipo de informação em vários outros.

Como resultado, tem-se um documento que possui dados em comum, mas com possíveis estruturas diferentes. Isso traz um desafio para um dos objetivos deste trabalho, a desidentificação. Os dados sensíveis podem estar expostos de diferentes maneiras num sumário de alta. Seja por tipos de informação bem explícitos como o nome do paciente ou número do CNS ou de forma implícita como citar o nome de algum profissional de saúde e/ou dado sensível de um paciente num tipo de informação em formato de texto livre.

Cabe ao profissional de informação entender o modelo computacional ao qual está submetido o sumário de alta e, assim, elaborar estratégias de proteção à privacidade para a disponibilização desses documentos para uso secundário. Esses desafios serão abordados no próximo capítulo, ao discutir técnicas de desidentificação e proteção à privacidade.

4 Desidentificação de documentos clínicos

A World Health Organization (2006) define um PEP como um conjunto de documentos que contém o perfil comportamental e de saúde de um paciente. Um PEP contém também o histórico de um paciente através de múltiplos episódios de atendimentos. Em outras palavras, é um documento longitudinal.

O PEP é escrito primariamente para uso interno em unidades de saúde com a finalidade de registro histórico. Outros profissionais de saúde que tenham contato com o paciente também podem escrever em seu prontuário. Outra razão da existência do PEP é legal, visto que a legislação de muitos países define a obrigatoriedade desse documento (DALIANIS, 2018).

Os PEPs contêm informações valiosas relacionadas aos tratamentos de saúde. Ter acesso a esses dados, no entanto, é uma tarefa onerosa devido aos trâmites necessários para obtenção desses dados no COEP. Dados sensíveis não podem ser manuseados sem o devido cuidado. Além disso, dizem respeito aos pacientes, donos da informação com direito à privacidade. Para resolver o problema, recorre-se à tarefa de desidentificação.

A desidentificação é o processo de remover todos os dados pessoais de um documento de forma que impossibilite a identificação de pessoas, locais, entidades, datas ocorridas etc. Existem diferentes usos do termo desidentificação na literatura. Há uma confusão entre os termos desidentificação, anonimização e pseudoanonimização. Alguns autores empregam os termos anonimização e desidentificação como sinônimos. Outros usam desidentificação como um processo e anonimização como um tipo específico de desidentificação irreversível. No contexto de saúde, alguns autores tratam os termos desidentificação e pseudoanonimização como equivalentes e o termo anonimização como a desidentificação irreversível (GARFINKEL, 2015). Para elucidar melhor o uso desses termos, o presente trabalho adotou as definições contidas na ISO 25237:2017, sobre pseudoanonimização, a saber (ISO, 2017):

- **desidentificação:** termo genérico para qualquer processo que remova a associação entre um conjunto de dados identificadores e o sujeito dos dados;
- **anonimização:** processo em que os dados pessoais são irreversivelmente removidos e o sujeito dos dados não pode ser mais identificado;
- **pseudoanonimização:** tipo particular de desidentificação que remove a associação entre o sujeito dos dados e os dados pessoais, mas adiciona outra associação que relaciona os dados com seu pseudônimo.

Portanto, pode-se afirmar que a desidentificação é o processo geral e pode ser dividido em duas categorias: anonimização e pseudoanonimização. A diferença entre dados anônimos e pseudônimos é que o primeiro tem toda a informação pessoal removida e é impossível associá-la novamente ao paciente original, garantindo, assim, sua proteção. O segundo tem os dados substituídos por pseudônimos ou códigos de segurança criptografados. Esses códigos podem ser descriptografados caso o paciente deseje ter acesso novamente aos seus dados (BERMAN, 2002). Autores como Stubbs et al. (2015) defendem que a desidentificação para fins de pesquisa seja feita sem prejudicar a leitura humana de registros médicos. Portanto é recomendado que a remoção de dados seja substituída por nomes ou códigos que deixem claro qual tipo de dado foi desidentificado. A substituição pode ser reversível (pseudoanonimização) ou irreversível (anonimização). A seguir, um exemplo fictício de como se faz a remoção de dados.

Elisa foi atendida pelo Dr. **Otávio** no **Hospital das Clínicas** no dia **25/07/2019**.
A paciente foi diagnosticada com dengue. (Texto Original)
(4.1)

Hellen foi atendida pelo Dr. **Samuel** no **Hospital Helton Cruz** no dia **13/06/2036**.
A paciente foi diagnosticada com dengue. (Texto Desidentificado)
(4.2)

A tarefa, no entanto, sofre com o **dilema da causalidade**. Dados não podem ser desidentificados sem que se tenha acesso a eles. Do mesmo modo, não se pode ter acesso aos dados sem que estejam devidamente protegidos. Para isso, é preciso que o interessado submeta-se a órgãos de controle e proteção de dados para que se tenha acesso a eles numa primeira vez. Então, o trabalho de desidentificação é realizado de forma que os dados já se encontrem tratados e prontos para uso nas próximas ocasiões em que houver necessidade de acesso a eles.

No Brasil e em outros países, para ter acesso aos documentos clínicos, é preciso realizar requisições ao comitê de ética solicitando permissão de pesquisa envolvendo animais ou seres humanos. No caso, o objeto de estudo são os dados referentes aos seres vivos e, dessa maneira, não há experimentos realizados diretamente com animais ou seres humanos. Ainda assim, essa característica de pesquisa não exime o pesquisador de ter que submeter o projeto de pesquisa a um comitê de ética. Os dados de pacientes contêm informações sensíveis de pessoas que não foram solicitadas a participar da pesquisa.

O procedimento padrão nesses casos seria solicitar a permissão para cada indivíduo, um Termo de Consentimento Livre e Esclarecido, ou TCLE. Esse termo foi definido pela primeira vez na resolução 196 do Conselho Nacional da Saúde (1996) e detalhado

posteriormente na resolução 466. A resolução 466 do Conselho Nacional da Saúde (2012) aprovou e regulamentou as diretrizes básicas para a realização de pesquisas entre seres humanos, constituindo-se referência para a ética aplicada à pesquisa no Brasil. Um dos destaques foi a formalização do TCLE e os passos necessários para requerer a autorização de pacientes alvos de pesquisa.

O TCLE, no entanto, pode inviabilizar pesquisas que tenham como objeto de estudo, milhares de pacientes. Isso tornaria qualquer pesquisa de análise de dados inviável, pois seria necessário um termo para cada usuário registrado numa base de dados. Uma forma de eliminar a necessidade de submissão de projetos de pesquisa no comitê de ética e pesquisa (CEP) é a desidentificação de PEPs. Conforme a resolução 510 do Conselho Nacional da Saúde (2016), está dispensada a avaliação do CEP "pesquisas que usem bancos de dados, cujas informações são agregadas, sem possibilidade de identificação individual".

O processo de desidentificação é útil em duas frentes diferentes. A primeira é garantir a privacidade de um paciente enquanto ele possui seus dados armazenados por terceiros. A ocultação de dados sensíveis garante que, em caso de vazamento, o paciente não seja exposto perante a sociedade. Informações como orientação sexual, religião, doenças sexualmente transmissíveis entre outras ainda enfrentam uma série de preconceitos na sociedade e podem trazer muitos prejuízos para um paciente. A segunda frente é a possibilidade de auxiliar as pesquisas que usam dados clínicos como fonte de informação. A ocultação desses dados, além de proteger os pacientes, permite que pesquisadores possam usá-los para extrair informações relevantes sem prejudicar as pessoas e as instituições envolvidas. A relevância de uso de dados clínicos como fonte de informação já é discutida entre muitos pesquisadores. Um estudo extensivo realizado por Meystre et al. (2017) lista os principais usos secundários para dados clínicos. Os principais benefícios de reuso, segundo os autores, tratam os seguintes problemas: privacidade e ética relacionada ao reuso de documentos; integração, interoperabilidade e construção de sistemas federados; modelos de dados e terminologias; extração de informação; mineração de dados; práticas clínicas e integração com pesquisas. A necessidade de desidentificação de documentos clínicos no Brasil já foi mencionada por Galvão e Ricarte (2011) que observam o potencial de uso na gestão, ensino e pesquisa do país. Para os autores, é preciso desenvolver, projetar e validar metodologias de desidentificação no Brasil, uma demanda antiga da informação clínica que precisa ser atendida.

Para isso, o presente capítulo faz uma revisão de literatura sobre o tema de desidentificação. Busca-se entender melhor o assunto da desidentificação para elaborar estratégias de uso dessa técnica para documentos clínicos na língua portuguesa do Brasil. A seção 4.1revela as principais categorias de dados sensíveis presentes em documentos clínicos. A seção 4.2 apresenta a importância da desidentificação e os riscos de reidentificação. A seção 4.3 apresenta o REM como principal técnica de desidentificação de linguagem natural.

A seção 4.4 mostra os métodos de desidentificação automatizados usados na literatura e suas respectivas *performances*. Por fim, a seção 4.5 sintetiza o presente capítulo trazendo as discussões necessárias para o trabalho.

4.1 Dados sensíveis em documentos clínicos

Para realizar o processo de desidentificação de documentos clínicos, é preciso, primeiramente, entender que dados sensíveis precisam ser removidos. Na literatura global, os casos de desidentificação são guiados pela *Health Insurance Portability and Accountability Act*, ou HIPAA (III; WEAVER; HUGHES, 2004). HIPAA é uma lei criada nos Estados Unidos em 1996 que lida com as principais questões relacionadas à indústria de saúde do país. As preocupações iniciais da HIPAA estavam endereçadas ao gasto relacionado aos seguros de saúde. No entanto, conforme mais discussões surgiam, novas questões foram levantadas, incluindo a privacidade dos dados. Desde a criação da HIPAA, mais de 400 medidas foram tomadas para a melhoria de dados de saúde.

Uma dessas medidas é o PHI, *Protected Health Information*, que define os elementos que devem ser removidos num documento clínico a ser considerado seguro para uso secundário da informação (Department of Health and Human Services, 2000, p.358). Esses elementos fazem parte de uma declaração que a HIPAA fez para garantir que os dados sejam tratados, desidentificados e prontos para serem manipulados sem oferecer riscos aos indivíduos. Essa declaração possui dois tópicos de esclarecimento (UZUNER; LUO; SZOLOVITS, 2007):

1. Um especialista deve determinar e documentar que o risco de manipulação de dados é muito pequeno, de forma que possam ser usados sozinhos ou em conjunto com outros dados, ou;
2. Os dados, de uma lista de elementos que identifiquem um paciente, parentes, profissionais de saúde entidades devem ser eliminados de forma que seja impossível a identificação de um indivíduo.

O segundo tópico dessa declaração é o mais importante para a pesquisa em desidentificação, pois ela esclarece quais são os dados sensíveis que precisam ser removidos para ser possível a manipulação de dados sem colocar em risco os indivíduos proprietários da informação. Ao todo são 18 elementos que compõem a lista de proteção de dados dos pacientes para disponibilização pública de seus documentos clínicos. São eles:

1. Nomes e sobrenomes;
2. Todas as subdivisões geográficas menores que estado, incluindo endereço, cidade, distrito, vila, código de endereçamento postal e seus códigos geográficos equivalentes;

3. Todos os elementos de datas diretamente relacionados ao indivíduo como nascimento, admissão, alta e óbito;
4. Todas as idades acima de 89, bem como elemento de dados indicativos de tal idade;
5. Números de telefone;
6. Números de fax;
7. Endereços de e-mail;
8. Números ou códigos de previdência social;
9. Números ou códigos de registros e prontuários de pacientes;
10. Números ou códigos de beneficiários de planos de saúde;
11. Números ou códigos de contas-correntes;
12. Números ou códigos de certificados ou licenças;
13. Números ou códigos seriais ou de identificação de veículos, incluindo placas e licenças para dirigir;
14. Números ou códigos seriais ou de identificação de dispositivos, incluindo patrimônio, IMEI e licenças;
15. Endereços da web de qualquer tipo, sejam URLs, URIs, IPs ou similares;
16. Identificadores biométricos, incluindo impressões digitas, voz e íris;
17. Fotografias da face e imagens parciais, mas comparáveis;
18. Qualquer outro código único de identificação, caracterização ou codificação.

Segundo Meystre et al. (2010), os elementos listados pela HIPAA podem ser divididos em sete categorias genéricas, as quais podem ser usadas para a criação de ferramentas de desidentificação. São elas:

1. **Nomes:** paciente, membro da família, médico, fornecedor do seguro de saúde etc;
2. **Localidades geográficas:** endereços, cidades, CEP etc;
3. **Idades:** maiores que 89 anos;
4. **Locais de tratamento:** hospitais, clínicas, casas de repouso etc;
5. **Datas:** quaisquer elementos menores que o período de um ano;

6. **Informações de contato:** telefone, fax, e-mail, redes sociais etc;
7. **Números de identificação:** RG, CPF, carteira de motorista, passaporte, etc;

Os PEPs são, geralmente, armazenados em bancos de dados relacionais. Essas estruturas de armazenamento dividem as informações de pacientes em campos diferentes, numa tabela, para cada tipo de dado como nome, endereço, idade etc. O processo de identificação e remoção de dados sensíveis é facilitado, sendo preciso apenas excluir ou proteger os dados referente a esses campos para ocorrer a desidentificação.

Há estudos, no entanto, que mostram que essa atitude não é suficiente para a remoção de todos os dados sensíveis de PEPs. Grande parte desses dados é encontrada em anotações de textos livres, em que profissionais de saúde descrevem a história de um paciente. As anotações são armazenadas em campos de banco de dados que não obedecem a uma estrutura padrão. Os PEPs disponibilizados para análise pública são armazenados em *corpora*, conjuntos de textos eletrônicos selecionados seguindo um critério que respeite a variação linguística (ALUÍSIO; ALMEIDA, 2006). Um desses *corpora*, o *Stockholm EPR corpus*, por exemplo, possui 46% dos *tokens* em anotações de textos livres. Um total de 1,6% de *tokens* do *corpus* é formado por dados sensíveis (DALIANIS, 2018).

A concentração de dados sensíveis em *corpora* de documentos clínicos varia conforme o tipo de documento e o idioma. A Tabela 6 ilustra diferentes densidades de dados sensíveis encontrados na literatura.

Tabela 6 – Concentração de dados sensíveis em diferentes *corpora* de pacientes

Referência	**Tamanho do *corpus* em *tokens***	**Frequência relativa de dados sensíveis**	**Idioma**
(DOUGLASS et al., 2004)	339150	0,5	Inglês
(DORR et al., 2006)	70552	2,9	Inglês
(NEAMATULLAH et al., 2008)	334000	0,5	Inglês
(UZUNER et al., 2008)	472315	6	Inglês
(HANAUER et al., 2013)	20500	3,7	Inglês
(CARRELL et al., 2013)	22525	1,5	Inglês
(KOKKINAKIS; THURIN, 2007)	14000	10	Sueco
(VELUPILLAI et al., 2009)	174000	2,5	Sueco
(GROUIN; NÉVÉOL, 2014)	29437	13,4	Francês
(PANTAZOS; LAUESEN; LIPPERT, 2017)	73150	1,8	Dinamarquês

Os estudos de Douglass et al. (2004) e Neamatullah et al. (2008) referem-se ao *corpus* MIMIC II em épocas diferentes. Nota-se que a densidade permaneceu a mesma. Dorr et al. (2006) também analisaram quais os profissionais de saúde que mais inseriam dados sensíveis nos documentos, sendo os médicos em primeiro lugar, seguido dos enfermeiros. Velupillai et al. (2009) encontraram nomes pessoais em 33% de dados sensíveis, enquanto o estudo de Grouin e Névéol (2014) encontrou 41% em seus respectivos *corpora*.

Os estudos feitos por Hanauer et al. (2013) e Henriksson, Kvist e Dalianis (2017) também analisaram quais documentos possuem a maior densidade de dados sensíveis e os que aparecem com mais frequência. Ambos chegaram ao mesmo resultado, apontando os **sumários de alta** como o documento com maior incidência de dados sensíveis, datas e nomes como os que aparecem mais frequentemente.

Os números calculados na Tabela 6 levam a um desvio padrão[1] σ de 4,08, um valor que mensura quanto os dados estão esparsos na distribuição de dados. Em outras palavras, há uma discrepância do percentual de dados sensíveis achados por diferentes autores. Algumas hipóteses podem ser levantadas a respeito. A primeira delas diz respeito aos tipos de documentos encontrados. Sumários de alta, narrativas clínicas, registros de pacientes e prontuários são exemplos dos documentos estudados pelos autores. A segunda hipótese refere-se à amostragem estatística. Todos os trabalhos apresentados são amostras de alguma unidade de saúde. Estatisticamente falando, diferentes amostras podem produzir diferentes estatísticas numa mesma população. Assim, pode-se dizer que os percentuais encontrados pelos autores são resultados do acaso de suas amostras. As hipóteses, no entanto, precisam ser confirmadas.

Para contornar essa dispersão de dados e oferecer um parâmetro de consulta, o presente trabalho optou por tirar a mediana dos dados sensíveis achados pelos diferentes autores. A mediana é a escolha usual de linguística de corpus para calcular o centro de uma distribuição de dados onde há valores extremos, como os da Tabela 6 (BREZINA, 2018). A mediana encontrada de dados sensíveis foi de 2,7%. Ou seja, podemos esperar uma média de **2,7%** de dados sensíveis num *corpus*. Essa análise traz um indicativo de quanto esperar de dados sensíveis em novos *corpora*.

4.2 Riscos da desidentificação

A desidentificação pode trazer dois riscos principais: (1) ter os dados sensíveis reidentificados e expostos ao público e (2) tornar os documentos incompreensíveis para a pesquisa (YOGARAJAN; PFAHRINGER; MAYO, 2019). Esses problemas precisam ser entendidos para que o processo de desidentificação obedeça a critérios de qualidade.

4.2.1 Reidentificação

A remoção de todos os dados sensíveis listados na seção 4.1 mostra-se necessária, pois autores como Sweeney (2000), Rothstein (2010), Li e Qin (2017) e Johnson et al. (2019) já demonstraram a possibilidade de reidentificar pacientes usando apenas três dos dezoito elementos da HIPAA. O processo de reidentificação envolve usar dados de uma pessoa

[1] O desvio padrão é dado pela equação $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

para identificá-la. Alguns dados mais explícitos como nome e endereço são exemplos claros de identificação. No entanto, é possível identificar pessoas usando semi-identificadores como CEP, data de nascimento e gênero, por exemplo. Esses dados podem ser cruzados com informações externas e, assim, identificar uma pessoa (EMAM et al., 2006; MAYO; YOGARAJAN, 2019).

Além disso, há casos em que indivíduos não desejam ser associados às informações sensíveis. Entre eles estão as doenças psiquiátricas, AIDS, câncer e doenças raras. Essas condições médicas geralmente requerem tratamentos e rotinas diferenciadas que, quando combinados com semi-identificadores, podem ser usados para identificar um paciente (EGUALE; BARTLETT; TAMBLYN, 2005; GKOULALAS-DIVANIS; LOUKIDES; SUN, 2014).

4.2.2 Compreensibilidade

A desidentificação de documentos clínicos envolve a remoção e/ou substituição de dados sensíveis. Para fins de pesquisas, no entanto, a substituição de dados por pseudônimos é tida como padrão, dado que a simples remoção de dados prejudicaria a compreensibilidade dos documentos. A compreensibilidade de um documento clínico envolve exatidão, consistência e legibilidade (PANTAZOS; LAUESEN; LIPPERT, 2017). A exatidão exige que a ordem dos textos não fique comprometida e que a informação disponibilizada seja idêntica ao documento original. Já a consistência exige que a substituição de dados sensíveis seja por pseudônimos de mesmo tipo. Por fim, a legibilidade exige que não haja comprometimento de interpretação do documento após a desidentificação. A tarefa, no entanto, não é trivial. Não é recomendado realizar substituições de dados sensíveis de forma aleatória, pois pode-se perder informações relevantes.

Ao alterar uma data, por exemplo, é importante pensar na estação do ano, para não perder a informação de uma doença sazonal (LI; QIN, 2017). Outro aspecto é a correlação entre data de nascimento de um paciente e sua idade. Ambas devem ser correspondidas, pois a idade é importante no diagnóstico de um paciente. A substituição da localidade de um paciente precisa considerar os aspectos sociais e étnicos da região, pois isso pode afetar também no diagnóstico de um paciente. O mesmo raciocínio vale para gênero e raça de um paciente.

Deve haver o cuidado para que a substituição do nome de um paciente não seja por outro semelhante ao seu. Também há a dificuldade de diferenciar alguns nomes de pessoas com nomes de doenças, visto que muitas doenças são nomeadas considerando o nome do profissional que a descobriu ou de um indivíduo que a contraiu. Ao substituir um nome de paciente por outro, deve-se tomar o cuidado de trocar todas as ocorrências pelo mesmo nome. Assim preserva-se a longitudinalidade dos documentos.

Além disso, há a dificuldade de dissociar pacientes de nomes iguais mas que não são a mesma pessoa. Um algoritmo poderia tratar ambas as pessoas como sendo a mesma e criando um registro médico não condizente com a realidade.

4.3 Reconhecimento de entidade mencionada

REM, é um subcampo da linguística computacional, uma técnica de tratamento de informação que detecta e extrai entidades mencionadas num texto. São consideradas entidades mencionadas quaisquer coisas que podem ser tratadas com um nome próprio (JURAFSKY; MARTIN, 2008). São geralmente considerados REM: pessoas, localizações, organizações, entidades geopolíticas, instalações e veículos. No entanto, para aplicações mais especializadas, o conceito de REM pode se estender. No contexto de saúde, por exemplo, o REM pode se estender para outras entidades como nome de doenças, medicamentos, instrumentos cirúrgicos etc. Além disso, o conceito de REM pode ser facilmente ampliado para trechos de textos que não são entidades, mas que possuem relevância prática para a resolução de algum problema dentro de um determinado contexto. Datas, idades, número de documentos, endereços de e-mail, telefones e outras expressões, geralmente numéricas, são exemplos de extensões aplicáveis ao REM.

Uma abordagem padrão para a tarefa de REM é a marcação sequencial de trechos de interesse num texto. A abordagem mais comum de marcação é denominada de IOB, ou *Inside-Outside-Beginning*, sendo uma análise, palavra a palavra, de fragmentos de texto para marcar os trechos de começo e fim de um REM. Além disso, trechos que estão fora dos limites de um REM são marcados como palavras sem interesse para aplicação (JURAFSKY; MARTIN, 2008). O método IOB pode ser implementado tanto de forma manual, quanto de forma automatizada.

O treinamento de sistemas de REM é realizado com uma série de características que podem ser usadas para identificar entidades de um texto. O uso ou não dessas características depende do método de aprendizado utilizado e da escolha pessoal do profissional da informação. Pode-se usar uma combinação dessas características, apenas uma delas ou nenhuma. A Tabela 7 mostra quais são as principais características identificadas na literatura.

Tabela 7 – Características de treinamentos de sistemas de REM

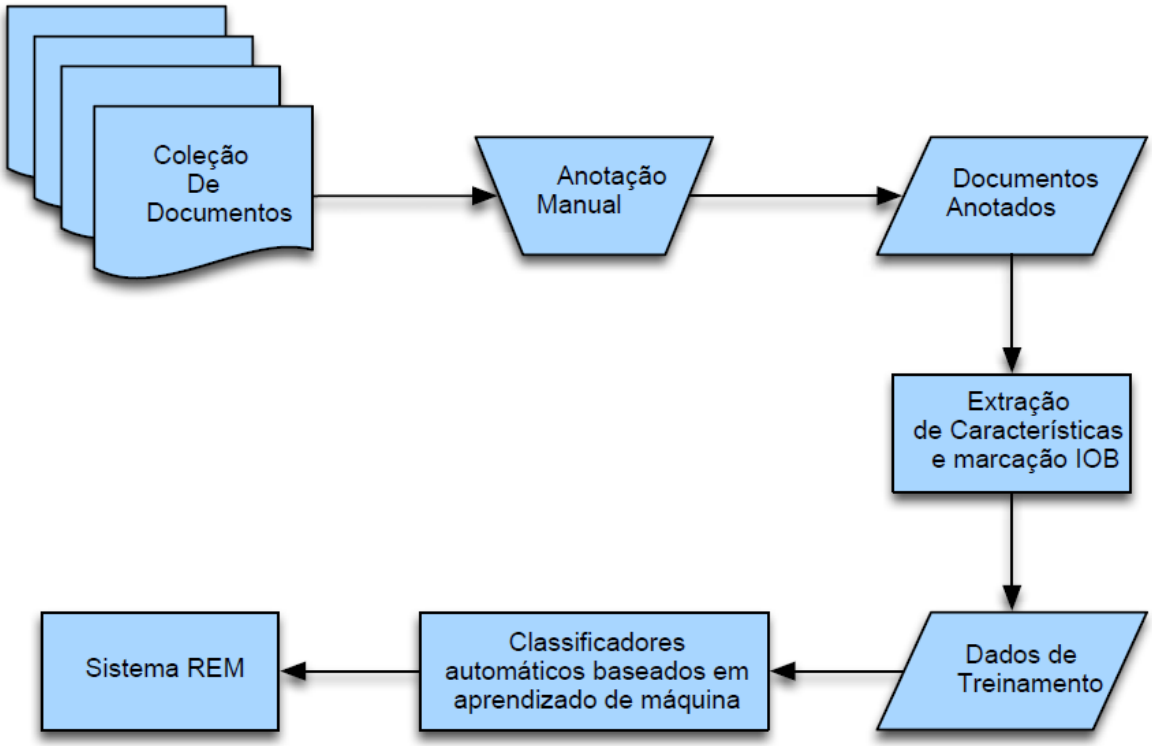
Característica	**Descrição**
Itens lexicais	O *token* será marcado. Pode ser uma palavra ou conjunto de palavras
Itens lexicais stemizados[2]	Versão stemizada do *token* marcado
Capitalização	Padrão ortográfico da palavra. Pode conter caixa alta no começo, meio ou fim d palavra
Afixos	Um elemento que é adicionado a um radical de forma que altera o sentido básico d palavra
Classe gramatical	Definição da classe gramatical do *token* (substantivo, verbo etc.)
Etiquetas de blocos sintáticos	Marcação básica da sintaxe de um bloco
Listas	Nomes pessoais, dicionários, dicionários geográficos etc
Tokens de predição	Alguns *tokens* podem predizer qual será a próxima palavra ou tipo de palavra qu virá a seguir
Bag of Words/N-Grams	Palavras ou conjunto de palavras que aparecem ao redor dos *tokens*

Fonte: Adaptado de Jurafsky e Martin (2008)

As características mostradas na Tabela 7 são, na prática, armazenadas numa matriz onde cada linha representa um *token* a ser classificado. As colunas dessa matriz são as características mostradas na Tabela 7 com exceção da última coluna que é a classificação referente a esse *token*. Essa tabela é transformada em variáveis numéricas e usada em algoritmos de aprendizado de máquina para fazer o treinamento de classificação de *tokens*. Depois que um sistema de REM é treinado, ele está pronto para classificar automaticamente os *tokens*. A Figura 2 ilustra o conceito básico de um sistema de REM.

[2] Stemização é o processo de redução de palavras para sua raiz. Dessa maneira, palavras com raízes iguais são identificadas como sendo similares. Ex: “pedra” e “pedregulho” compartilham a mesma raiz “pedr”.

Figura 2 – Conceito básico de um sistema de REM.



Fonte: Adaptado de Jurafsky e Martin (2008).

Pode-se dizer que a desidentificação nada mais é do que uma tarefa de REM com objetivos ligeiramente distintos: o REM identifica e classifica as entidades enquanto a desidentificação vai além, removendo-as ou criptografando-as. Além disso, a desidentificação de documentos clínicos sofre com problemas de ambiguidades e com palavras inidentificáveis como acrônimos, erros de digitação e estrangeirismos. A palavra Chagas, por exemplo, pode ser uma doença ou um nome próprio. Para desidentificação de documentos clínicos o nome próprio deve ser removido, mas o nome da doença não. Uma solução para evitar essas ambiguidades é utilizar ontologias biomédicas que fornecem vocabulários controlados para a área da saúde. Elas oferecem um bom complemento para as características apresentadas na Tabela 7 (JANOWICZ; KESSLER, 2008).

4.4 Métodos de desidentificação

Na literatura é possível encontrar uma série de métodos de desidentificação de textos. Fazendo uma análise das revisões de literatura dos últimos anos, realizada pelos autores Uzuner, Luo e Szolovits (2007), Meystre et al. (2010), Yogarajan, Mayo e Pfahringer (2018), podemos dividir esses métodos em três categorias principais: (1) métodos baseados em regras; (2) métodos de aprendizado de máquina e (3) métodos híbridos (DALIANIS, 2018).

Os métodos baseados em regras extraem as características de uma categoria de dados sensíveis e constroem algoritmos para identificar e tratar esses dados. A identificação de regras utiliza recursos léxicos, dicionário de palavras, caracteres especiais, números, morfologia, parte de discurso etc. para construir o algoritmo de desidentificação. É uma área com bastante afinidade com a pesquisa em linguística computacional (NASRABADI, 2007). As principais vantagens dessa técnica são que ela possui alto índice de precisão quando as regras são definidas corretamente, e não requer, na maioria das vezes, uma quantidade razoável de dados para treinamento de algoritmos. Em contrapartida, o aumento da complexidade de padrões exige cada vez mais *expertise* de profissionais envolvidos com a desidentificação de dados nos documentos.

Os métodos de aprendizado de máquina utilizam uma base com dados potenciais para serem desidentificados. Como o campo de estudo em inteligência artificial é muito amplo, diferentes técnicas são aplicadas e testadas na literatura. Essas técnicas analisam documentos e geram regras automáticas baseadas em aprendizado para identificar dados sensíveis. A vantagem desses métodos é que eles conseguem generalizar regras complexas de serem definidas, como nomes próprios, por exemplo. No entanto, para que os algoritmos aprendam eficientemente, são necessárias grandes quantidades de dados para treinamento (MICHALSKI; CARBONELL; MITCHELL, 2013).

Os métodos de aprendizado de máquina e baseados em regras podem ser combinados. Dá-se a essa abordagem o nome de métodos híbridos. Isso ocorre com frequência pois, geralmente, é preferível implementar regras em dados de fácil desidentificação devido à alta precisão desse método. Os métodos combinados estão presentes numa série de aplicações desenvolvidas para esse propósito. Os principais documentos que têm sido utilizados nas tarefas de desidentificação são os **sumários de alta** e **relatórios de patologia**. A justificativa é que esses documentos possuem estrutura formal, contém poucas abreviações técnicas, erros de digitação e outros problemas comuns em outros documentos clínicos (MEYSTRE et al., 2010).

A análise de *performance* desses algoritmos é medida pelo cálculo de F-Score. Também chamado *F-Measure*, é a principal medida de acurácia usada em recuperação de informação para medir buscas, identificação de padrões em textos, classificação de documentos, entre outros (SASAKI et al., 2007). O F-Score foi citado pela primeira vez por Chinchor (1992) numa conferência em que o objetivo era discutir métricas de avaliação para PLN.

O F-Score é a média harmônica entre precisão e revocação. A idealização do F-Score veio como uma forma de unificar tanto a precisão quanto a revocação numa única unidade de medida (CHINCHOR; SUNDHEIM, 1993).

A precisão é uma técnica de avaliação que mede a quantidade de dados relevantes encontrados entre as instâncias recuperadas. É calculada, segundo Baeza-Yates e Ribeiro-

Neto (2011), por:

$$Precisão = \frac{\text{Positivos verdadeiros}}{\text{Positivos verdadeiros} + \text{Positivos falsos}} \tag{4.3}$$

Em outras palavras, é a capacidade de desidentificar corretamente os dados. Já a revocação é a técnica de avaliação que mede a quantidade de dados relevantes encontrados dentre todos os possíveis dados relevantes passíveis de recuperação. É calculada, ainda segundo Baeza-Yates e Ribeiro-Neto (2011), por:

$$Revocação = \frac{\text{Positivos verdadeiros}}{\text{Positivos verdadeiros} + \text{Falsos negativos}} \tag{4.4}$$

Em outras palavras, é a capacidade de identificar os dados que precisam realmente ser desidentificados. E, por fim, o F-Score é calculado, segundo Sasaki et al. (2007), por:

$$F - Score = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4.5}$$

O que torna o F-Score uma medida relevante na área da informação é a sua fórmula flexível que envolve tanto precisão quanto revocação. Dessa forma, é possível encontrar o uso do F-Score na literatura em diferentes aplicações (CHINCHOR; SUNDHEIM, 1993).

Para os métodos de desidentificação, apresentar uma boa revocação é mais importante que uma precisão, visto que é preferível identificar todos os dados que necessitam ser removidos (DALIANIS, 2018). Embora haja a discussão de que a perfeição do processo de desidentificação não possa ser alcançada, há o consenso de que um F-Score de 95% é considerado como valor aceitável (STUBBS; KOTFILA; UZUNER, 2015; STUBBS; FILANNINO; UZUNER, 2017).

Segundo as revisões de literatura de Uzuner, Luo e Szolovits (2007), Meystre et al. (2010) e Yogarajan, Mayo e Pfahringer (2018) os métodos desenvolvidos apresentaram um F-Score entre 59% e 99%.

4.4.1 Métodos de aprendizado de máquina

Aprendizado de máquina possui duas definições clássicas. Segundo Samuel (1959) é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados para isso. Para Mitchell (1999), é a capacidade de um programa melhorar seu desempenho após passar por um processo de treinamento.

Ambas definições se completam. Elas revelam que os computadores são capazes de aprender e melhorar a sua *performance*. Um algoritmo de aprendizado de máquina é treinado usando uma base de dados. Esse treinamento permite que padrões de dados sejam analisados e uma função matemática seja criada para a representação desses padrões.

Essa função é utilizada para analisar conjuntos de dados e definir, com certo grau de acurácia, os padrões identificados. A vantagem de algoritmos de aprendizado de máquina é o aprendizado automático de padrões complexos de desidentificação. A desvantagem, como já citado anteriormente, é a necessidade de grandes quantidades de dados para treinar esses algoritmos.

Os métodos mais comuns de aprendizado de máquina, encontrados na literatura de desidentificação, são os baseados na cadeia de Markov, em especial os campos aleatórios condicionais, máquina de vetores de suporte, árvores de decisão e redes neurais recorrentes. Esses métodos e trabalhos correlatos são explicados nas próximas seções.

4.4.1.1 Métodos baseados em cadeias de Markov

Uma cadeia de Markov é um processo estocástico, ou seja, um modelo matemático que evolui temporalmente de maneira probabilística (KEMENY; SNELL, 1976). As cadeias de Markov são representadas por um autômato finito de estados em que os conjuntos de transições entre os estados são medidos por probabilidades de um evento acontecer dada uma observação. Essas probabilidades mensuram qual é o provável comportamento que o modelo matemático terá numa situação específica. Em outras palavras, as cadeias de Markov calculam qual é a probabilidade do próximo evento ocorrer, dado o último evento ocorrido.

Para os problemas de PLN, mais especificamente a desidentificação, as cadeias de Markov funcionam analisando as palavras anteriores do texto, para prever qual será a próxima palavra. Por exemplo: dado que a última palavra lida foi “Sr.” qual é a probabilidade da próxima palavra ser um nome próprio?

As técnicas mais usadas para a desidentificação de textos são o modelo oculto de Markov, máxima entropia e os campos aleatórios condicionais. O modelo oculto de Markov é composto por classificadores discriminativos que analisam os *tokens* em sequência para identificar as suas características e marcá-los com a categoria correspondente. Já a máxima entropia e os campos aleatórios condicionais são classificadores gerativos que consideram o contexto onde os *tokens* estão inseridos. Os modelos são descritos com mais detalhes nas próximas subseções.

4.4.1.1.1 Modelo oculto de Markov

O modelo oculto de Markov, também conhecido como *Hidden Markov Model* ou HMM, é um método estatístico bastante usado no processamento de texto e fala, uma das técnicas mais importantes no aprendizado de máquina para esse fim (JURAFSKY; MARTIN, 2008). O modelo oculto é útil quando não é possível observar todas as probabilidades que interferem na escolha dos estados. Em PLN, as probabilidades ocultas são as PoS que

as palavras carregam, tais como substantivo, verbo etc., ou seja, características adicionais inerentes que cada palavra carrega e que têm seu peso na hora do aprendizado. De acordo com Rabiner (1989), as cadeias de Markov atacam três problemas fundamentais:

1. **Probabilidade:** dada uma cadeia de Markov, qual é a probabilidade de ocorrer uma sequência de eventos?
2. **Decodificação:** dada uma sequência de eventos, qual é o melhor estado oculto que reflete essa tomada de decisão?
3. **Aprendizado:** dada uma sequência de eventos e uma cadeia de Markov, aprenda quais decisões foram realizadas pelo modelo.

Para desidentificação, o método usado é o de aprendizado em que a base de treinamento possui a sequência de eventos pré-definida. Basta o modelo de Markov aprender a tomada de decisão e replicá-la na base de testes. Wellner et al. (2007) realizaram o processo de desidentificação usando o LingPipe[3], uma ferramenta de PLN baseada em HMM. O resultado encontrado pelos autores foi de 96,64% de F-Score. Chen, Cullen e Godwin (2015) usam o processo de Dirichlet[4] para modelar os estados ocultos da cadeia de Markov e classificar os *tokens* dos documentos. Os autores alcançaram um F-Score de 90,9% nesse método que é puramente baseado em cadeias de Markov.

4.4.1.1.2 Máxima entropia

Máxima entropia, ou MaxEnt, é um classificador que funciona extraindo características de *tokens* e combinando-as linearmente por multiplicação e soma de probabilidades (JURAFSKY; MARTIN, 2008). A MaxEnt realiza classificações analisando todas as probabilidades de um *token* e escolhendo a que possui a maior probabilidade. O método, no entanto, não é um classificador sequencial. Ele apenas analisa uma única observação e a classifica.

Para a tarefa de desidentificação, ser um classificador sequencial importa, pois a análise de contexto das palavras precedentes pode dar informações preciosas sobre o significado de um *token*. Para isso, usa-se a variação da Máxima Entropia chamada *Maximum-Entropy Markov Model* ou MEMM. Essa técnica combina a análise de MaxEnt com a modelagem de HMM. Dessa maneira, o MEMM consegue calcular, analisar o *token* atual, extrair as características e calculá-las na análise do estado anterior da máquina de estados.

[3] Disponível em http://alias-i.com/. Acessado em março de 2022

[4] Dirichlet é um processo de distribuição estocástico usado para fazer inferências baseadas em eventos antecedentes de uma mesma observação.

Taira, Bui e Kangarloo (2002) desenvolveram um algoritmo somente para a desidentificação de nomes em inglês. Esse algoritmo usa a MaxEnt para calcular a probabilidade de uma palavra ser um nome próprio. O método proposto pelos autores analisa o léxico e as restrições semânticas do texto para atribuir as probabilidades dos *tokens* com base no contexto. O léxico contém mais de 64 mil nomes e sobrenomes, e as restrições semânticas são compostas por situações como pronomes de tratamento que enquadrariam um *token* como sendo um nome. O algoritmo dos autores alcançou 97% de F-Score.

4.4.1.1.3 Campos aleatórios condicionais

Também conhecido como *Conditional Random Field* ou CRF, essa é uma técnica de aprendizado que usa o contexto para fazer predições. Portanto é indicada para PLN, dado que o significado das palavras também é baseado no contexto em que estão inseridas. O CRF funciona definindo probabilidades condicionais para cada *token* lido previamente numa sequência e associando-o ao *token* atual. Isso permite que, a cada *token* lido, inferências sejam feitas baseadas nos aprendizados anteriores. Em tarefas de desidentificação, o CRF identifica estruturas padrões de textos que acompanham os dados sensíveis e aprende a identificá-los baseados no contexto ao qual estão inseridos.

O CRF se mostra uma técnica bastante popular para desidentificação devido ao número de trabalhos encontrados na literatura. Alguns autores, como Aramaki et al. (2006) e Wellner et al. (2007), afirmam que o CRF possui desempenho superior ao de outras técnicas de aprendizado de máquina como o HMM e as máquinas de vetores de suporte. O sistema criado por Aramaki et al. (2006) usa o CRF para extrair dados sensíveis do *corpus* i2b2 e obteve 97,47% de F-Score. Já o sistema apresentado por Wellner et al. (2007), trabalha de forma híbrida, construindo regras para campos mais padronizados. Em campos mais complexos, o CRF foi usado. Essa combinação obteve um resultado de 97,36% de F-Score, ligeiramente melhor que seu experimento usando HMM.

Apesar de ser popular, o CRF raramente é usado sozinho. Na maioria das vezes, é combinado com soluções híbridas usando regras para desidentificar os campos mais comuns. Autores, como Lee et al. (2017), fizeram um experimento comparando a abordagem puramente baseada em CRF com a híbrida, realizarando dois testes. O primeiro usou os recursos do software CLAMP Toolkit[5], um conjunto de ferramentas para tarefas de PLN. O segundo teste adicionou regras para gerar mais especificidade ao *corpus* de documentos clínicos. O melhor resultado encontrado pelos autores foi do sistema híbrido, obtendo 90,74% de F-Score, enquanto o puramente baseado em CRF obteve 74,5%. Podem ser citados os trabalhos de Gardner e Xiong (2008), He et al. (2015), Yadav et al. (2017) e Dehghan et al. (2017), como sistemas de desidentificação puramente baseados em CRF.

[5] Disponível em https://clamp.uth.edu/. Acessado em março de 2022.

Gardner e Xiong (2008) analisaram características como palavras precedentes e subsequentes do *token* além de verificar se um *token* está capitalizado ou se é um número. Por meio dessas análises um *corpus* de treinamento foi construído. O treinamento do CRF é iterativo, ou seja, um mesmo *token* é classificado várias vezes durante o processo. O resultado encontrado pelos autores foi de 98,32% de F-Score.

Já He et al. (2015) e Yadav et al. (2017) usaram regras somente no pré-processamento de *tokens*. Para treinamento e análise apenas o CRF foi usado e a *performance* obtida foi de 92,32% e 97,47% de F-Score respectivamente.

Por fim, o trabalho de Dehghan et al. (2017) realizou testes usando dois sistemas diferentes: o mDEID[6] e o CliDEID[7]. Ambos os sistemas são desidentificadores de textos médicos baseados em CRF. Os autores realizaram três testes. O primeiro usando o CliDEID, o segundo o mDEID e o terceiro combinando os resultados dos dois primeiros testes. Os resultados encontrados pelos autores foram uma média dos três testes, correspondendo a aproximadamente 90% de F-Score.

Os métodos híbridos apresentados pelos autores a seguir usam regras para identificar dados sensíveis mais padronizados. As regras são construídas usando expressões regulares para campos mais simples como telefones, datas, IDs, e-mails entre outros, ou uso de dicionários e listas de nomes para identificação de médicos, pacientes etc.

O método híbrido de Yang e Garibaldi (2015) criou características de identificação para extrair dados sensíveis de documentos clínicos. As características que possuíam dados suficientes para treinamento foram usadas com CRF para aprendizado e classificação. As demais características foram identificadas usando dicionários e expressões regulares. O resultado alcançado pelos autores foi um F-Score de 93,6%.

Liu et al. (2015) usaram o método híbrido com CRF em duas frentes distintas: a primeira no nível de *tokens* e a segunda no nível de caracteres. Os autores argumentam que o aprendizado no nível de *token* pode gerar erros de limites entre sentenças, prejudicando a análise de contexto. Para isso, foram separadas as sentenças e essas, por fim, foram analisadas no nível de *tokens*. O resultado obtido no estudo foi um F-Score de 91,2%.

Dehghan et al. (2015) usaram regras e dicionários para identificar todas as categorias da HIPAA. O CRF foi usado como um sistema de apoio na tomada de decisão para decidir se as regras identificaram os dados sensíveis corretamente. Os autores fizeram três submissões de avaliação. A primeira sem o uso do CRF, a segunda com o CRF e a terceira é uma tentativa de melhoria das duas submissões anteriores. O melhor resultado obtido pelos autores foi um F-Score de 91% usando-se o método híbrido. De maneira similar, Phuong e Chau (2016) usaram regras no pré e pós-processamento, quando os dados sensíveis são

[6] Disponível em: https://svn-us.apache.org/repos/asf/ctakes/sandbox/ctakes-clinical-deid/. Acessado em março de 2022.

[7] Disponível em https://github.com/kovacevica/CliDEID. Acessado em março de 2022.

identificados. Essas regras foram usadas como fonte para o treinamento de CRF, otimizado para identificar dados que não foram detectados anteriormente. Os autores testaram a sua abordagem em cinco *corpora* diferentes, conseguindo um F-Score médio de 91,32%.

O método híbrido, elaborado por Bui, Wyatt e Cimino (2017) baseia-se no sistema de peneiras, no qual o processamento vai passando por diferentes filtros até chegar ao resultado. As peneiras são estruturadas na seguinte ordem: detecção de padrão, identificação através de dicionários e aprendizado de máquina usando o CRF. Os autores obtiveram um F-Score de 93%.

4.4.1.2 Máquina de vetores de suporte

Também conhecido por SVM, do inglês *Support-vector Machine*, é um método de aprendizado supervisionado usado para classificação. O SVM funciona traçando, num hiperplano, as entidades classificadas e depois ajustando uma função polinomial de forma que essas entidades estejam separadas em dois grupos distintos. Dessa maneira, o algoritmo usa esta função para classificar uma nova entidade verificando em qual lado do polinômio ela está situada (MÜLLER; GUIDO et al., 2016).

Por ser um classificador binário, o SVM deve ser usado na desidentificação como um algoritmo genérico, classificando os *tokens* como dados sensíveis ou não. Outra forma é dividir a tarefa de classificação para categorias diferentes da HIPAA. Isso permite a criação de sistemas mais especializados para cada tipo de dado sensível enquanto aumenta a complexidade de implementação devido à quantidade de elementos que compõem a lista da HIPAA. Todos os trabalhos encontrados na literatura usam métodos híbridos combinados com o SVM. As metodologias desenvolvidas seguem um padrão de usar as regras em categorias de dados mais simples como datas, idades, telefones e IDs.

Guo et al. (2006) usaram o SVM para fazer classificações de IOB dos sumários de alta e identificá-los como dados sensíveis. O sistema foi dividido em categorias de classificação conforme os tipos de dados sensíveis como nome, idade, localização etc. Também foram adicionados prefixos e sufixos como características para ajudar no reconhecimento de nomes. O resultado foi uma média ponderada da *performance* de cada sistema, obtendo um F-Score de 98,69%.

Hara et al. (2006) usaram as regras, com o apoio do aprendizado de máquina, com características como interpretação de cabeçalho de seções, categorias de frases, PoS, versão stemizada do *token* e ortografia para identificar nomes de hospitais, pacientes, médicos, localidades e idades. As informações combinadas foram usadas para desidentificação de documentos clínicos. A *performance* do trabalho foi de um F-Score aproximado de 92%.

4.4.1.3 Árvores de decisão

Árvores de decisão é uma ferramenta de suporte à tomada de decisão que se destaca de outros métodos de aprendizado de máquina pela fácil implementação e compreensão do aprendizado (MÜLLER; GUIDO et al., 2016). Trata-se de uma estrutura hierárquica de dados de forma que, o dado mais ao topo da estrutura, também denominado raiz da árvore, é o ponto de partida para consultas e navegações. As árvores de decisão são criadas por análise e aprendizado de uma base de dados de treinamento. O algoritmo cria uma série de estruturas condicionais que servirão como meios de navegação da árvore.

Por exemplo, ao analisar uma base de treinamento, o algoritmo de criação de árvores de decisão detectou que 80% dos tokens cuja categoria é "telefone" possui somente caracteres numéricos. Dessa maneira, um ramo de decisão é construído onde é testado se o *token* possui somente caracteres numéricos. Em caso positivo, ele é classificado como "telefone". Caso contrário, a árvore pula para outro ramo de decisão em que outra condição é analisada.

Szarvas, Farkas e Busa-Fekete (2007) usam árvores de decisão para classificar dados sensíveis. Cada ramo da árvore, num total de três, usa um algoritmo de aprendizado diferente e é treinado usando dicionários, análise léxica e outras características locais do texto como ortografia, tamanho do *token*, etc. Os autores conseguiram uma *performance* de 99,75% de F-Score.

4.4.1.4 Redes neurais recorrentes

As redes neurais recorrentes, do inglês *recurrent neural networks*, ou RNNs, são modelos especializados em dados sequenciais, assim como o CRF. O algoritmo é uma técnica de aprendizado profundo, subárea do aprendizado de máquina. Ele funciona lendo uma sequência de *tokens* e produzindo um vetor de tamanho fixado que representa uma generalização dessa sequência (GOLDBERG, 2017). A generalização possui diferentes significados, dependendo do objetivo do algoritmo. Para desidentificação, por exemplo, a RNN deseja realizar classificações gramaticais e identificar quais *tokens* são dados sensíveis ou não.

As RNNs, no entanto, são raramente usadas sozinhas. O grande destaque para essa técnica está no treinamento de bases, que serão usadas para a alimentação de outros sistemas (GOLDBERG, 2017). A técnica mais comum de RNNs para desidentificação é o *Long Short-Term Memory* ou LSTM. Na literatura, é muito comum o uso combinado de LSTM com CRF, chamado CRF-LSTM.

A adoção de RNN como abordagem para resolver problemas de desidentificação é uma prática recente. Autores, como Goldberg (2017), afirmam que as redes neurais e o aprendizado profundo possuem resultados superiores ao do popular CRF. Para sustentar

esta afirmativa, autores como Jiang et al. (2017), Dernoncourt et al. (2017), Richter-Pechanski et al. (2019) e Trienes et al. (2020) realizaram comparações com diferentes abordagens para testar o desempenho das redes neurais para desidentificação de documentos clínicos.

Jiang et al. (2017), Dernoncourt et al. (2017), Richter-Pechanski et al. (2019) e Trienes et al. (2020) compararam o CRF diretamente com o LSTM. Jiang et al. (2017) preprocessaram e classificaram os documentos com blocos sintáticos. Os autores obtiveram 89,86% de F-Score para o sistema com o LSTM e 88,02% com o CRF. Richter-Pechanski et al. (2019) compararam documentos clínicos alemães. Os autores encontraram uma ligeira superioridade do LSTM, com F-Score de 96%, em relação ao CRF, com 93%. Dernoncourt et al. (2017) foram além e compararam também a *performance* do CRF-LSTM. Para isso, usaram dois *corpora* diferentes para comparar os algoritmos. Para o primeiro *corpus*, o melhor resultado foi do CRF-LSTM com F-Score de 97,83%. Para o segundo *corpus*, o método que obteve a melhor *performance* foi o LSTM com F-Score de 99,22%. Em ambos os *corpora*, o CRF foi o algoritmo que obteve pior desempenho. Trienes et al. (2020) também compararam o CRF com o CRF-LSTM e foram além adicionando um método puramente baseado em regras na comparação. Para os testes eles usaram *corpora* em inglês e holandês. O sistema com CRF-LSTM, obteve 88,13% de F-Score médio, seguido do CRF, com 83,03% de F-Score médio, e do baseado em regras, com 66,4% de F-Score. Indo um pouco na contramão de outros estudos, Saluja et al. (2019) compararam o SVM com o CRF-LSTM. Os resultados encontrados pelos autores foi um F-Score de 86,1% para o SVM e 93,5% para o CRF-LSTM.

Além dos trabalhos comparativos citados acima, há os trabalhos que desenvolveram técnicas puramente baseadas em redes neurais. Lee et al. (2016) usam o LSTM num sistema de três camadas distintas: (1) vetorizador de palavras; (2) preditor de categorias e (3) otimizador de categorias. Estas camadas são usadas para detectar e desidentificar dados sensíveis de documentos clínicos. Os autores conseguiram 99,2% de F-Score. Dernoncourt, Lee e Szolovits (2017) apresentam um software chamado NeuroNER[8]. Os autores argumentam que o NeuroNER é uma alternativa viável para iniciantes que desejam usar redes neurais para realizar desidentificação de dados. O software acompanha um sistema de anotação usado para marcar dados sensíveis. A *performance* máxima do NeuroNER demonstrada pelos autores é de F-Score de 97,7%.

Como trabalhos híbridos, destacam-se o de Liu et al. (2017) e Zhao et al. (2018). Liu et al. (2017) utilizaram uma abordagem híbrida que mistura CRF-LSTM, SVM e regras. Os autores fizeram a extração de características usando o CRF-LSTM. Em seguida foram feitas representações matemáticas das palavras usando-se redes neurais para que os algoritmos de aprendizado pudessem processar o material. As características e

[8] Disponível em http://neuroner.com/. Acessado em março de 2022.

representações matemáticas foram combinadas usando o SVM. Os resultados obtidos pelo SVM e pela extração baseada em regras, que atuou independentemente, constituíram o resultado deste trabalho, um F-Score de 91,43%. Zhao et al. (2018) usaram o CRF-LSTM combinado como uma técnica chamada *text skeleton*, que extrai frases, retendo apenas as palavras mais comuns. Essas frases constituem-se de base de treinamento para o CRF-LSTM e se mostram úteis para análise de contexto em diferentes perspectivas. O resultado encontrado pelos autores foi um F-Score médio de 94,34%.

O LSTM não é a única alternativa de redes neurais para desidentificação de documentos, apesar de ser a mais popular. Srivastava et al. (2016) usam duas variantes de redes neurais conectadas, diferentes do LSTM: Elman e Jordan. As redes neurais de Elman usam três camadas de dados conectadas por uma camada oculta que é responsável por armazenar valores propagados durante a execução do algoritmo e ajustar o aprendizado da rede. As redes neurais de Jordan são similares às de Elman, com a diferença de que a camada oculta é substituída por uma camada externa que pode ser alimentada via outros métodos. Os autores identificaram que ambas as técnicas foram superiores ao CRF sendo as redes neurais de Jordan a que melhor se adaptou à tarefa de desidentificação com o F-Score de 93,84%. Já Yadav et al. (2017) usaram o conceito de *transfer learning*, técnica que aproveita o aprendizado de um sistema para resolver outros problemas. O *transfer learning* foi usado para enriquecer o aprendizado do processo de desidentificação sendo aplicado sobre uma rede neural e a desidentificação ocorre em dois *corpora* distintos. O primeiro obteve um resultado de 98% de F-Score e o segundo, 96%.

4.4.2 Métodos baseados em regras

Os métodos baseados em regras são condições construídas manualmente para serem comparadas durante o processo de análise de desidentificação. Todo *token* é analisado com regras criadas e, caso corresponda a alguma delas, é marcado naquela categoria à qual a regra pertence. As regras mais comumente criadas são o uso de expressões regulares e estruturas condicionais de se-senão. Expressões regulares são escritas em linguagem formal para identificar uma cadeia específica de caracteres, podendo ser números, letras ou símbolos (CHAPMAN; WANG; STOLEE, 2017). Seu uso em desidentificação é majoritariamente para identificar dados sensíveis que possuem formas padronizadas, ou seja, é possível construir uma regra para identificação de uma sequência única de caracteres. As estruturas condicionadas são lógicas de programação muito usadas na computação. Por meio delas, é possível programar o comportamento das regras usando perguntas simples com respostas de verdadeiro ou falso. O uso de estruturas condicionadas é comum quando as expressões regulares não conseguem obter êxito. As condições são úteis, pois elas conseguem abranger diferentes *tokens* e estruturas simultaneamente, criando regras mais complexas. No entanto, ao aumentar a complexidade das regras, perde-se também a sua eficiência. A construção

de regras complexas deixa a detecção de dados sensíveis muito específica para o problema, tornando-a ineficiente em *corpora* variáveis. Há também a dificuldade de construí-las, devido ao número de condições necessárias para funcionar.

Os algoritmos baseados em regras podem ser construídos usando diferentes raciocínios. Sweeney (1996) criou regras paralelas para identificar categorias de dados sensíveis separadamente. A autora usa detectores baseados num modelo de prioridades. Um detector de localidade, por exemplo, engloba regras de detecção de cidades, estados e países e possui alta prioridade na identificação dessas categorias. Isso é usado em conjunto com outros detectores. Cada detector atribui um valor de certeza para um *token* em relação ao dado sensível que está sendo analisado. Os detectores com maior prioridade possuem o resultado na hora de avaliação. Guillen et al. (2006) usaram a posição do texto para identificação de dados sensíveis. Segundo os autores, o início e o final do texto possuem mais indícios de dados como nomes de hospitais, médicos, IDs e datas. Essa informação foi usada para construir regras de identificação que se concentram nessas partes do documento para fazer a desidentificação. Beckwith et al. (2006) desenvolveram a ferramenta HMS Scrubber[9], um depurador de dados sensíveis baseados em regras. O HMS Scrubber funciona em três etapas. A primeira busca por identificadores conhecidos por carregar dados sensíveis como nomes, número do cartão de saúde etc. A segunda usa um total de cinquenta expressões regulares para detectar dados como datas, número de acessos, endereços etc. Na terceira etapa, os dados identificados nas etapas anteriores são comparados a dicionários e listas para identificar nomes próprios e localizações. O algoritmo de Friedlin e McDonald (2008) usa ao todo cinquenta expressões regulares para identificar padrões numéricos. Algumas regras tentam identificar pronomes de tratamento para detectar nomes de pessoas. Quando isso não é possível, uma lista de nomes e análises de cabeçalhos de documentos são usadas para complementar as regras criadas.

O uso de dicionários e listas de nomes também é muito comum em métodos baseados em regras. As listas são usadas para fazer comparação simples de *tokens* com nomes preexistentes. Se um *token* corresponde a algum nome da lista, é possível afirmar que ele pertence àquele grupo específico. Seu uso é principalmente para identificar nomes próprios, de localidades e empresas. Thomas et al. (2002) usaram dicionários, de diversas fontes, para identificação de nomes próprios com o auxílio de expressões regulares que identificam pronomes de tratamento que precedem os nomes. Fielstein, Brown e Speroff (2004) criaram um algoritmo de detecção de categorias usando fontes públicas de endereços de e-mail, nomes, localizações, entre outros, baseadas nas listas fornecidas pelo *United States Department of Veterans Affairs*[10], instituição que fornece serviços de saúde para veteranos e outros pacientes dos Estados Unidos. Além disso, foram criadas expressões

[9] Disponível em https://sourceforge.net/projects/spin-chirps/files/hms-scrubber/hms-scrubber-2.0/. Acessado em março de 2022.

[10] Disponível em https://www.va.gov/. Acessado em março de 2022.

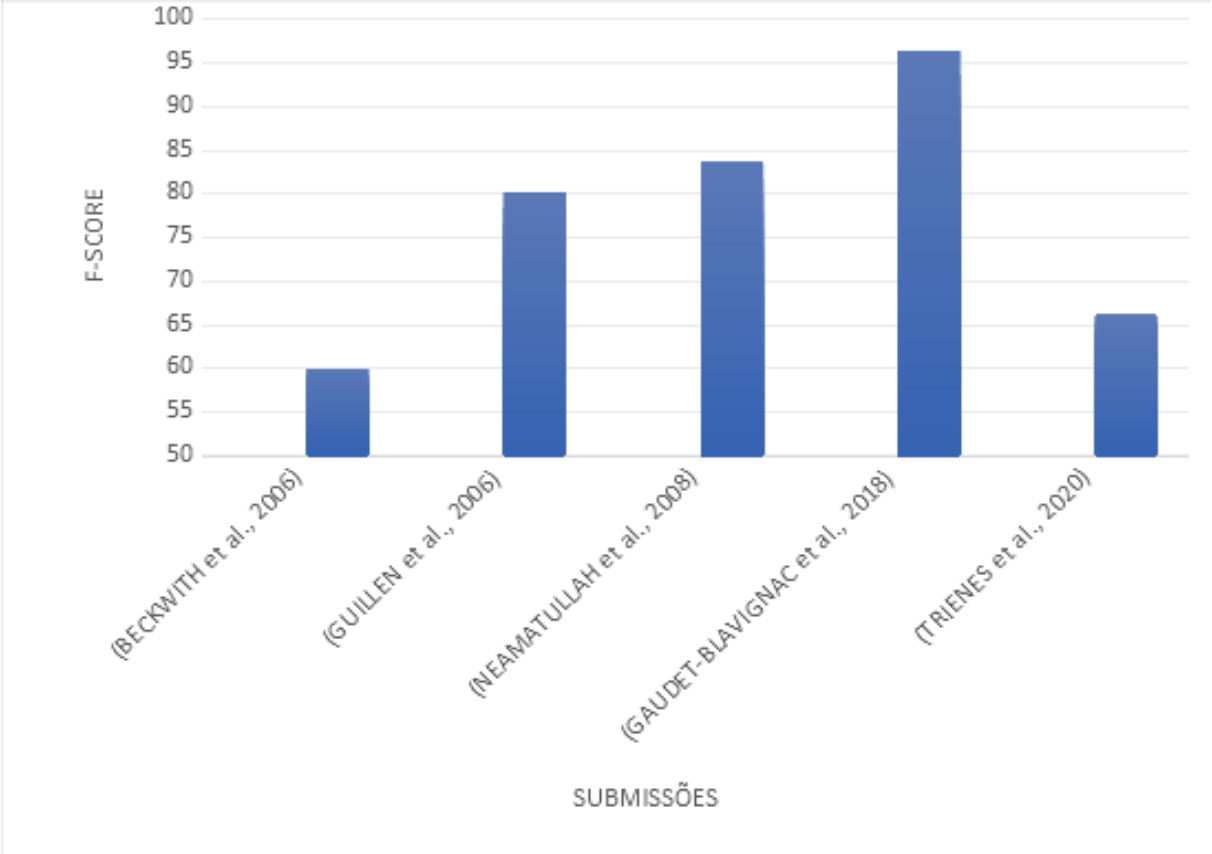
regulares para identificar números de seguro social e registros médicos, entre outros. Gaudet-Blavignac et al. (2018) usaram o *corpus* Unitex[11] e outras fontes de dados como dicionários. Os dicionários foram integrados com um autômato finito para a criação de regras de análise de contexto.

Alguns estudos também recorrem a ontologias e outros vocabulários controlados para identificar o que não é dado sensível e reduzir as ocorrências de falsos positivos. Ruch et al. (2000) usam o framework chamado MEDTAG que possui uma abordagem mais próxima da linguística. Primeiramente são realizadas marcações nos *tokens* usando classes gramaticais. Em seguida é realizada uma segunda marcação usando uma desambiguação de sentido. Baseando-se nessas marcações, o algoritmo avalia se um *token* é um dado sensível ou não. O MEDTAG é equipado com o vocabulário do UMLS (*Unified Medical Language System*) e mais de quarenta regras de desambiguação de termos. O trabalho de Berman (2003) usa o UMLS para substituir o vocabulário controlado encontrado em documentos clínicos por códigos correspondentes. O sistema desenvolvido pelo autor funciona em três etapas: (1) análise de texto para identificar palavras, frases e *stop-words*; (2) as palavras que não são *stop-words* são mapeadas no vocabulário controlado da UMLS; (3) os termos mapeados são substituídos por códigos correspondentes na UMLS e o restante é desidentificado. O algoritmo de Gupta, Saul e Gilbertson (2004) usa o UMLS para identificar termos que não precisam ser desidentificados. Para a desidentificação eles usaram expressões regulares para dados numéricos e listas do censo para identificar nomes e localidades. O diferencial do trabalho, segundo os autores, está na análise de cabeçalho de documentos clínicos para identificar dados relevantes na ficha de um paciente. Neamatullah et al. (2008) usaram o UMLS para descartar termos médicos encontrados na análise de documentos. O algoritmo foi dividido em três partes: (1) dados sensíveis numéricos foram identificados usando expressões regulares; (2) dados sensíveis não numéricos foram identificados usando listas, análise de contexto e expressões regulares e (3) os dados sensíveis identificados em processos anteriores são substituídos por classificações que indicam a qual categoria de dados sensíveis da HIPAA eles pertencem. Morrison et al. (2009) usaram um sistema chamado MedLEE (*Medical Language Extraction and Encoding System*) para realizar o preprocessamento de documentos clínicos. Em seguida esses documentos foram comparados usando o UMLS para identificar e extrair termos médicos. Demais termos que não foram identificados foram descartados e desidentificados. Em seguida foi realizada uma revisão manual para remover os dados sensíveis que não foram descartados pelo MedLEE.

A Figura 3 apresenta a *performance* dos trabalhos submetidos que usaram o F-Score como métrica de avaliação.

[11] Disponível em https://unitexgramlab.org/. Acessado em março de 2022.

Figura 3 – Análise de *performance* em F-Score dos métodos baseados em regras.



Dentre os trabalhos avaliados, apenas o de Gaudet-Blavignac et al. (2018) obteve *performance* de F-Score superior aos 95% considerados aceitáveis pela literatura. Os demais trabalhos obtiveram uma *performance* abaixo dos 85%, o que pode ser considerado insatisfatório.

Uma característica que destaca os trabalhos que usam métodos baseados em regras são as métricas de avaliação usadas que destoam do tradicional F-Score. Sweeney (1996), Ruch et al. (2000), Thomas et al. (2002), Gupta, Saul e Gilbertson (2004), Friedlin e McDonald (2008) e Morrison et al. (2009) avaliaram apenas a porcentagem de dados sensíveis removidos. Os autores atingiram a *performance* de 99%, 98-99%, 92,7%, 99,1%, 99,47% e 96,8% respectivamente. A avaliação de Berman (2003) considera apenas a velocidade de processamento de desidentificação, conseguindo processar mais de meio milhão de documentos clínicos em menos de uma hora. Por fim, Fielstein, Brown e Speroff (2004) avaliaram seu trabalho apenas usando os positivos verdadeiros e positivos falsos do algoritmo chegando a 92% e 99% de *performance* respectivamente.

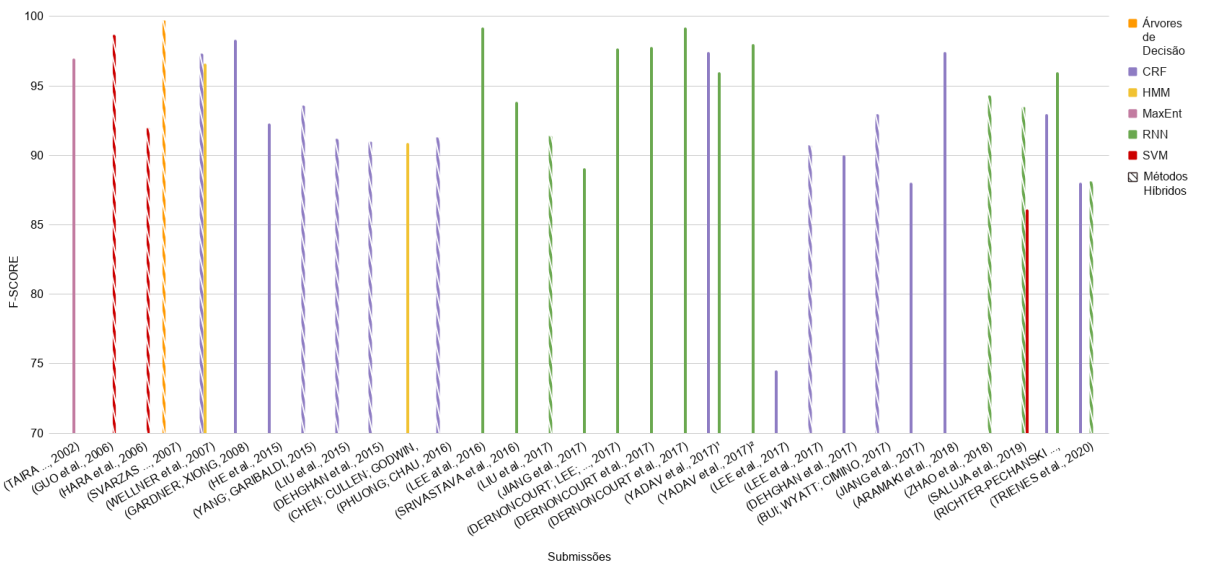
4.4.3 Análise de métodos de desidentificação

Os resultados encontrados por Uzuner, Luo e Szolovits (2007), Meystre et al. (2010) e Yogarajan, Pfahringer e Mayo (2019) apontam para a mesma direção. Esses autores

identificaram que os métodos híbridos ofereceram uma abordagem melhor para a solução do problema de desidentificação, seguidos pelos métodos de aprendizado de máquina e os métodos baseados em regras respectivamente. Os resultados encontrados por Uzuner, Luo e Szolovits (2007) mostraram que ambiguidades podem afetar a *performance* enquanto a identificação de dados sensíveis incomuns foi eficientemente detectada. Na análise de Yogarajan, Pfahringer e Mayo (2019), ressalta-se que sistemas desenvolvidos posteriormente aos eventos analisados favorecem as abordagens de aprendizado profundo em relação aos métodos híbridos.

A Figura 4 traz uma série de percepções sobre a evolução de métodos de desidentificação neste século. O primeiro deles é a mudança de paradigma ao longo do tempo. Na primeira década, os métodos predominantes eram baseados em HMM e SVM. A partir de 2015, houve uma predominância dos métodos que usam CRF e aprendizado profundo como RNN. Esses métodos se mostraram mais eficientes justificando sua larga adoção. O aprendizado profundo vem ganhando notoriedade na área de inteligência artificial em aplicações distintas. Seu uso em desidentificação vem-se mostrando promissor e visto com bons olhos, embora ainda seja preciso mais tempo para aperfeiçoar essas técnicas e consolidar os resultados na literatura.

Figura 4 – Análise de *performance* em F-Score dos métodos baseados em aprendizado de máquina.



Dos resultados encontrados na subseção 4.4.1, apenas quinze obtiveram a *performance* de 95%, considerada aceitável para esses sistemas. Desses quinze sistemas, onze foram desenvolvidos usando CRF ou RNN, o que reforça ainda mais a eficácia dessas abordagens. **Desses quinze métodos que obtiveram *performance* aceitável, apenas três foram híbridos e todos datam de antes de 2010.** Este resultado vai na contramão das análises dos autores que citam as abordagens híbridas como sendo as mais

eficientes. Em contrapartida, os piores resultados apresentados vêm de métodos puramente baseados em aprendizado de máquina. Isso favorece a abordagem mista como tendo a melhor *performance* média.

Embora os resultados mostrados nos últimos cinco anos se mostrem promissores, ainda há muito o que ser feito nessa área de pesquisa. A primeira questão a ser levantada é o teste desses sistemas em dados heterogêneos. Grande parte dos sistemas avaliados neste trabalho foi desenvolvidos em ambientes controlados, com *corpora* previamente fornecidos para os pesquisadores. A alteração de um *corpus* usado para a tarefa de desidentificação pode implicar queda de *performance*, como visto nos trabalhos de Dehghan et al. (2015) e Dehghan et al. (2017). Uma alternativa é a utilização do *transfer learning*, usada por Yadav et al. (2017) para o treinamento de desidentificação. Essa técnica pode ser aplicada em outros domínios além do médico, como em documentos legais, jornais, literatura, entre outros. Estes demais domínios podem ajudar no treinamento de desidentificação de dados como profissão, localizações, nomes próprios etc. Além disso, as políticas de proteção de dados diferem ao redor do mundo e tendem a se tornarem mais exigentes conforme os documentos são publicados. Embora haja o consenso de que alguns dados como nome pessoal, endereço, entre outros, sejam elegíveis para desidentificação, qualquer mudança fora do escopo de pesquisa torna a desidentificação inviável, na prática.

Outro aspecto é a manutenção da compreensibilidade de documentos. O processo de desidentificação deve assegurar que os documentos continuem legíveis e interoperáveis do ponto de vista clínico. Além disso, há o problema de excesso de desidentificação. Nesses casos os falsos positivos, dados erroneamente desidentificados, podem afetar a legibilidade de documentos. Os trabalhos analisados focam quase em sua totalidade a avaliação de desempenho por meio da métrica do F-Score. O F-Score, por sua vez, foca apenas na precisão e revocação na identificação de dados sensíveis. Nenhum trabalho analisou a desidentificação sob a ótica da compreensibilidade de documentos desidentificados.

A terceira questão do processo de desidentificação é a análise do valor gerado em documentos desidentificados para pesquisa em saúde. O processo de remoção e substituição de dados-chave podem prejudicar a análise clínica de um documento. Pode-se citar as datas como exemplo. A simples substituição de uma data por outra totalmente aleatória pode afetar o julgamento de um profissional de saúde ao interpretar um documento desidentificado. Datas possuem informações relevantes como a sazonalidade de algumas doenças, identificação de surtos e epidemias, relação com a idade do paciente, entre outras questões que devem ser consideradas quando for implementado um processo de desidentificação para as categorias da HIPAA.

Há também o problema de reprodutibilidade dos trabalhos avaliados. Sistemas de desidentificação são difíceis de ser replicados para estudos científicos e nem podem ser comparados diretamente, pois os *corpora* usados estão diretamente ligados às *performances*

obtidas. Apesar disso, a análise desses trabalhos oferece uma visão macro da literatura.

Por fim, nenhum sistema citado neste trabalho usou *corpora* em português e nem foi desenvolvido para a língua portuguesa. Isso sugere a falta de pesquisa na área e uma oportunidade para o desenvolvimento de ações que colaborem com o avanço da pesquisa médica em território brasileiro.

4.5 Discussões

A desidentificação como campo de pesquisa tem-se mostrado uma área ampla e multidisciplinar com conhecimentos que envolvem a linguística, a computação e a recuperação de informação. No contexto médico, a desidentificação se torna mais relevante. Apesar das questões levantadas na subseção 4.4.3, o compartilhamento de documentos clínicos é crucial para o avanço da pesquisa em saúde. Mas essa é uma área sujeita às implicações de legislações locais, pois não é possível o compartilhamento de informações médicas sem quebrar sigilos.

Sob esse ponto de vista, a tarefa de desidentificação assume um aspecto social e legal. Proteger a privacidade de pessoas e instituições de saúde é fundamental. A tarefa, no entanto, não é trivial. Além de estar submissa à legislação no que tange à privacidade, apontar qual dado precisa ser desidentificado envolve muitas nuances e interpretações. Como foi discutido neste trabalho, pequenos fatores podem ser suficientes para a identificação de um indivíduo. Há também a forma que esses dados são desidentificados de maneira que não sejam possíveis de serem reidentificados e também não prejudiquem a compreensibilidade nem a legibilidade de documentos clínicos.

Os sistemas desenvolvidos para automatizar a tarefa de desidentificação foram baseados em regras linguísticas e/ou técnicas de inteligência artificial, como aprendizado de máquina e aprendizado profundo. A maioria dos sistemas citados neste trabalho ocorreu em eventos criados para desenvolver competições de desidentificação. Essas competições se mostram importantes para o desenvolvimento dessa área de pesquisa e oferecem oportunidade de acesso aos *corpora* de documentos clínicos. Também elevam o nível de dificuldade de desidentificação assumindo que os elementos da HIPAA são insuficientes para realizar uma completa proteção de dados clínicos. Além disso, essas competições possuem uma natureza aberta de dados, fornecendo acesso à comunidade. Isso é importante para o avanço da pesquisa, mas, em contrapartida, os resultados das publicações originárias dessas competições são limitados aos *corpora* fornecidos, visto que é difícil ter acesso a esses dados em outras ocasiões.

Os resultados das análises mostram que ainda há muito o que ser feito para que a automatização atinja *performance* confiável e se mostre adaptável na heterogeneidade que compõem os documentos clínicos. Essa demanda é essencial, dado que o mundo todo

vem adotando os documentos eletrônicos como medidas para desafogar o espaço físico de unidades de saúde e promover interoperabilidade entre diferentes setores.

No Brasil não é diferente. No entanto, a falta de trabalhos similares na área voltados para a língua portuguesa preocupa. Isso pode deixar a pesquisa nacional em desidentificação e proteção à privacidade defasada em relação ao resto do mundo. É importante que iniciativas neste sentido ocorram para fomentar a pesquisa no país. Para isso sugere-se duas medidas prioritárias: (1) criação de um *corpus* de documentos clínicos com dados sensíveis substituídos por pseudônimos de forma que possa ser disponibilizado publicamente para a pesquisa, e (2) fomentar competições de desidentificação com esse *corpus* para desenvolver técnicas de desidentificação especificamente voltadas para a língua portuguesa. Dessa maneira é possível construir uma comunidade científica profícua em torno desse campo de pesquisa e colher resultados que contribuam com o avanço da área de saúde no país.

Um pesquisador de posse do panorama fornecido por este trabalho estará apto a fazer parte das iniciativas propostas nesta seção. Acredita-se que este trabalho contribuirá para guiar os pesquisadores que desejam fomentar o campo de pesquisa de desidentificação de documentos clínicos.

5 Metodologia

Os capítulos anteriores deste trabalho forneceram os materiais necessários para que a metodologia fosse construída. O Capítulo 2 permitiu identificar os tipos mais relevantes de documentos que compõem os *corpora* de documentos clínicos, as diferentes aplicações em que são usados, como é feita a política de disponibilização, o tamanho médio e as anotações feitas. O Capítulo 3 debruçou-se sobre o sumário de alta, um dos potenciais documentos para aplicação desta pesquisa. O sumário consiste num documento formal, bem estruturado e com informações relevantes sobre a história clínica de um paciente. O Capítulo 4 permitiu analisar sob uma ótica ampla a tarefa de desidentificação, suas origens, seus riscos, como são feitas as técnicas de desidentificação e quais são as que obtiveram melhor *performance*. Dessa maneira, é possível delinear uma estratégia para o desenvolvimento dos objetivos propostos no Capítulo 1.

Este trabalho possui um viés filosófico pós-positivista que acredita que nenhum conhecimento é irrefutável, mas sim especulativo (PHILLIPS; PHILLIPS; BURBULES, 2000, p.9). O método científico pós-positivista é fundamentado em hipóteses que apoiam uma teoria. Portanto, pretende-se recorrer ao teste de teorias através de observações empíricas para validá-las ou não de acordo com o cenário proposto pela pesquisa.

O presente capítulo está dividido da seguinte maneira: a seção 5.1 discute a metodologia **de** pesquisa. Entende-se por metodologia de pesquisa a sua classificação como instrumento de estudo. É a caracterização da pesquisa em seus pontos de vista como a natureza, abordagem, objetivo e procedimentos técnicos. Trata-se do desenho, da modelagem conceitual e a abordagem filosófica que permeiam a condução da pesquisa (CRESWELL; CRESWELL, 2017). A seção 5.2 apresenta a metodologia **da** pesquisa que diz respeito aos passos que serão executados durante a sua condução. Ela contém todos os objetos de estudo e processos adotados para alcançar o resultado proposto. Por fim, a seção 5.3 traz um resumo do capítulo apresentando uma tabela e um fluxograma das etapas de desenvolvimento.

5.1 Metodologia de pesquisa

A pesquisa pode ser considerada, de acordo com Marconi e Lakatos (2017) e Gil (2019), com as seguintes características:

- **segundo sua natureza:** aplicada;
- **segundo sua abordagem:** quantitativa;

- **segundo seu objetivo:** descritiva;
- **segundo os procedimentos técnicos:** experimental.

Trata-se de uma pesquisa **aplicada**, pois tem como objetivo primário uma aplicação prática que é desenvolver métodos de desidentificação de dados sensíveis que possibilitem a criação de um *corpus* desidentificado de documentos clínicos na língua portuguesa do Brasil. Está dirigida a um problema específico, visto que a literatura trata apenas de textos em outros idiomas, mas não na língua portuguesa.

A pesquisa é **quantitativa**, pois o objeto de pesquisa é quantificável. Busca-se traduzir o desempenho de métodos de desidentificação por análise estatística. Avaliações de desempenho como taxas de precisão, revocação e acurácia serão realizadas.

Trata-se de pesquisa **descritiva** porque descreve, como o objeto de pesquisa, dados sigilosos, comporta-se quando submetido a algoritmos de desidentificação. A *performance* da desidentificação será avaliada durante a execução do projeto, relatando as mudanças que trouxeram melhor ou pior desempenho para o método.

E, por fim, é uma pesquisa **experimental**, pois experimenta técnicas de desidentificação dentro de um cenário controlado. Como visto nos capítulos anteriores, o *corpus* pode trazer resultados diferentes para um mesmo algoritmo de desidentificação.

5.2 Metodologia da pesquisa

A metodologia da pesquisa terá duas etapas práticas: criação de *corpus* e desenvolvimento de algoritmo de desidentificação de documentos clínicos. As subseções seguintes detalham melhor o objeto de pesquisa e as medidas tomadas para a realização dessas etapas.

5.2.1 O objeto da pesquisa e a coleta de dados

Dados clínicos do hospital das clínicas (HC) da UFMG são o objeto da pesquisa. Uma pesquisa preliminar indicou a existência de mais de 1 300 000 documentos armazenados em 140 GB de espaço, número que cresce diariamente pois o HC-UFMG/EBSERH é um hospital de referência localizado no centro de Belo Horizonte com grande demanda por parte de cidadãos belorizontinos e também do Estado de Minas Gerais.

Há, no entanto, uma série de documentos que não serão contemplados nesta pesquisa, como termos de consentimento, exames clínicos e de imagem, dentre outros. A revisão de literatura sugeriu que o melhor tipo de documento para a tarefa de desidentificação é o **sumário de alta**, pois ele contém informações relevantes de pacientes estruturadas de maneira formal, conforme justificado no Capítulo 3.

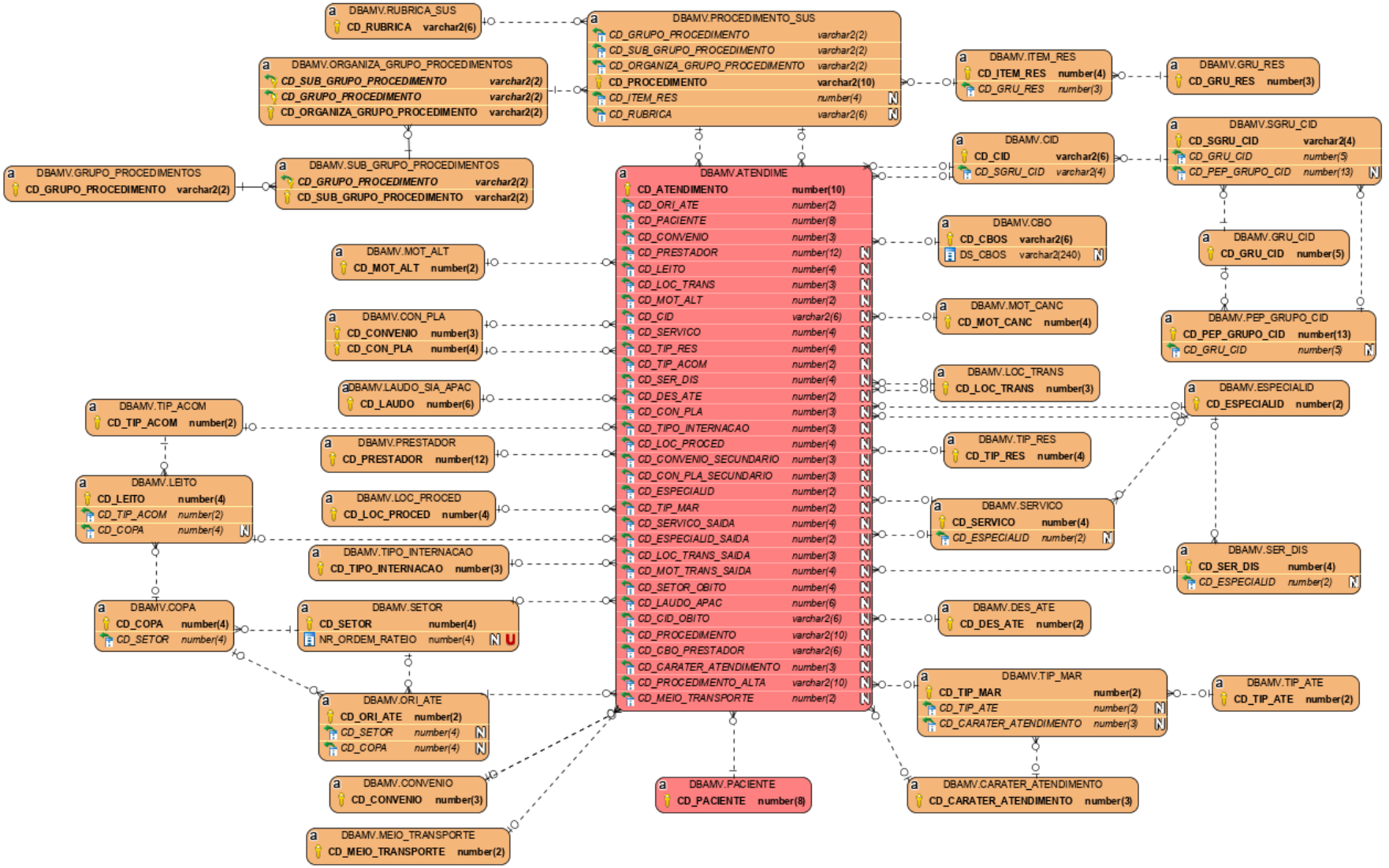
Os documentos do HC-UFMG/EBSERH são armazenados num banco de dados relacional. Esse banco alimenta o sistema que gerencia o hospital. O sistema, no entanto, é terceirizado e o HC-UFMG/EBSERH tem posse dos dados produzidos pelo próprio hospital e não da infraestrutura por trás do desenvolvimento. Dessa maneira, o acesso ao banco de dados fica limitado por consultas para acessar as informações produzidas pelo hospital.

Uma conversa informal com uma profissional de saúde ligada ao HC-UFMG/EBSERH revelou que o sumário de alta do hospital contém os seguintes campos:

- Cabeçalho com identificação do paciente;
- Motivo da admissão e diagnósticos relevantes;
- Procedimentos diagnósticos;
- Procedimentos cirúrgicos;
- Procedimentos terapêuticos;
- Resumo da internação;
- Alergias e reações adversas;
- Diagnósticos à alta;
- Prescrição e orientações;
- Informações da alta;
- Rodapé com identificação do profissional de saúde.

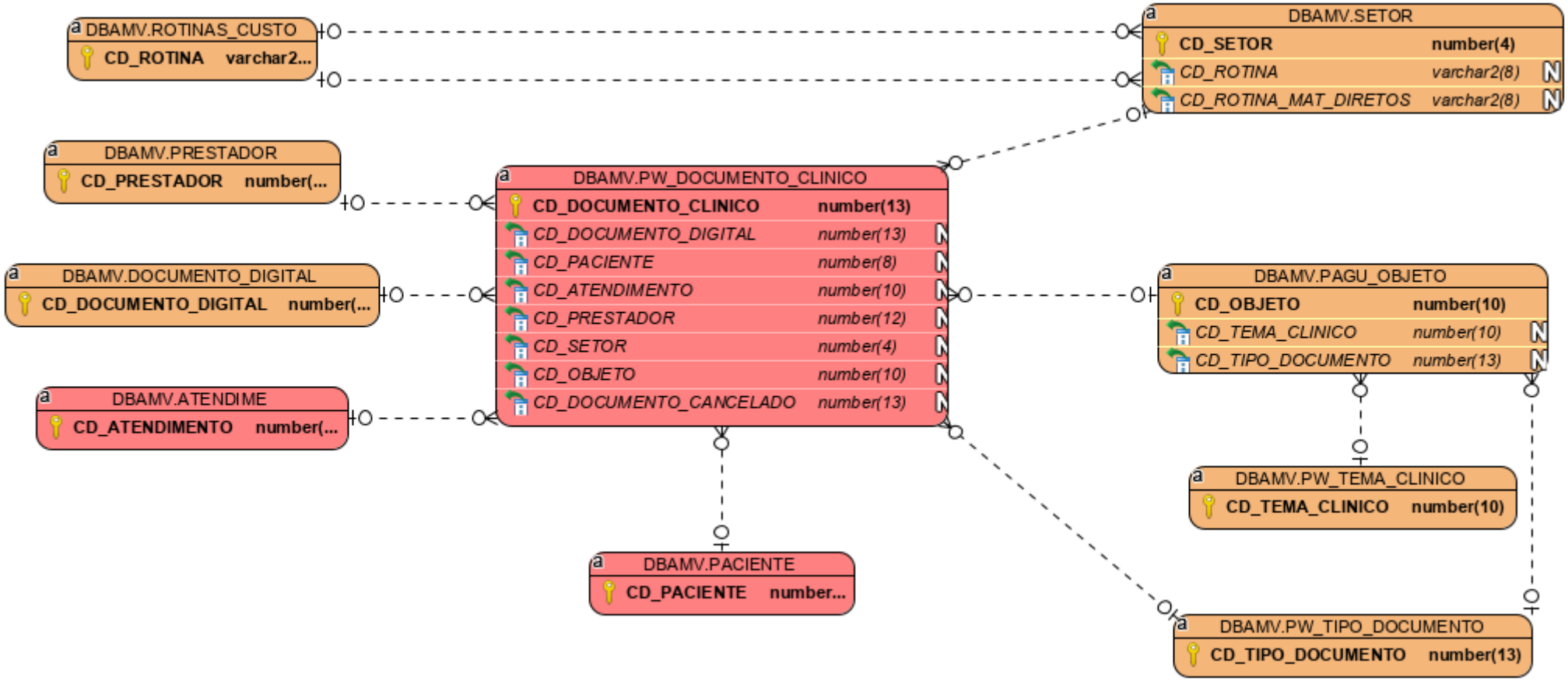
Todos os campos, listados acima, estão na forma não estruturada. Uma investigação preliminar apontou que o banco de dados possui mais de 3000 tabelas, sendo que nenhum DER foi fornecido para saber como elas se relacionam. Um processo de engenharia reversa foi usado para identificar alguns relacionamentos entre as principais tabelas de interesse deste trabalho, como podem ser observado nas figuras 5, 6 e 7. No entanto, o excesso de tabelas torna esse processo inviável a curto prazo, fazendo a parte de coleta de dados uma tarefa experimental em busca dos sumários de alta.

Figura 5 – DER para a tabela Atendimento.



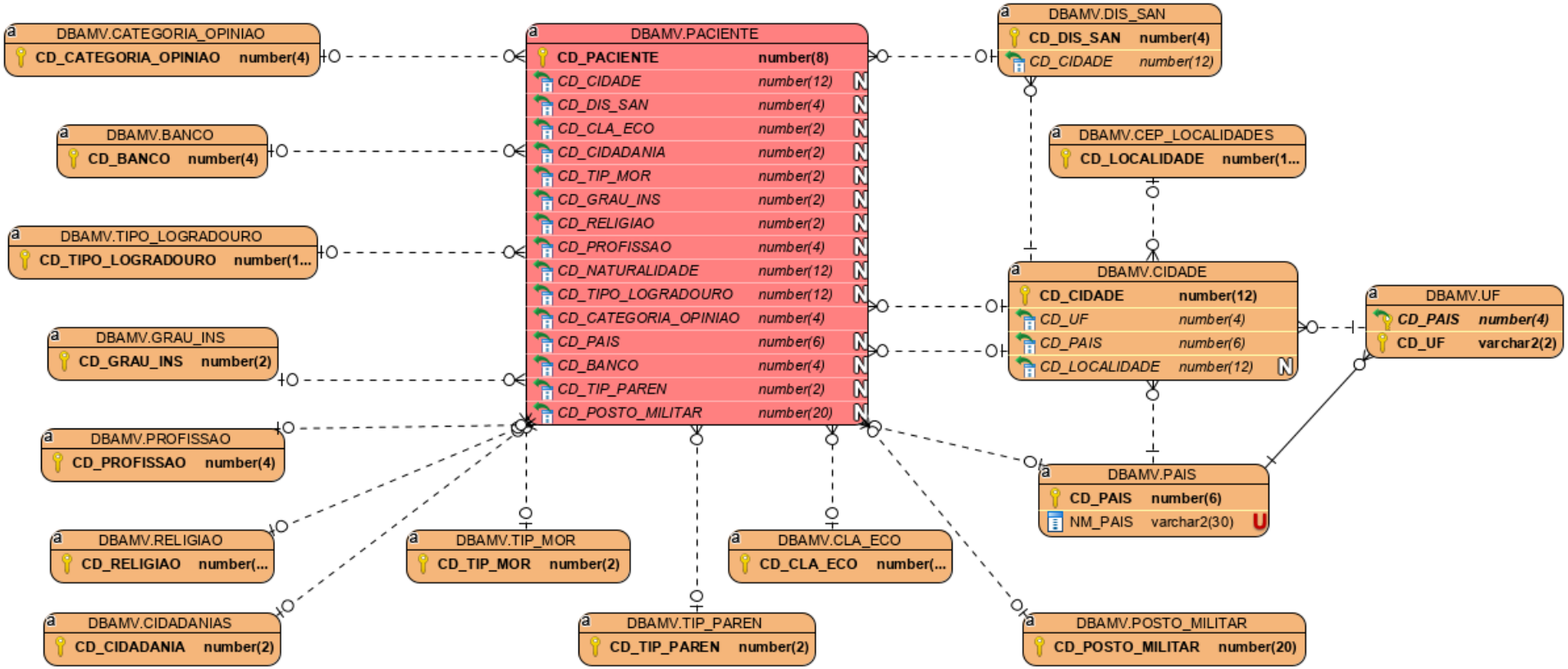
Fonte: Elaborado pelo próprio autor.

Figura 6 – DER para a tabela Documento Clínico.



Fonte: Elaborado pelo próprio autor.

Figura 7 – DER para a tabela Paciente.



Fonte: Elaborado pelo próprio autor.

As figuras 5, 6 e 7 mostram o fluxo inicial de aquisição de dados do HC-UFMG/EBSERH. Na primeira etapa ocorre o atendimento onde são recolhidos dados preliminares sobre um paciente, como esse paciente deu entrada no hospital, se é atendimento público ou privado, se possui convênio de saúde, qual leito será encaminhado etc. Este atendimento também recolhe as informações sobre o paciente como endereço, religião, profissão etc. Por fim, é gerado o documento clínico que conterá todo o histórico de procedimentos adotados no tratamento. As tabelas em vermelho das figuras 5, 6 e 7 representam o cerne principal que se relaciona com as demais tabelas. Os campos das tabelas mostradas nas figuras 5 e 6 foram omitidos para simplificar as imagens. Nota-se, no entanto, que diferentes dados passíveis de proteção estão armazenados em colunas simples das tabelas como, por exemplo, o nome da cidade onde o paciente mora está diretamente atrelado à tabela "Cidade".

Desse modo optou-se pela estratégia de localizar campos de textos livres usados pelos profissionais de saúde para descrever detalhes de tratamentos aplicados aos pacientes. Uma pesquisa preliminar indicou que esse campo encontra-se na tabela "Documento Digital" que está diretamente relacionada à tabela "Documento Clínico".

Devido à dificuldade de acesso à base de dados do HC-UFMG/EBSERH esta estratégia não foi executada. A obtenção de dados no HC-UFMG/EBSERH tornou-se inviável, embora o projeto tivesse sido aprovado no CEP, como mostra o Anexo A. O primeiro motivo para a interrupção dessa estratégia é a já apresentada inviabilidade de elaborar consultas no banco de dados sem um DER adequado. O segundo motivo é a indisponibilidade de profissionais de tecnologia do HC-UFMG/EBSERH para atender as demandas de extração de dados necessárias para a condução deste projeto.

Uma alternativa pensada para contornar esse impasse foi solicitar os dados de outro hospital. Um novo projeto foi escrito e submetido ao CEP, dessa vez para obter os dados do Felício Rocho[1], hospital geral particular da região de Belo Horizonte. O projeto foi aprovado como consta no Anexo B, mas a aquisição de dados foi negada pela diretoria. Os casos são discutidos com mais detalhes na seção 7.1.

A solução encontrada utilizou dados do trabalho de SÁ (2018), previamente extraídos do HC-UFMG/EBSERH. O Anexo A permitiu a utilização desses dados, pois sua aprovação cobre qualquer uso de dados do HC-UFMG/EBSERH.

Os documentos do trabalho de SÁ (2018) são de evolução médica de emergência e documentos de avaliação à admissão hospitalar. Para entender a estrutura do documento, é preciso, antes de mais nada, conceitualizar campos estruturados e não estruturados. Os campos estruturados são aqueles definidos através de uma estrutura rígida, para armazenar um tipo específico de dado previamente planejado. Dados sensíveis como nomes, e-mail,

[1] Disponível em https://www.feliciorocho.org.br/fundacao. Acessado em março de 2022.

endereço, data de nascimento, entre outros são armazenados em campos estruturados do banco de dados. Já os campos não estruturados são compostos por uma estrutura flexível ou até mesmo ausente. Eles foram projetados para armazenar dados que podem ter diferentes formas como documentos de texto, arquivos, imagens, vídeos etc.

Optou-se por lidar apenas com campos não estruturados porque os Sistemas Gerenciadores de Banco de Dados (SGBDs) já possuem recursos nativos para a proteção de dados sensíveis estruturados. A Tabela 8 mostra os campos não estruturados obtidos da coleta de dados.

Tabela 8 – Campos não estruturados de prontuários eletrônicos de pacientes

Nome do Campo	**Descrição**
ANAMNESE ESPECÍFICA	Sintomas que um paciente apresenta.
HISTÓRIA ATUAL	História atual de um paciente no momento do atendimento.
HISTÓRIA DA MOLÉSTIA ATUAL (CONSIDERAR ASPECTOS FÍSICOS, PSÍQUICOS, SOCIAIS E ESPIRITUAIS)	História atual de um paciente no momento do atendimento.
HISTÓRIA FAMILIAR	Dados familiares de um paciente.
HISTÓRIA PREGRESSA	Histórias anteriores de um paciente.
HISTÓRIA SOCIAL	História comportamental de um paciente (tabagismo, religião, orientação sexual etc.).
HISTÓRICO ALIMENTAR	Dieta e hábitos alimentares de um paciente.
IMPRESSAO_DIAGNOSTICA	Relatório por escrito de exames de imagem de um paciente.
MEDICAMENTOS EM USO	Medicamentos de que um paciente está fazendo uso.
MOTIVO DO ATENDIMENTO	Motivo que levou um paciente a procurar atendimento de saúde.
PROBLEMAS IDENTIFICADOS	Identificação de problemas pelo profissional de saúde no momento do atendimento
PROCEDIMENTOS TERAPÊUTICOS	Decisão tomada pelo profissional de saúde em relação aos problemas identificados.
TRANSFERÊNCIA DO CUIDADO	Identificação da necessidade de transferência do cuidado de um paciente para outra pessoa.
TRATAMENTO_PROPOSTO	Prescrição do profissional de saúde para combater os problemas identificados.

Fonte: Elaborado pelo autor.

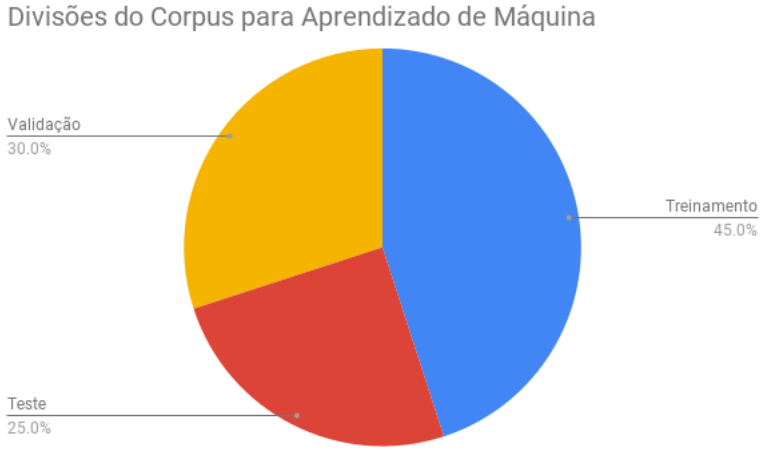
Os campos foram obtidos através de um arquivo em formato de planilha eletrônica exportada do banco de dados. A estrutura do arquivo traz os campos de maneira desordenada onde cada linha possui três colunas distintas: número do atendimento, campo e valor. A coluna número do atendimento é referente à chave primária do banco de dados e possui um valor único que identifica um prontuário de saúde dos demais. A coluna campo é o nome de um dos campos listados na Tabela 8 e a coluna valor é o campo de texto com os dados, que pode estar preenchido ou não.

A revisão apresentada no Capítulo 2 mostrou que *corpora* possuem tamanhos e aplicações distintos. Apesar da conclusão da revisão sugerir uma quantidade inicial de 3254 documentos para a construção de um *corpus* de estudo para escopo único, essa etapa não pôde ser executada pela inviabilidade de aquisição de dados, já justificada nessa seção, e também devido aos prazos e limitações desta pesquisa.

5.2.2 Construção de *corpus*

O presente trabalho teve como objetivo construir um *corpus* para aplicação de PLN. Esse *corpus* contém dados citados na subseção 5.2.1. Ele foi dividido em três subcorpora: um para calibrar o algoritmo de desidentificação, outro para fazer um teste de *performance* e o terceiro para a execução do algoritmo de desidentificação. A Figura 8 indica que a proporção sugerida para essa divisão seja de 45% para calibração ou treinamento do algoritmo, 25% para teste de *performance* e 30% com dados livres para validação do modelo de desidentificação. (PUSTEJOVSKY; STUBBS, 2013).

Figura 8 – Divisão de um *corpus* para execução de algoritmos de aprendizado de máquina

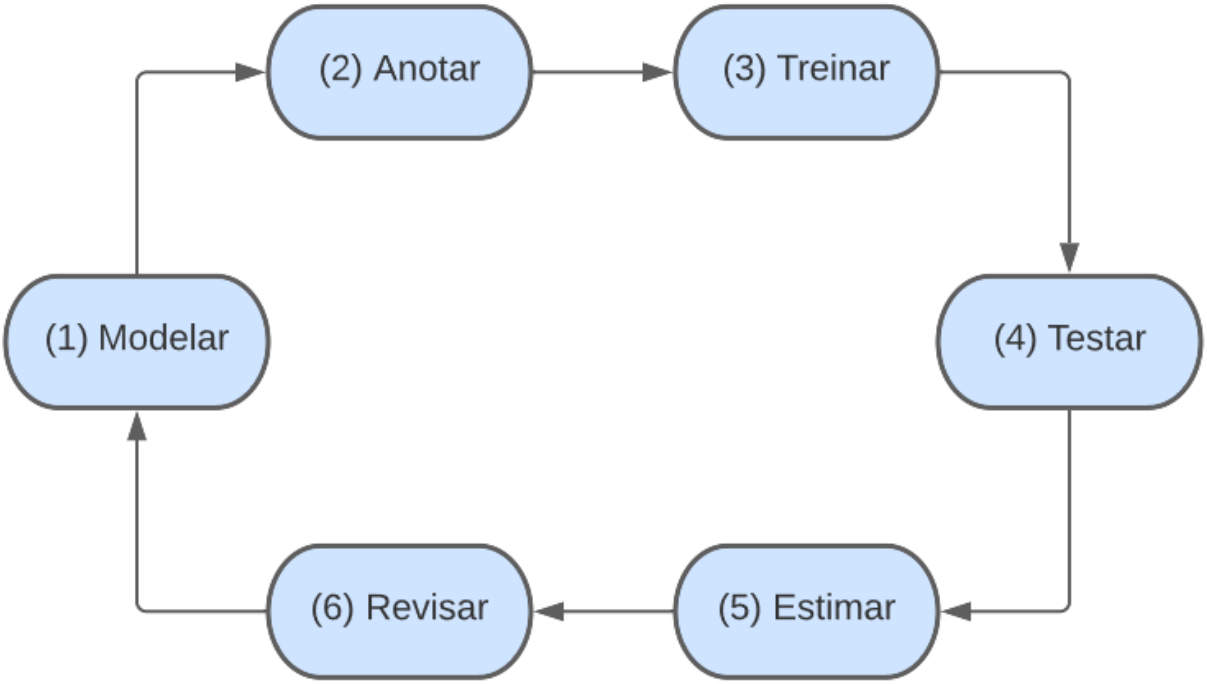


Fonte: Adaptado de Pustejovsky e Stubbs (2013).

A construção dos *corpora* segue a metodologia MATTER (*Model, Annotate, Train, Test, Evaluate, Revise*) apresentada por Pustejovsky e Stubbs (2013). Mais precisamente as etapas de *Model* e *Annotate*, sendo as demais etapas usadas no restante do projeto.

Segundo os autores, a metodologia MATTER é adequada para desenvolver aplicações que usam PLN como fonte base de estudo, como a proposta por este trabalho. Trata-se de um ciclo de etapas, mostrados na Figura 9, executadas durante o desenvolvimento do escopo proposto.

Figura 9 – Etapas do modelo MATTER.



Fonte: Adaptado de Pustejovsky e Stubbs (2013).

A primeira etapa da metodologia MATTER é modelar os dados. Escolhe-se quais são os tipos de dados que serão anotados e quais as etiquetas que serão usadas na anotação. Os tipos de dados e as etiquetas foram adaptados de outro projeto de desidentificação da literatura que usa as categorias da HIPAA para anotação. Tendo em vista que as categorias são padronizadas, baseadas nas generalizações de Meystre et al. (2010), adotaram-se as seguintes etiquetas de Stubbs e Uzuner (2017) mostrados na Tabela 9.

Os nomes das categorias representadas na Tabela 9 estão em caixa alta e sem acentos para diferenciar as classificações dos *tokens*. Cada uma dessas categorias terão duas subcategorias pertencentes ao modelo IOB apresentado no Capítulo 4. A subcategoria "B" indica o começo de uma categoria e a "I" indica a sequência de uma categoria. A terceira categoria, "O", do modelo IOB é usada distintamente das demais. Sua aplicação é para indicar que uma palavra não pertence a nenhuma das categorias da Tabela 9. O texto 4.1 mostrado originalmente no Capítulo 4 seria marcado da seguinte forma, como

Tabela 9 – Especificações de etiquetagem de *tokens* sensíveis para a construção dos *corpora*

Categoria do PHI	**Categorias de anotação**
NOME	Usuário, profissional de saúde, paciente
PROFISSAO	Nomes de profissão
LOCAL	Quarto, departamento, hospital, empresa, rua, cidade, estado, país, CEP, outros
IDADE	Números que indiquem a idade algum indivíduo
DATA	Datas
CONTATO	Telefone, fax, e-mail, endereços web como URLs, URIs, IPs, outros
ID	RG, CPF, carteira de motorista, passaporte, identificação do SUS, registro de saúde, número de plano de saúde, IMEI, qualquer número ou código serial

Fonte: Adaptado de Stubbs e Uzuner (2017)

mostrado nos exemplos a seguir.

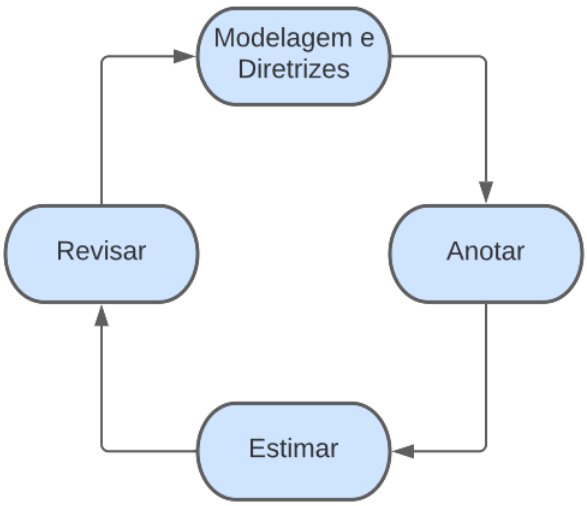
Elisa foi atendida pelo Dr. **Otávio** no **Hospital das Clínicas** no dia **25/07/2019**.
A paciente foi diagnosticada com dengue. (Texto Original)
(5.1)

B-NOME O O O O **B-NOME** O **B-LOCAL I-LOCAL I-LOCAL** O O **B-DATA**
O O O O O O. (Texto Anotado)
(5.2)

Essas categorias podem sofrer alterações em trabalhos futuros. A ABNT/CEE-78) (2019), por meio da NBR ISO 25 237, prevê outras variáveis de desidentificação para documentos clínicos gerados no Brasil. Alguns exemplos são as variáveis atípicas, que possuem potencial de identificação do paciente por ocorrências raras como diagnósticos raros, procedimentos incomuns, fato noticiado publicamente etc. Entende-se que essas categorias fogem do escopo deste trabalho por requererem auxílio de um profissional de saúde qualificado para identificar e anotar estes dados. A metodologia apresentada nesse trabalho, no entanto, pode ser adequada para quaisquer quantidades de categorias que forem necessárias para a desidentificação.

Isso já está previsto na metodologia MATTER. A etapa Modelar da metodologia possui uma subetapa chamada MAMA (*Model-Annotate Model-Annotate* em inglês e Modelar-Anotar-Estimar-Revisar em português). Nessa subetapa a modelagem é anotada, estimada e revisada para detectar que se o processo de modelagem está de acordo com o previsto ou se será necessário aperfeiçoar a etapa de modelagem. A Figura 10 mostra os ciclos da subetapa MAMA.

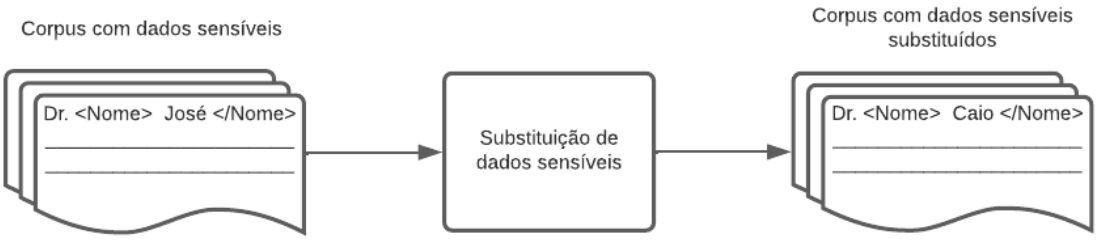
Figura 10 – Etapas da subetapa MAMA



Fonte: Adaptado de Pustejovsky e Stubbs (2013).

A anotação, segunda etapa da metodologia MATTER, consiste em aplicar as etiquetas, definidas na modelagem, no corpus. Essa anotação foi feita usando o software Label Studio[2]. A escolha justifica-se por ser um programa gratuito, de código aberto, de característica geral para anotação de dados e possuir integração com métodos de aprendizado de máquina. Como resultado da anotação, há documentos selecionados com os dados sensíveis substituídos por pseudônimos, conforme ilustram as Figuras 11 e 12.

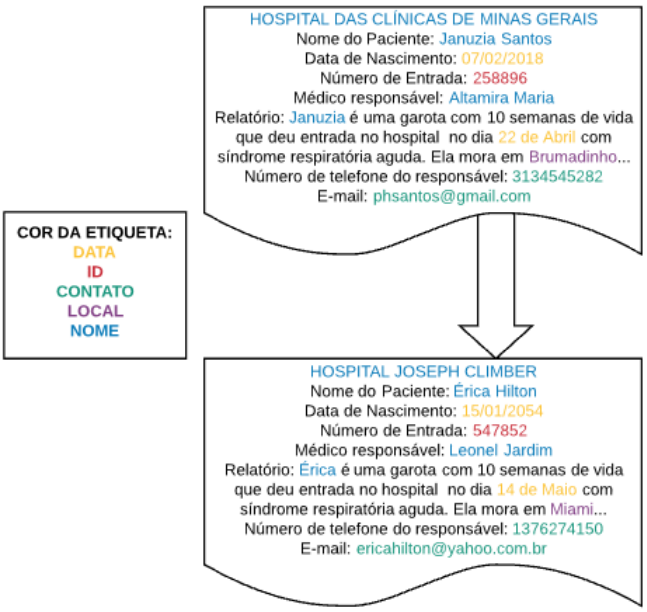
Figura 11 – Processo de substituição de dados sensíveis



Fonte: Adaptado de Deleger et al. (2014).

[2] Disponível em https://labelstud.io/. Acessado em março de 2022.

Figura 12 – Exemplo de anotação de documentos clínicos com substituições de dados sensíveis



Fonte: Adaptado de Deleger et al. (2014).

As anotações foram feitas por um único anotador, o autor deste trabalho. Embora a literatura sugira a presença de, pelo menos, dois anotadores, esse método não foi aplicado neste trabalho por limitações de confidencialidade e acesso aos dados do HC-UFMG/EBSERH. Múltiplas anotações de um mesmo *corpus* por anotadores diferentes fornecem os insumos necessários para a modelagem de dados. O modelo, no entanto, já foi testado e validado por Stubbs e Uzuner (2017).

Após o processo de anotação, os dados foram exportados num formato compatível com o ilustrado no texto desidentificado 5.2. Esse formato foi usado no processo de desidentificação discutido com mais detalhes na subseção 5.2.3.

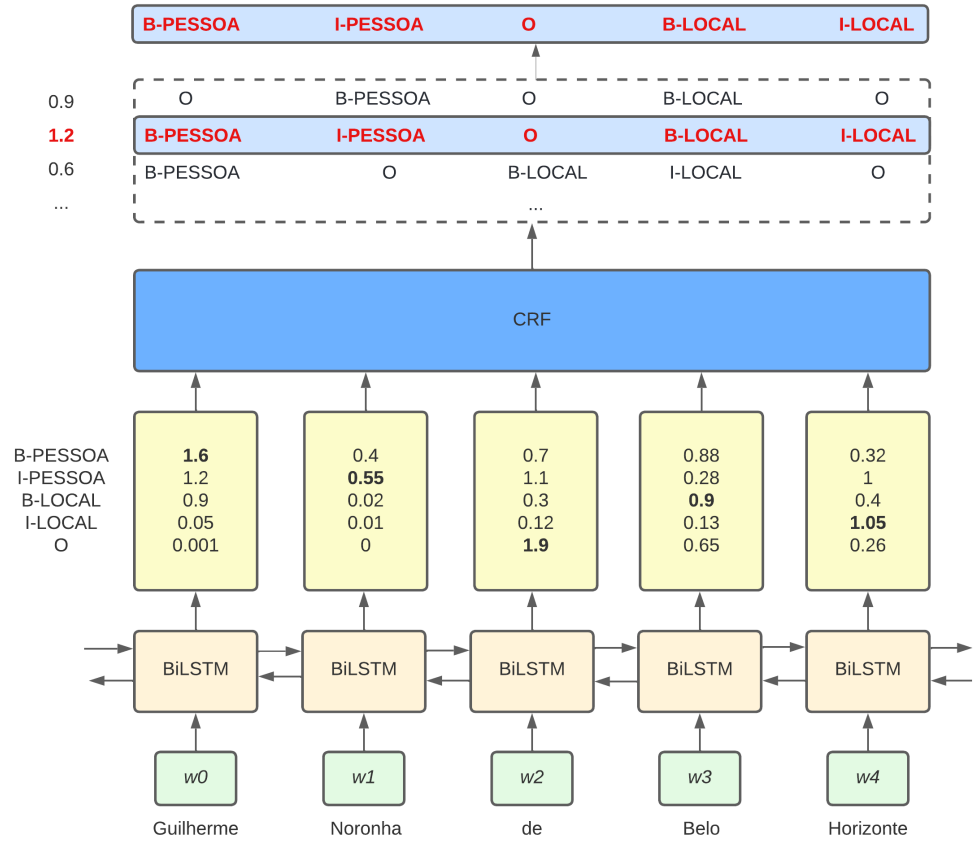
5.2.3 Desidentificação

O processo de desidentificação é automatizado e usa o aprendizado de máquina como recurso. A decisão é apoiada pela revisão apresentada no Capítulo 4, ao mostrar que as técnicas puramente baseadas em aprendizado de máquina possuem resultados superiores em relação às demais abordagens. Ainda citando o Capítulo 4, a técnica escolhida para implementação do algoritmo de desidentificação foi a RNN, mais precisamente a BI-LSTM-CRF, apresentada por Huang, Xu e Yu (2015). Esse algoritmo é uma variação bidirecional do CRF-LSTM apresentado no Capítulo 4 em que o aprendizado ocorre sequencialmente tanto do começo para o fim, quanto usando o caminho inverso.

A Figura 13 mostra a arquitetura por trás desse algoritmo. Nesse exemplo, são cinco categorias usadas para fazer anotações: *B-PESSOA*, *I-PESSOA*, *B-LOCAL*, *I-LOCAL* e

O. O algoritmo BI-LSTM é executado sobre uma sentença X, composta por palavras ou *tokens*, w_i. A sentença é processada pelo algoritmo partindo do começo da frase e depois pelo caminho contrário. O BI-LSTM calcula, então, as probabilidades de cada palavra w pertencer a uma determinada categoria. O resultado é usado na camada CRF responsável por adicionar restrições no aprendizado e garantir que as classificações estejam corretas. Como explicado no Capítulo 4, ambas as camadas desse algoritmo usam aprendizado de máquina para realizar as classificações.

Figura 13 – Arquitetura do BI-LSTM-CRF.



Fonte: Adaptado de CreateMoMo[3]

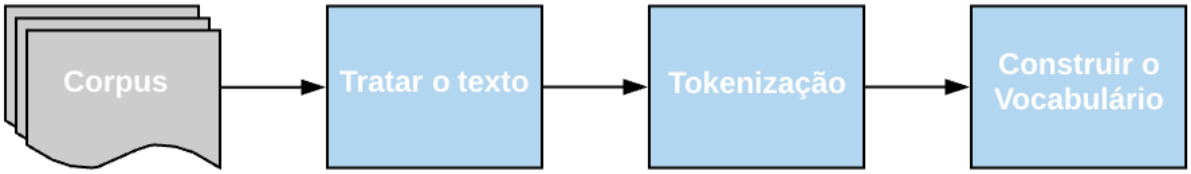
O algoritmo BI-LSTM-CRF é usado por meio da biblioteca *bi-lstm-crf 0.2.0*[4], disponibilizada para a linguagem de programação Python[5]. Para usar a biblioteca, o *corpus* de treinamento deve estar devidamente preparado e formatado. Para isso é necessário possuir três arquivos: "vocab.json", "tags.json" e "dataset.txt".

O arquivo "vocab.json" é composto por todas as palavras que definem o *corpus*. Para isso, deve se obedecer a um procedimento padrão de tratamento e preprocessamento de texto como mostra a Figura 14. A etapa de tratamento do texto envolve separa-lo em partes atômicas que são os *tokens*. Por fim, o vocabulário é construído a partir

[3] Disponível em https://createmomo.github.io/2017/09/12/CRF_Layer_on_the_Top_of_BiLSTM_1/. Acessado em março de 2022.
[4] Disponível em https://pypi.org/project/bi-lstm-crf/. Acessado em março de 2022.
[5] Disponível em https://www.python.org/. Acessado em março de 2022.

dos *tokens* separados e identificados na etapa anterior. Esse processo foi feito usando o Python e bibliotecas de suporte ao tratamento de palavras como o spaCy[6]. O resultado foi armazenado no arquivo “vocab.json”.

Figura 14 – Etapas de preprocessamento de texto antes de aplicações em PLN.



Fonte: elaborado pelo próprio autor

O arquivo “tags.json” é composto por todas as etiquetas usadas nas anotações apresentadas na Tabela 9. Cada etiqueta deve estar entre aspas e todas elas devem estar em colchetes e separadas por vírgulas como mostra o exemplo a seguir.

$$\text{["B-NOME", "I-NOME", "B-PROFISSAO", } \cdots \text{ , "B-ID", "I-ID", "O"]} \tag{5.3}$$

Por fim, o arquivo “dataset.txt” é composto pelo *corpus* anotado. Cada linha deste arquivo representa um documento do *corpus* e suas anotações no mesmo formato do exemplo 5.3. O exemplo a seguir ilustra como deve ser uma linha do arquivo “dataset.txt”.

$$\text{["Hector", "contraiu", "asma", "no", "dia", "12/08/2015"] ["B-NOME", "O", "O", "O", "O", "B-DATA"]} \tag{5.4}$$

O exemplo anterior demonstra com exatidão como o processo de anotação explicado na subseção 5.2.2 deve ser executado. Depois que o *corpus* estava devidamente formatado, o BI-LSTM-CRF foi aplicado sobre ele. O algoritmo processou os textos para classificar os *tokens* conforme as categorias da Tabela 9. As etiquetas servirão como um indicativo para ser a substituição futura de dados sensíveis. A *performance* de aprendizado do algoritmo foi medida pelo F-Score, a média harmônica entre precisão e revocação, calculada da seguinte maneira:

$$F - Score = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{5.5}$$

A taxa de F-Score ideal almejada por este trabalho segue as recomendações apresentadas no Capítulo 4, sugerindo que o valor de 95% é considerado aceitável para sistemas de desidentificação.

[6] Disponível em https://spacy.io/. Acessado em março de 2022.

5.3 Resumo da metodologia

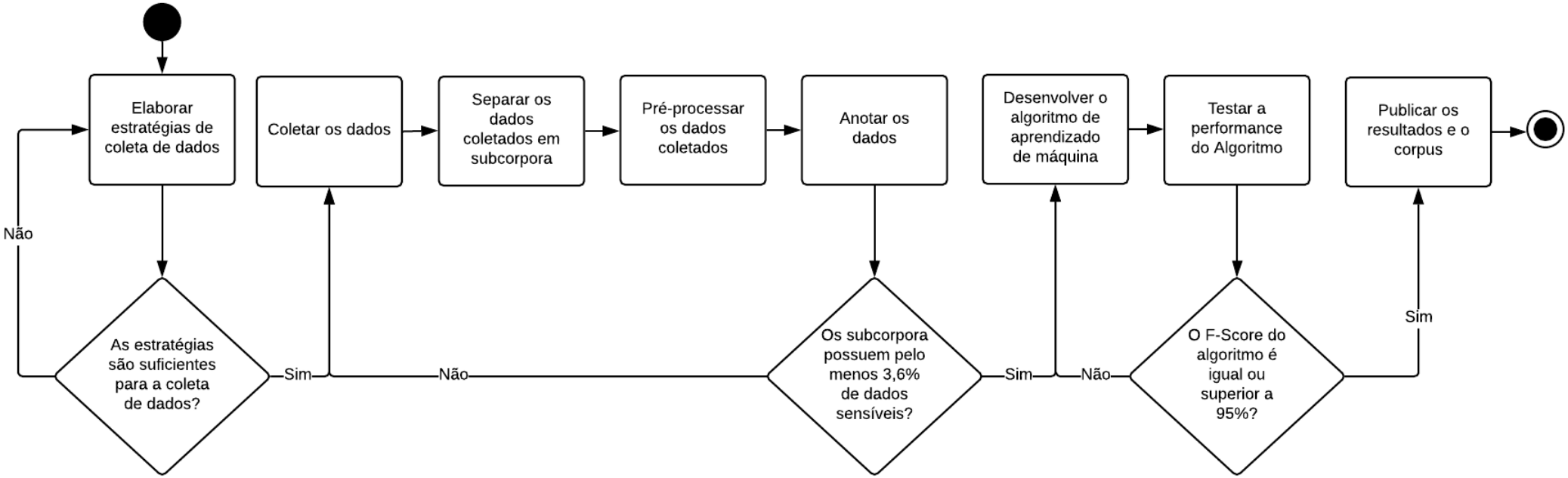
As seções anteriores do Capítulo 5 apresentaram detalhadamente a metodologia executada pelo presente projeto. Esta seção sintetiza os conceitos apresentados anteriormente e fornece uma visão geral do processo metodológico. A sintetização é mostrada melhor na Tabela 10 e na Figura 15.

A Tabela 10 mostra resumidamente as etapas de desenvolvimento para a metodologia proposta enquanto a Figura 15 ilustra as etapas principais de execução da metodologia e as principais tomadas de decisão.

Tabela 10 – Descrição das etapas da metodologia de pesquisa

Etapa	**Subetapa**	**Breve descrição da etapa**	**Objetivo específico**
Coleta de dados	Elaboração de estratégias de consultas ao banco de dados	Extrair o DER das principais tabelas e traçar planos de consultas eficientes para identificar que dados serão extraídos.	Construção do corpus.
	Extração de dados	Fazer uma consulta que retorne os dados e extraí-los para um arquivo de texto.	Construção do corpus.
	Separação dos *corpora*	Separar 45% dos dados para treinamento, 25% para teste e 30% para execução	Criação do algoritmo automatizado.
Construção do *corpus*	Preprocessamento	Remover ruídos do texto e prepará-lo para anotação.	Criação do algoritmo automatizado.
	Anotação	Colocar etiquetas em cada *token* nos *corpora* de treinamento e teste conforme a Tabela 9	Construção do corpus.
Desidentificação	Preparação do algoritmo	Desenvolver um código em Python, usando a biblioteca bi-lstm-crf, para processar os *corpora* anotados.	Criação do algoritmo automatizado.
	Testes	Verificar taxa de *performance* através de F-Score. Repetir a etapa de desidentificação caso os resultados não sejam alcançados	Criação do algoritmo automatizado.
Publicação	Publicação dos resultados	Divulgar os resultados da pesquisa em revistas e eventos científicos de interesse.	

Figura 15 – Fluxograma de execução da metodologia.



Fonte: elaborado pelo próprio autor

6 Resultados obtidos

O presente capítulo descreve cada etapa de execução da metodologia proposta no Capítulo 5. Para isso, foi dividido da seguinte maneira: a seção 6.1 descreve como foi feito o tratamento de dados e sua preparação até que os documentos estivessem prontos para anotação; a seção 6.2 descreve o processo de anotação realizado; a seção 6.3 mostra o processo de preparação dos dados anotados para execução pelo algoritmo de aprendizado; a seção 6.4 detalha a execução do algoritmo de aprendizado BI-LSTM-CRF; seção 6.5 mostra os resultados obtidos e o processo de extração necessário para obtê-los e a seção 6.6 mostra o processo de aprimoramento de resultados usando a técnica de busca em *grid*.

6.1 Tratamento de dados

Os dados coletados foram entregues num arquivo em formato de planilha eletrônica. A estrutura do arquivo traz os campos de maneira desordenada em que cada linha possui três colunas distintas: número do atendimento, campo e valor. A coluna número do atendimento é referente à chave primária do banco de dados e possui um valor único que identifica um documento clínico dos demais. A coluna campo é o nome de um dos campos listados na Tabela 8 e a coluna valor é o campo de texto em que os dados podem estar preenchidos ou não. São, ao todo, 1573 linhas divididas em duas planilhas, uma de documentos de admissão e outra para documentos de urgência, com um total de 49 719 *tokens*.

A primeira etapa para construção e estruturação do *corpus* é remover possíveis dados duplicados. O Código 6.1 detalha essa rotina. Após a remoção de dados duplicados, a quantidade de *tokens* presente no código caiu para 49 514.

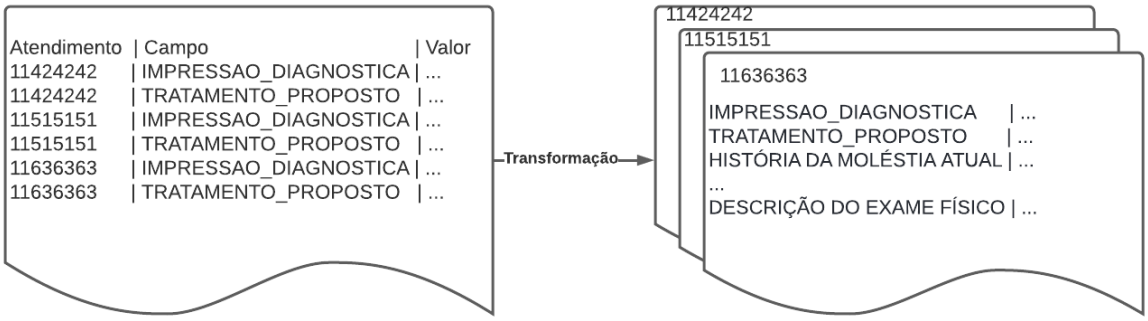
```
1 import pandas as pd
2
3 # Carrega a planilha eletronica dentro de um manipulador de tabelas chamado
      Pandas
4 dfs = []
5 xls = pd.ExcelFile('corpus/Dados.xlsx')
6
7 for sheet_name in xls.sheet_names:
8     dfs.append(pd.read_excel(xls, sheet_name))
9
10 df = pd.concat(dfs, ignore_index=True)
11
12 # Remove as linhas duplicadas
```

```
13 df.drop_duplicates(inplace=True)
```

Código 6.1 – Carga de dados e remoção de dados duplicados.

A próxima etapa para construção e estruturação do *corpus* é agrupar os dados coletados pelo número de atendimento. Essa etapa consiste em transformar os dados de forma que o conteúdo de cada prontuário esteja agrupado com seus respectivos campos. O arquivo que possui apenas uma tabela com múltiplas linhas é transformado em várias tabelas e cada uma delas corresponde a um prontuário diferente. Dessa maneira, a tarefa de anotação pode ser feita através de cada prontuário individualmente. A Figura 16 ilustra essa transformação enquanto o Código 6.2 mostra a rotina executada para que o objetivo fosse alcançado.

Figura 16 – Transformação dos dados coletados para múltiplos documentos.



Fonte: elaborado pelo próprio autor

```
1 df = df.pivot_table(index='atendimento', columns='campo', values='valor',
2                     aggfunc=lambda x: ' '.join(x.dropna()))
```

Código 6.2 – Transformação de dados.

O Código 6.2, embora pareça simples à primeira vista, executa uma série de transformações essenciais para que o formato dos dados seja agrupado em múltiplos documentos. A função `pivot_table` é usada gerar uma nova tabela cujos itens são agregados sob uma variável discreta, nesse caso a variável `atendimento`. As colunas dessa nova tabela são retiradas da variável `campo` e seus valores são retirados da variável `valor`. O parâmetro da função `aggfunc` atua para agregar possíveis entradas duplicadas que o conjunto de dados possa ter. Ele cria uma função temporária que junta os dados duplicados sem considerar possíveis campos vazios.

Após a transformação executada pelo Código 6.2, os dados foram agrupados em 198 linhas e 15 colunas. Cada linha representa um atendimento e um documento diferente. Já as colunas são os dados coletados em cada atendimento. Cada coluna é um dos campos

não estruturados mostrados na Tabela 8, num total de 14. O 15.º campo da tabela é o número do atendimento que atua como índice.

Uma análise dos dados foi feita após a transformação e observou-se que alguns campos estavam preenchidos com ruídos. Acredita que os profissionais da saúde preencheram determinados campos com caracteres simples, como o uso de "-", por exemplo, para atender a um prerrequisito de obrigatoriedade de preenchimento dos SGBDs. Para resolver esse problema, optou-se por remover esses ruídos e preservar somente o que é conteúdo de texto livre propriamente dito. Para isso o Código 6.3 foi executado.

```
df.replace(['-', '--', '---', '-----', ''], np.nan, inplace=True)
df.fillna('', inplace=True)
```

Código 6.3 – Remoção de ruídos.

Este código substituiu as ocorrências dos caracteres listados para uma variável nula. Em seguida, essa variável nula é substituída novamente por um campo vazio. Após a execução da transformação e remoção de ruídos, a quantidade de *tokens* dos dados foi reduzida para 49 268.

Após a execução de todas as etapas até então descritas, o preprocessamento de dados está concluído. A próxima etapa é salvar os dados preprocessados de duas maneiras: a primeira é num SGBD para garantir a persistência de dados e consultas futuras; a segunda é num formato de arquivo de texto que será usado como insumo para que os dados sejam carregados no software de anotação.

Para a persistência de dados, optou-se por armazená-los no MongoDB[1], um banco de dados orientado a documentos. A escolha se deu pelo tipo de documento manipulado e sua estrutura flexível. Bancos de dados orientados a documentos manipulam melhor esses tipos de dados do que os bancos de dados relacionais. A persistência de dados foi executada segundo o Código 6.4.

```
import json
from pymongo import MongoClient
from decouple import config

# Transforma o DF do Pandas num documento JSON
json_documents  = df.to_json(orient='index')
parsed = json.loads(json_documents)

# Recupera usuario e senha de uma fonte externa
u_name = config('MONGO_USERNAME', default='')
pass_d = config('MONGO_PASSWORD', default='')

# Conecta no banco de dados
client = MongoClient('mongo', 27017, username=u_name, password=pass_d)
```

[1] Disponível em https://www.mongodb.com/. Acessado em março de 2022.

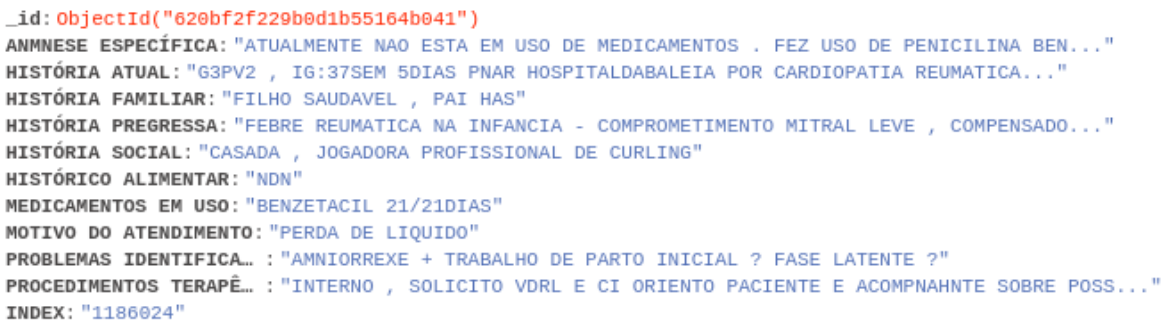
```
db = client['DocumentosClinicos']
collection = db['Admissao-Urgencia']

#Insere os documentos
collection.insert_many(parsed)
```

Código 6.4 – Ingestão de dados.

Os dados foram ingeridos num SGBD local por questões de segurança e apenas para demonstrar a persistência de dados. A Figura 17 mostra um dos documentos ingeridos no MongoDB. Nota-se que o referido documento possui apenas onze de quinze campos preenchidos. Neste caso os outros quatro campos não foram preenchidos pelos profissionais de saúde. O MongoDB, por ser um banco orientado a documentos, permite que cada registro em sua base tenha tamanho e formas flexíveis.

Figura 17 – Ingestão de dados no MongoDB.



Fonte: elaborado pelo próprio autor

Por fim, a última etapa do preprocessamento é armazenar os dados em arquivos de texto de forma que eles consigam ser carregados na ferramenta Label Studio para ser anotados. Essa etapa foi executada com o Código 6.5.

```
from spacy.lang.pt import Portuguese

# Inicializa o tokenizador para o Portugues
nlp = Portuguese()
tokenizer = nlp.tokenizer

# Loop que percorre todos os documentos agrupados
for row_index, row in df.iterrows():
    tokenized_fields = []
    for column_index, column in row.items():
        if (not pd.isnull(column)):
            # Tokeniza o texto e reagrupa ele novamente em sentencas.
            doc = tokenizer(column)
            tokenized_fields.append(f'{column_index}: ' + ' '.join(token.
    text for token in doc) + '\n\n')
```

```
15          # Salva o documento num arquivo de texto onde o nome e' o numero do
        atendimento.
16          with open(f'Preprocessed \textit{corpus}/{row_index}.txt', 'w',
      encoding="UTF-8") as f:
17              for tf in tokenized_fields:
18                  f.write(tf)
```

Código 6.5 – Criação de arquivos de textos.

O Código 6.5 utiliza do tokenizador em português do SpaCy para separar o texto antes de salvá-lo novamente no arquivo de texto. Essa etapa é importante para separar o texto em partes atômicas que serão anotadas. Após a tokenização, o *corpus* ficou com 65 005 *tokens*.

6.2 Anotação

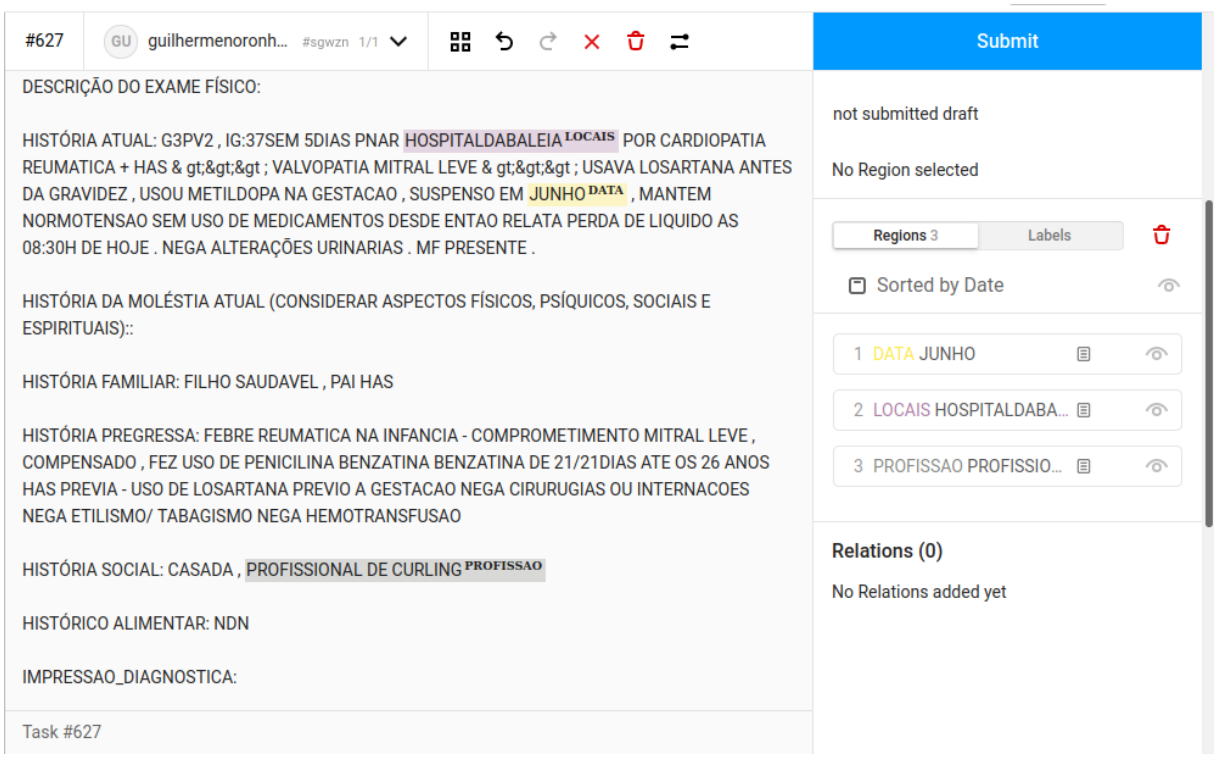
O processo de anotação dos dados foi executado no *software* Label Studio. Para isso foi necessário configurá-lo com as etiquetas definidas na Tabela 9. O Código 6.6 mostra as configurações necessárias para a utilização do Label Studio na anotação dos dados coletados neste trabalho.

```
1  <View>
2    <Labels name="label" toName="text">
3      <Label value="NOME" background="#ef2929"/>
4      <Label value="IDADE" background="#fcaf3e"/>
5      <Label value="DATA" background="#fce94f"/>
6      <Label value="CONTATO" background="#8ae234"/>
7      <Label value="IDENTIFICACAO" background="#729fcf"/>
8      <Label value="PROFISSAO" background="#888a85"/>
9      <Label value="LOCAIS" background="#ad7fa8"/>
10   </Labels>
11   <Text name="text" value="$ner"/>
12 </View>
```

Código 6.6 – Configuração de etiquetas para anotação de texto.

Os arquivos de textos gerados na etapa de preprocessamento, demonstrado na seção 6.1, foram carregados no Label Studio. A estrutura criada pelo Código 6.6 permitiu que uma tarefa fosse criada para cada arquivo diferente. Dessa maneira, o Label Studio criou 198tarefas distintas de anotação, cada uma delas contendo todo o texto dos documentos gerados. A tarefa de anotação foi feita de forma manual após a importação de dados. A Figura 18 ilustra um texto anotado.

Figura 18 – Texto anotado no Label Studio.



Fonte: elaborado pelo próprio autor

Algumas considerações devem ser feitas sobre o processo de anotação. Optou-se por anotar apenas as datas que fossem absolutas ou que indicassem algum tempo específico. Datas como “23/05/2025”, “23/05”, “junho” foram anotadas como dados sensíveis. Textos que fazem referências a datas passadas de forma relativa, como “há 2 anos”, “às 08:30h de hoje”, “em 6 meses” não foram anotados como dados sensíveis.

Optou-se por anotar apenas a idade atual das pessoas. Textos que indiquem um evento ocorrido com determinada idade foram anotados como dados sensíveis. Também optou-se por marcar qualquer tipo de dado sensível ainda que estivesse sob o uso de uma abreviação ou acrônimo.“HC” como acrônimo de “Hospital das Clínicas” e “CS Barreiro”, como abreviação de “Centro de Saúde Barreiro” foram marcados como dados sensíveis. Essas considerações podem ser anotadas e inseridas no Label Studio como normas de anotação nos casos em que mais de um anotador faça a anotação.

Quatro documentos foram excluídos durante o processo de anotação. Esses documentos não continham nenhum tipo de dado sensível e, portanto, não servem ao propósito desse trabalho. Após o término da etapa de anotação, os dados foram exportados para serem processados pelo algoritmo de aprendizado de máquina. O formato escolhido foi o CONLL2003, formato usado para tarefas de reconhecimento de entidade mencionada e também por ser facilmente manipulado. A Figura 19 exemplifica a estrutura do resultado da exportação.

Figura 19 – Trecho da exportação em CONLL2003 dos documentos anotados.

```
 1 -DOCSTART- -X- O
 2 HISTÓRIA -X- _ O
 3 DA -X- _ O
 4 MOLÉSTIA -X- _ O
 5 ATUAL -X- _ O
 6 (CONSIDERAR -X- _ O
 7 ASPECTOS -X- _ O
 8 FÍSICOS, -X- _ O
 9 PSÍQUICOS, -X- _ O
10 SOCIAIS -X- _ O
11 E -X- _ O
12 ESPIRITUAIS):: -X- _ O
13 PACIENTE -X- _ O
14 55 -X- _ B-IDADE
15 ANOS -X- _ I-IDADE
16 , -X- _ O
17 G2PN1A0 -X- _ O
18 , -X- _ O
19 25S+2D -X- _ O
20 PELO -X- _ O
21 US -X- _ O
22 DE -X- _ O
23 23/07/28 -X- _ B-DATA
```

Fonte: elaborado pelo próprio autor

A Figura 19 mostra como será a manipulação de dados para preparar o processo de aprendizagem de máquina. Na coluna mais à esquerda, tem-se o *token* que foi anotado, e na coluna mais à direita, o tipo de anotação que esse *token* recebeu. Na linha 23, por exemplo, o *token* “23/07/28” recebeu a categoria de anotação “B-DATA”.

De posse das anotações exportadas em CONLL2003, as próximas etapas de processamento são a preparação dos dados para serem executadados no algoritmo de aprendizado de máquina. Esses processos são discutidos na próxima seção.

6.3 Preparando dados para treinamento

A biblioteca de aprendizado de máquina exige que os dados estejam em três arquivos de textos diferentes:

1. dataset.txt: onde estão os dados e suas anotações;
2. tags.json: contém a lista de todas as etiquetas usadas na anotação;
3. vocab.json: dicionário de todas as palavras que aparecem no *corpus*.

A primeira etapa dessa preparação é carregar os *tokens* e suas respectivas *tags* no Python. Esse processo é dado pelo Código 6.7.

1

```
text = []
text_list = []
tags = []
tag_list = []

# Carrega o arquivo exportado do Label Studio
with open('Annotated\textit{corpus}/annotated-\textit{corpus}.conll', 'r',
    encoding='UTF-8') as f:
     full_text = f.readlines()

# Separa os tokens e as tags em variaveis diferentes.
for line in full_text[1:]:
    if line != '\n':
        words = line.split()
        tags.append(f'{words[3]}')
        if words[0] == '"':
            text.append(f'\\{words[0]}')
        elif words[0] == '\\':
            text.append(f'\\{words[0]}')
        else:
            text.append(f'{words[0]}')
    else:
            text_list.append(text)
            tag_list.append(tags)
            text = []
            tags = []
```

Código 6.7 – Carregando as anotações.

A próxima etapa é transformar essas variáveis carregadas em dois conjuntos diferentes de dados: o arquivo "dataset.txt", sendo o *subcorpus* de treinamento e validação e o arquivo "test.txt" sendo o *subcorpus* de teste. Nota-se que a biblioteca do BILSTM-CRF não requer que esse segundo arquivo seja criado. Ele será usado, no entanto, para testar o modelo após ser treinado.

O particionamento das bases de testes foi feito respeitando a metodologia e a Figura 8. A divisão foi de 70% para treinamento e 30% para teste. Os 25% relacionados para teste de *performance* citados na Figura 8 serão obtidos durante a execução do algoritmo. Esse percentual será retirado da base de treinamento. A separação dos corpora foi executada segundo Código 6.8

```
from sklearn.model_selection import train_test_split

# Criando as bases de treinamento e teste
X_text_train, X_text_test = train_test_split(text_list, test_size=0.30,
    random_state=42)
X_tag_train, X_tag_test = train_test_split(tag_list, test_size=0.30,
    random_state=42)
```

```
6
7 # Salvando a base de treinamento
8 with open ('BILSTM data/dataset.txt', 'w', encoding='UTF-8') as f:
9         for i in range(len(X_text_train)):
10             f.write('[%s]\t[%s]\n' % (', '.join(f'"{token}"' for token in
    X_text_train[i]),
11                                       ', '.join(f'"{tag}"' for tag in
    X_tag_train[i])))
12
13 # Salvando a base de teste
14 with open ('BILSTM data/test.txt', 'w', encoding='UTF-8') as f:
15         for i in range(len(X_text_test)):
16             f.write('[%s]\t[%s]\n' % (', '.join(f'"{token}"' for token in
    X_text_test[i]),
17                                       ', '.join(f'"{tag}"' for tag in
    X_tag_test[i])))
```

Código 6.8 – Criando os subcorpora de treinamento e teste.

Um ponto interessante a ser observado no Código 6.8 é a variável `random_state=42`. Em aprendizado de máquina, as variáveis escolhidas manualmente pelo desenvolvedor são chamadas hiperparâmetros. Esses hiperparâmetros são ajustáveis e sua escolha pelo desenvolvedor é arbitrária em busca da melhor solução. Neste caso, `random_state` determina como os dados serão aleatoriamente particionados. A escolha pelo número 42 foi arbitrária, podendo ser qualquer outro número, e não impacta no desenvolvimento do projeto.

A próxima etapa é criar os arquivos "tags.json" e "vocab.json". O primeiro arquivo é estático para essa proposta e baseia-se nas etiquetas usadas durante a etapa de anotação. O segundo arquivo é dinâmico e depende dos dados anotados e como foram divididos. Optou-se por montar o vocabulário apenas com as palavras presentes no *subcorpus* de treinamento. Assim, evita-se o viés de aprendizado em que o modelo conhece *tokens* que ele não viu durante a modelagem. Para executar essa etapa usou-se o Código 6.9

```
1 # Criando o arquivo tags.json
2 tags = ('B-NOME', 'I-NOME', 'B-IDADE' , 'I-IDADE', 'B-DATA' , 'I-DATA',
3         'B-CONTATO' , 'I-CONTATO', 'B-IDENTIFICAÇÃO' , 'I-IDENTIFICAÇÃO',
4         'B-PROFISSÃO' , 'I-PROFISSÃO',
5         'B-LOCAIS' , 'I-LOCAIS', 'O')
6 with open('BILSTM data/tags.json', 'w', encoding='UTF-8') as f:
7     f.write('[%s]' % (', '.join(f'"{tag}"' for tag in tags)) )
8
9 # Criando o arquivo vocab.json a partir da base de treinamento
10 vocab = []
11 for doc in X_text_train:
12     vocab.extend(doc)
13 vocab = list(dict.fromkeys(vocab))
14 with open('BILSTM data/vocab.json', 'w', encoding='UTF-8') as f:
```

```
15      f.write('[%s]' % (', '.join(f'"{token}"' for token in vocab)) )
```

Código 6.9 – Criando os subcorpora de treinamento e teste.

De posse dos arquivos necessários para a execução do algoritmo de aprendizagem, a próxima etapa é o aprendizado.

6.4 Executando o algoritmo de aprendizagem

A biblioteca escolhida para uso não requer implementações extras para a execução do modelo de aprendizagem. Basta ser chamada num terminal que contenha o Python instalado e que os arquivos necessários estejam no mesmo diretório em que a biblioteca é executada. O programa é chamado conforme o Código 6.10.

```
1 python -m bi_lstm_crf "BILSTM data" --model_dir "deid_clinical_docs" --
    num_epoch 2000 --save_best_val_model --val_split=0.25 --test_split=0
```

Código 6.10 – Executando o algoritmo de aprendizagem.

O Código 6.10 possui uma série de hiperparâmetros importantes para configurar o algoritmo de aprendizagem[2]. Eles são explicados a seguir:

- `--model_dir`: indica o nome do modelo;
- `--num_epoch`: o número de vezes (épocas) que o algoritmo será treinado para otimizar o aprendizado.
- `--save_best_val_model`: para salvar a melhor época que o algoritmo obteve o melhor resultado.
- `--val_split`: tamanho da base de validação.
- `--test_split`: tamanho da base de teste.

O número de épocas foi configurado para ser o maior possível. Haverá uma época em que o resultado deixará de ser otimizado e o modelo salvará esse resultado como sendo o melhor, ignorando as demais épocas. O parâmetro `--val_split` foi configurado para ser 25% da base respeitando as sugestões da Figura 8. O parâmetro `--test_split` foi configurado como 0, pois a base de teste já havia sido previamente separada. O Código 6.11 mostra o relatório parcial de execução do BI-LSTM-CRF.

```
1 config BILSTM data/vocab.json loaded
2 config BILSTM data/tags.json loaded
3 tag dict file => deid_clinical_docs/tags.json
```

[2] Mais informações sobre esses e outros hiperparâmetros em https://github.com/jidasheng/bi-lstm-crf/wiki/training-options. Acessado em março de 2022.

```
tag dict file => deid_clinical_docs/vocab.json
loading dataset BILSTM data/dataset_cache_100.npz ...
datasets loaded:
        train: torch.Size([102, 100]), torch.Size([102, 100])
        val: torch.Size([33, 100]), torch.Size([33, 100])
        test: torch.Size([0, 100]), torch.Size([0, 100])
 1/2000 loss: 400.77, val_loss:  0.00: 100% 1/1 [00:01<00:00,  1.53s/it]
eval: 100% 1/1 [00:00<00:00,  9.47it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 0) => deid_clinical_docs/loss.csv
 2/2000 loss: 395.78, val_loss: 392.91: 100% 1/1 [00:01<00:00,  1.39s/it]
eval: 100% 1/1 [00:00<00:00, 17.83it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 1) => deid_clinical_docs/loss.csv
 3/2000 loss: 390.83, val_loss: 388.18: 100% 1/1 [00:01<00:00,  1.44s/it]
eval: 100% 1/1 [00:00<00:00, 26.17it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 2) => deid_clinical_docs/loss.csv
 4/2000 loss: 385.87, val_loss: 383.44: 100% 1/1 [00:01<00:00,  1.29s/it]
eval: 100% 1/1 [00:00<00:00, 22.35it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 3) => deid_clinical_docs/loss.csv
 5/2000 loss: 380.83, val_loss: 378.62: 100% 1/1 [00:01<00:00,  1.45s/it]
eval: 100% 1/1 [00:00<00:00, 19.56it/s]
........
290/2000 loss:  2.27, val_loss: 13.56: 100% 1/1 [00:00<00:00,  1.03it/s]
eval: 100% 1/1 [00:00<00:00, 27.80it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 289) => deid_clinical_docs/loss.csv
291/2000 loss:  2.24, val_loss: 13.56: 100% 1/1 [00:00<00:00,  1.10it/s]
eval: 100% 1/1 [00:00<00:00, 18.18it/s]
save model => deid_clinical_docs/model.pth
save model(epoch: 290) => deid_clinical_docs/loss.csv
292/2000 loss:  2.22, val_loss: 13.56: 100% 1/1 [00:01<00:00,  1.30s/it]
eval: 100% 1/1 [00:00<00:00, 21.62it/s]
293/2000 loss:  2.20, val_loss: 13.56: 100% 1/1 [00:01<00:00,  1.27s/it]
eval: 100% 1/1 [00:00<00:00, 21.75it/s]
294/2000 loss:  2.17, val_loss: 13.56: 100% 1/1 [00:01<00:00,  1.22s/it]
eval: 100% 1/1 [00:00<00:00, 21.79it/s]
......
2000/2000 loss:  0.03, val_loss: 19.37: 100% 1/1 [00:00<00:00,  1.01it/s]
eval: 100% 1/1 [00:00<00:00, 22.33it/s]
```

Código 6.11 – Criando os subcorpora de treinamento e teste.

O Código 6.11 fornece muitos detalhes da tarefa de aprendizagem. Um desses detalhes é a variável `val_loss`, que mede o quão longe o aprendizado está da resposta ótima. Na prática, quanto menor esse valor, melhor o modelo. Na época 1, esse valor é 0,

pois nenhum aprendizado ainda foi realizado e não há parâmetros para comparação. A partir da época 2, esse valor é estipulado e a cada época o algoritmo tenta otimizar esse valor.

Observa-se que, na época 291, o algoritmo atinge seu pico de otimização e salva o modelo pela última vez obtendo um `val_loss` de 13,56. Ou seja, para essa execução, o BI-LSTM-CRF levou 291 épocas para ajustar o seu modelo ao máximo. Na época 2000, o `val_loss` era de 19,37.

6.5 Extraindo os resultados

De posse do modelo de aprendizado treinado e salvo, foi possível realizar testes e extrair métricas de análise. Para isso, o Código 6.12 extraiu o F-Score e montou uma matriz de confusão para análise.

```
from sklearn.metrics import f1_score, confusion_matrix
from bi_lstm_crf.app import WordsTagger

test_tokens = []
test_tags = []

for i in range(len(X_text_test)):
    test_tokens.append(X_text_test[i])
    test_tags.append(X_tag_test[i])

y_true = []
y_pred = []
for i, token in enumerate(test_tokens):
    true_tags = tags_to_numbers(tags, test_tags[i])
    y_true += true_tags
    model = WordsTagger(model_dir='deid_clinical_docs')
    pred_tags, _ = model([token])
    pred_tokens = tags_to_numbers(tags, pred_tags[0])
    y_pred += pred_tokens

f1 = f1_score(y_true, y_pred, average='micro')
matrix  = confusion_matrix(y_true, y_pred)
```

Código 6.12 – Executando o algoritmo de aprendizagem.

O F-Score obtido como resultado do treinamento e teste desse algoritmo de aprendizado foi de 97,65% usando a opção `micro` para calcular o F-Score. Essa opção calcula o F-Score de forma global. Usando a opção `macro`, em que o F-Score de cada categoria é computado separadamente e depois tira-se a média de todos, o valor obtido foi de 36,31%. A Tabela 11 mostra as características do *corpus* de teste em que os testes ocorreram.

Tabela 11 – Características do *corpus* de teste

Característica	**Frequência Absoluta**	**Frquência Relativa**
Total de *tokens*	22200	100%
Tokens sensíveis	709	3,19%
B-NOME	10	0,045%
I-NOME	3	0,013%
B-IDADE	64	0,288%
I-IDADE	62	0,0279%
B-DATA	400	1,8%
I-DATA	9	0,04%
B-CONTATO	0	0%
I-CONTATO	0	0%
B-IDENTIFICAÇÃO	0	0%
I-IDENTIFICAÇÃO	0	0%
B-PROFISSÃO	14	0,063%
I-PROFISSÃO	12	0,054%
I-LOCAIS	72	0,324%
I-LOCAIS	63	0,283%
O	21491	96,8%

A matriz de confusão, Tabela 12, é um método de análise de dados em forma de tabela que permite a interpretação do desempenho de um algoritmo de aprendizado supervisionado. Cada linha da matriz representa a etiqueta correta de classificação enquanto as linhas representam a predição que foi feita pelo algoritmo. A diagonal principal, em negrito, indica as predições que foram feitas corretamente enquanto as demais células representam as predições erradas (POWERS, 2020). Nota-se que o cabeçalho superior teve alguns caracteres suprimidos por conta de espaço, mas seus valores são os mesmos do cabeçalho a esquerda. Nota-se também que as etiquetas de contato e identificação foram suprimidas, pois, não foram encontradas ocorrências no *corpus*.

Tabela 12 – Matriz de confusão gerada pelas predições feitas pelo BILSTM-CRF.

	B-NOM	**I-NOM**	**B-IDA**	**I-IDA**	**B-DAT**	**I-DAT**	**B-PRO**	**I-PRO**	**B-LOC**	**I-LOC**	**O**
B-NOME	0	0	0	0	0	0	0	0	0	0	10
I-NOME	0	0	0	0	0	0	0	0	0	0	3
B-IDADE	0	0	**43**	0	0	0	0	0	0	0	21
I-IDADE	0	0	0	**51**	0	0	0	0	0	0	11
B-DATA	0	0	0	0	**59**	0	0	0	0	0	341
I-DATA	0	0	0	0	5	0	0	0	0	0	4
B-PROFISSÃO	0	0	0	0	0	0	0	0	0	0	14
I-PROFISSÃO	0	0	0	0	0	0	0	0	0	0	12
B-LOCAIS	0	0	0	3	0	0	0	0	**34**	0	35
I-LOCAIS	0	0	0	1	0	0	0	0	1	**17**	44
O	0	0	0	0	14	0	0	0	1	0	**21476**

6.6 Sintonizando os hiperparâmetros para aprimorar os resultados

O resultado apresentado na seção 6.5 foi a primeira execução usando os hiperparâmetros padrões da biblioteca, com exceção do número de épocas configurado em 2000. Esta seção descreve uma tentativa de aprimorar o resultado obtido por uma técnica de otimização chamada busca em *grid*, do inglês *grid search*. Essa técnica consiste em criar combinações possíveis de um conjunto discreto de hiperparâmetros para execução (BURKOV, 2019).

Esse conjunto foi definido arbitrariamente pelo autor. As escolhas tiveram como base a limitação de tempo e os parâmetros padrão já conhecidos. O fator tempo limitou a quantidade de valores discretos, pois mais quantidade de valores traria um aumento significativo no número de combinações. Os valores escolhidos foram:

- *embedding_dim*: transforma cada *token* do *corpus* num vetor de dimensão n. Os valores escolhidos para esse parâmetro foram 100 e 250;
- *batch_size*: tamanho do bloco de *tokens* que serão processados de uma só vez. Os valores escolhidos para esse parâmetro foram 1024 e 2048;
- *hidden_dim*: quantidade de neurônios que cada camada da rede terá. Os valores escolhidos para esse parâmetro foram 128 e 256;
- *epochs*: número de vezes que o algoritmo de aprendizagem ajustará o modelo em busca do melhor resultado. Os valores escolhidos para esse parâmetro foram 300 e 600;
- *lr*: determinação da taxa de aprendizado que o modelo muda de uma época para outra. Os valores escolhidos para esse parâmetro foram 0,001 e 0,1;
- *num_rnn_layer*: quantidade de camadas que a rede neural terá. Os valores escolhidos para esse parâmetro foram 1 e 2;
- *max_seq_len*: controle do tamanho máximo da sequência de *tokens* que o modelo treinará por vez. Os valores escolhidos para esse parâmetro foram 100 e 150;
- *weight_decay*: contrapesos que evitam que o modelo generalize bem a base treino e performe mal a base de teste. Os valores escolhidos para esse parâmetro foram 0 e 0,3.

O conjunto dos valores escolhidos para cada hiperparâmetro do modelo de aprendizagem permitiu um total de 256 combinações diferentes de modelos de treinamento. A implementação da busca em *grid* é mostrada no Código 6.13.

```
embedding_dim = [100, 250]
batch_size = [1024, 2048]
hidden_dim = [128, 256]
epochs = [300, 600]
lr = [0.001, 0.1]
num_rnn_layer = [1, 2]
max_seq_len = [100, 150]
weight_decay = [0, 0.3]

for e in embedding_dim:
    for b in batch_size:
        for h in hidden_dim:
            for ep in epochs:
                for l in lr:
                    for n in num_rnn_layer:
                        for m in max_seq_len:
                            for w in weight_decay:
                                bash_command = f'bash bilstm.sh {e} {b} {h} {ep} {l} {n} {m} {w}'
                                process = subprocess.Popen(bash_command.split(), stdout=subprocess.PIPE)
                                output, error = process.communicate()
```

Código 6.13 – Uso da busca em *grid* para aprimorar os hiperparâmetros do modelo.

Após a execução das 256 combinações diferentes, a busca em *grid* conseguiu aprimorar o F-Score `micro` de 97,94% e F-Score `macro` de 39,83%. A matriz de confusão está disposta na Tabela 13 e os parâmetros selecionados estão dispostos na Tabela 14.

Tabela 13 – Matriz de confusão gerada resultante da otimização feita pela busca em *grid*.

	B-NOM	**I-NOM**	**B-IDA**	**I-IDA**	**B-DAT**	**I-DAT**	**B-PRO**	**I-PRO**	**B-LOC**	**I-LOC**	**O**
B-NOME	0	0	0	0	0	0	0	0	0	0	10
I-NOME	0	0	0	0	0	0	0	0	0	0	3
B-IDADE	0	0	**48**	0	0	0	0	0	0	0	16
I-IDADE	0	0	0	**48**	0	0	0	0	0	0	14
B-DATA	0	0	0	0	**87**	0	0	0	1	0	312
I-DATA	0	0	0	0	4	0	0	0	0	0	5
B-PROFISSÃO	0	0	0	0	0	0	0	0	0	0	14
I-PROFISSÃO	0	0	0	0	1	0	0	0	0	0	11
B-LOCAIS	0	0	0	2	0	0	0	0	**40**	0	30
I-LOCAIS	0	0	0	1	0	0	0	0	2	**29**	32
O	0	0	2	2	24	0	0	0	3	3	**21457**

6.7 Discussões sobre os resultados

A primeira execução do algoritmo de aprendizado obteve o F-Score de 97,65%, resultado esperado pela metodologia proposta. Ao calcular o F-Score usando-se a opção `macro`, o resultado cai para 36,31%, bem abaixo do valor proposto pela metodologia.

Tabela 14 – Melhor seleção de parâmetros na busca em *grid*.

Nome do parâmetro	**Valor escolhido**
embedding_dim	250
batch_size	2048
hidden_dim	256
epochs	600
lr	0.1
num_rnn_layer	1
max_seq_len	150
weight_decay	0

A diferença entre os dois métodos de cálculo de F-Score está no peso que eles dão para cada categoria. O cálculo `micro` usa todos os *tokens* igualmente para obter o F-Score. Ou seja, se existe uma quantidade muito superior de *tokens* de uma determinada categoria em relação às demais, essa categoria terá um peso muito maior sobre o F-Score. Já o cálculo *macro* calcula o F-Score de cada categoria separadamente e depois faz-se a média dos resultados obtidos. Esta abordagem trata cada categoria igualmente, dando-se a mesma importância para cada uma. Os trabalhos apresentados na literatura do Capítulo 4 não realizam uma distinção clara de como o F-Score é calculado. Alguns trabalhos detalham melhor essa etapa enquanto outros generalizam mais. Cabe aos leitores a interpretação desses resultados.

Para analisar mais a fundo os resultados obtidos, gerou-se uma matriz de confusão. Esta matriz é ideal para ver como a classificação dos dados foi feita pelo algoritmo de aprendizado. Essa análise traz uma noção se a desidentificação está ocorrendo eficazmente ou não.

Ao analisar a matriz de confusão na Tabela 12, observou-se que somente 204 de 709 *tokens* foram corretamente desidentificados em suas respectivas categorias. Isso representa 28,77% dos dados sensíveis. Outra abordagem de análise seria considerar a tarefa de desidentificação puramente simples. Nesse caso, observa-se a quantidade de dados que foram desidentificados corretamente, independentemente, se a categoria estava correta ou não. Essa abordagem desidentificou 214 de 709 *tokens* ou 30,18%. Ainda analisando a matriz de confusão, observou-se que 495 *tokens* ou 69,81% não foram desidentificados. Numa situação hipotética em que o conteúdo desse *corpus* pudesse ser disponibilizado ao público, falhas graves de vazamento estariam sendo cometidas.

Numa tentativa de melhorar os resultados preliminares, tentou-se modificar os hiperparâmetros do algoritmo de aprendizagem. Para isso, adotou-se a técnica de busca em *grid*, em que uma série de hiperparâmetros são combinados e testados em diferentes execuções. Foram ao todo 256 combinações diferentes para gerar o melhor resultado que

foi um F-Score `micro` de 97,94% e `macro` de 39,83%.

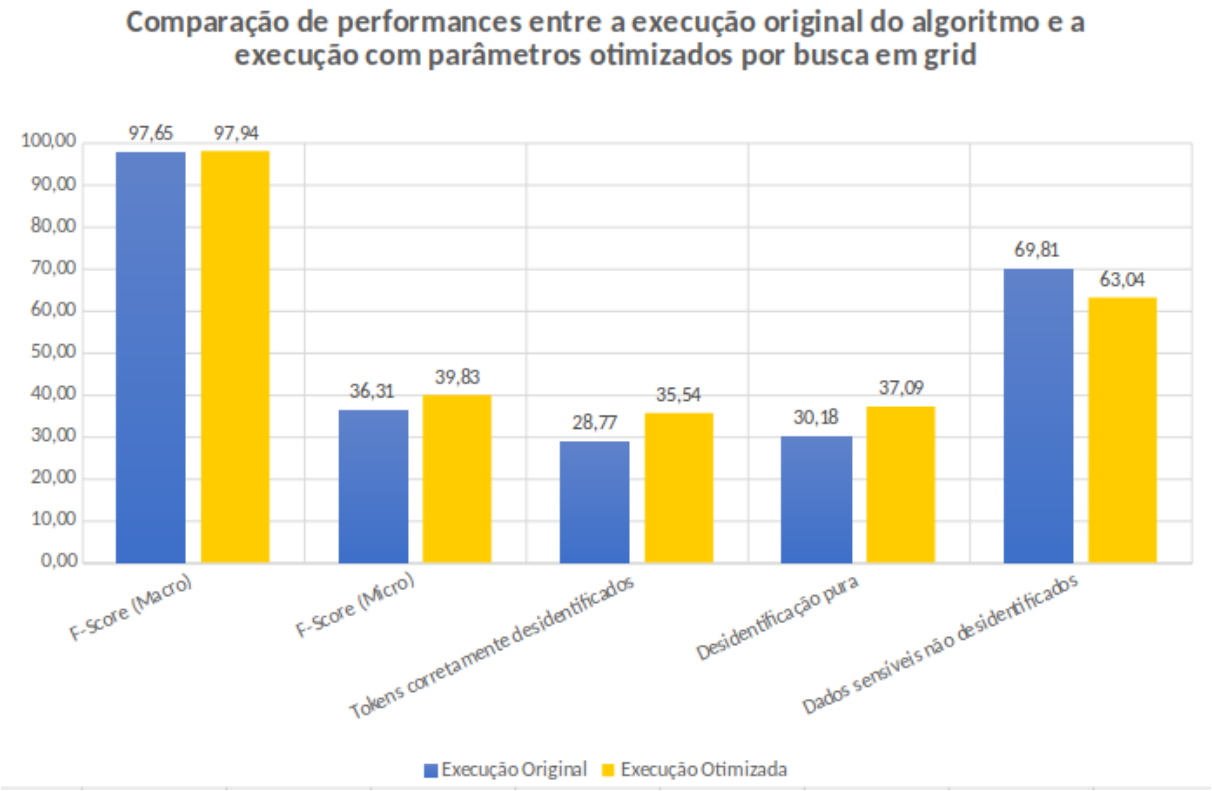
A matriz de confusão gerada para essa melhoria, mostrada na Tabela 13, mostra que 252 de 709, ou 35,54%, *tokens* foram corretamente desidentificados. Considerando a abordagem de desidentificação puramente simples, obtém-se 263 *tokens* ou 37,09%. A quantidade de *tokens* que não foram corretamente desidentificados é de 447 ou 63,04%. A Tabela 15 condensa as matrizes de confusão para comparação.

Tabela 15 – Matrizes de confusão condensadas. Na diagonal de baixo, o primeiro resultado e na diagonal de cima, o resultado gerado pela otimização da busca em *grid*.

Matriz original \ **Matriz otimizada**	**B-NOM**	**I-NOM**	**B-IDA**	**I-IDA**	**B-DAT**	**I-DAT**	**B-PRO**	**I-PRO**	**B-LOC**	**I-LOC**	**O**
B-NOME	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	10 10
I-NOME	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	3 3
B-IDADE	0 0	0 0	**48** **43**	0 0	0 0	0 0	0 0	0 0	0 0	0 0	16 21
I-IDADE	0 0	0 0	0 0	**48** **51**	0 0	0 0	0 0	0 0	0 0	0 0	14 11
B-DATA	0 0	0 0	0 0	0 0	**87** **59**	0 0	0 0	0 0	1 0	0 0	312 341
I-DATA	0 0	0 0	0 0	0 0	4 5	0 0	0 0	0 0	0 0	0 0	5 4
B-PROFISSÃO	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	14 14
I-PROFISSÃO	0 0	0 0	0 0	0 0	1 0	0 0	0 0	0 0	0 0	0 0	11 12
B-LOCAIS	0 0	0 0	0 0	2 3	0 0	0 0	0 0	0 0	**40** **34**	0 0	30 35
I-LOCAIS	0 0	0 0	0 0	1 1	0 0	0 0	0 0	0 0	2 1	**29** **17**	32 44
O	0 0	0 0	2 0	2 0	24 14	0 0	0 0	0 0	3 1	3 0	**21457** **21476**

O ganho na perspectiva `macro` foi de 3,5%. Cada execução do algoritmo teve um tempo médio de execução de 20 minutos num computador com 16GB de memória RAM e um processador Core i5-3570K (4 CPUs com 3.4GHz de frequência). As 256 execuções tiveram um tempo aproximado de 85 horas. Do ponto de vista de recursos computacionais e considerando que esses testes rodariam num ambiente de produção, a busca em *grid* não produziu resultados significativos para o modelo. A Figura 20 faz um comparativo entre as *performances* obtidas.

Figura 20 – Comparação de *performances* entre a execução original do algoritmo e a execução com hiperparâmetros otimizados por busca em *grid*.



Fonte: elaborado pelo próprio autor

A busca em *grid*, no entanto, teve um papel importante na análise dos resultados ao chegar a duas conclusões mostradas a seguir. Essas conclusões e outras sugestões de aprimoramento deste trabalho são discutidas na seção 7.3:

1. a escolha dos hiperparâmetros para o conjunto de dados utilizado pouco teve efeito no resultado obtido. Nesse caso, conclui-se que o *corpus* criado para esse problema foi insuficiente, seja pela quantidade de dados coletados, pela densidade de dados sensíveis encontrados ou pelas abordagens de preprocessamentos adotadas;

2. os hiperparâmetros não foram escolhidos adequadamente. Nesse caso, seria necessário executar mais testes utilizando novas combinações de hiperparâmetros.

A tarefa de desidentificação de dados sensíveis pode ser entendida como uma tarefa de nível crítico. Ao fazer essa afirmação, é necessário entender que os dados sensíveis precisam ser desidentificados a qualquer custo. Ou seja, não há espaço para algoritmos onde dados sensíveis não são devidamente desidentificados. Portanto, esse trabalho optou por considerar o F-Score `macro` como resultado, tornando os frutos desta pesquisa insuficientes para uma generalização.

Há ainda que debater se os 95% de F-Score adotados pela literatura como número aceitável de pesquisa é suficiente para a tarefa de desidentificação de dados sensíveis. Um algoritmo que alcançasse 97, 98 ou 99% de dados sensíveis desidentificados ainda teria problemas com dados revelados. A liberação desses dados ainda careceria de uma revisão manual.

Isso leva à discussão inicial deste projeto sobre a possibilidade de disponibilizar dados de saúde publicamente num formato de *corpus* para uso secundário. Os resultados levantados nesta pesquisa mostram que isso é uma tarefa muito difícil, embora não impossível. Acredita-se que disponibilizar os dados via um processo completamente automatizado seja um objetivo ambicioso demais. Isso demandaria a construção de um algoritmo de desidentificação infalível. Embora não seja tarefa impossível, demandaria esforços corporativos intensos, como ter equipes especializadas em aprendizado de máquina, anotação de dados, além de quantidade massiva de dados para treinamento e teste. Outra possibilidade mais acessível é disponibilizar esses dados de forma semiautomática. Nessa possibilidade, o algoritmo de aprendizado desidentificaria os dados que passariam por uma curadoria humana antes de sua disponibilização pública. O algoritmo atuaria como um sistema de apoio ao profissional. Para tal, seria aceitável uma *performance* de 95% ou mais de F-Score `macro`.

7 Conclusões e trabalhos futuros

Este capítulo conclui o presente trabalho. Através dos resultados obtidos no Capítulo 6, discutem-se aqui as conclusões e apresentam-se ideias para futuras pesquisas. Para isso optou-se por dividir o capítulo da seguinte maneira: a seção 7.1 apresenta as dificuldades encontradas na execução deste projeto; a seção 7.2 apresenta as contribuições geradas por esta pesquisa; a seção 7.3 fornece algumas sugestões de trabalhos futuros a partir desta pesquisa e na seção 7.4 são dados alguns comentários finais.

7.1 Problemas e dificuldades encontrados

Para discutir algumas seções deste capítulo optou-se por falar primeiramente dos problemas e dificuldades encontrados. Não que eles sejam maiores que o restante do capítulo. Porém, algumas discussões farão mais sentido se o leitor estiver de posse de alguns percalços encontrados na execução deste projeto.

O primeiro e mais importante deles é a obtenção de um conjunto de dados para a execução do trabalho. Durante a revisão de literatura, advogou-se a coleta representativa de documentos clínicos para que o problema de desidentificação na área da saúde pudesse ser amplamente atacado. Na metodologia, foi preciso realizar um recorte e a escolha mais sugerida na revisão de literatura seria o sumário de alta ou o sumário de admissão. Durante a execução da metodologia, no entanto, teve-se que trabalhar com evolução médica de emergência e avaliação à admissão hospitalar. O uso desses documentos se deu pela incapacidade de obter os dados almejados durante a elaboração da metodologia.

Embora o projeto tenha contado com a aprovação do COEP, isso não colaborou muito para a obtenção de dados. No caso do Apêndice A, o parecer deu acesso ao setor de tecnologia do Hospital das Clínicas, que era responsável pelo acesso e cuidado aos dados. Isso não foi suficiente, pois os funcionários alegaram diversas vezes indisponibilidade para realizar a extração de dados. Um dos problemas alegados por esses profissionais é a falta de conhecimento de como o sumário de alta é construído no sistema. Segundo eles, o sistema de informação é terceirizado, sendo o hospital proprietário apenas dos dados. Qualquer intervenção no sistema deve ser comunicada à empresa terceirizada através de um chamado ao suporte. Deu-se a entender que esses chamados são serviços pagos pelo hospital.

A solução oferecida pela equipe do hospital seria elaborar a própria consulta do sumário de alta em busca dos campos que seriam desidentificados. Surge então outro problema. Os profissionais que deram acesso à base de dados do hospital não sabiam informar quais seriam os campos ou as consultas a serem feitas no banco de dados. Isso

teria de ser descoberto por tentativa e erro. Uma tentativa foi feita e documentada na subseção 5.2.1. Detectaram-se mais de 3000 tabelas sem uma perspectiva clara de quais seriam as necessárias para elaborar a consulta certa. Depois de um processo de engenharia reversa para tentar montar um DER das tabelas do hospital, documento esse inexistente e/ou inacessível, chegou-se à conclusão de que o processo era inviável para este projeto.

Optou-se, então, por tentar obter esses dados de outra fonte, como mostra o Apêndice B. Dessa vez foi escolhido um hospital particular cujos sistema e dados são de sua propriedade. A dificuldade encontrada aqui foi outra. Embora o projeto tenha sido devidamente aprovado, a diretoria do hospital vetou a coleta de dados. A justificativa dada foi o risco de vazamento de dados sensíveis durante a pesquisa. A prerrogativa usada é que a palavra final é do hospital. Ou seja, de nada valeu todo o trâmite requisitado em ambos os COEPs.

Todo esse processo tomou cerca de dois anos de pesquisa. Durante esse tempo, protótipos da metodologia foram construídos. Algumas tarefas-chave como anotação e preprocessamento não puderam ser adiantadas, pois isso é uma característica peculiar de cada conjunto de dados. Sobrou então menos tempo do que o desejado para a execução adequada da metodologia proposta. O trabalho, no entanto, foi executado, seus resultados discutidos na seção 6.7 e as contribuições e trabalhos futuros são discutidas nas seções seguintes.

7.2 Contribuições do trabalho

Embora os resultados obtidos pela metodologia do trabalho não tenham sido considerados satisfatórios, acredita-se que o esse produziu algumas contribuições significativas para a área de modo geral. O trabalho nasceu a partir da perspectiva de que a produção científica de PLN para a língua portuguesa na área de saúde representa apenas 0,3% de toda pesquisa feita mundialmente. A discussão do Capítulo 2 trouxe à tona a importância dos corpora de documentos clínicos ao redor do mundo como fonte de uso secundário da informação. Constatou-se também que não existem fontes dessa natureza para o português do Brasil. Sob essa perspectiva, buscou-se identificar como esse problema poderia ser atacado.

Uma das hipóteses levantadas pelo trabalho seria a disponibilização de mais bases de textos, ou corpora, para serem usados como instrumentos de pesquisa. Em outras palavras, fornecer dados para uso secundário da informação. Um dos desafios para disponibilização de dados é a proteção de dados sensíveis. Novas regulamentações estão surgindo ao redor do mundo com a preocupação de proteger o cidadão de ter seus dados expostos publicamente. Para isso, o trabalho esclareceu os pontos de interesse para proteger os dados sensíveis em documentos clínicos. Num contexto de *big data* onde quantidades massivas de dados

são geradas diariamente, a desidentificação manual de documentos é uma tarefa fora de questão O Capítulo 4, então, esclareceu quais são as técnicas mais utilizadas na literatura para fazer a desidentificação automática de dados. Aqui surge o profissional da informação com conhecimentos em computação, estatística e linguística para atuar na produção de documentos desidentificados prontos para serem consumidos em tarefas de uso secundário da informação.

Essas discussões levaram à primeira contribuição do trabalho, a metodologia da pesquisa. Essa metodologia detalha como o objeto de pesquisa deve ser escolhido, a coleta de dados deve ser efetuada, como construir o *corpus*, como anotá-lo e a técnica de aprendizado a ser utilizada. A Figura 15 trouxe uma síntese da metodologia que pode ser replicada, não apenas para documentos clínicos, mas para qualquer documento que necessite da remoção de dados sensíveis.

A outra contribuição desse trabalho é a documentação da execução da metodologia. Embora o *corpus* não possa ser compartilhado, todo o código utilizado para execução do trabalho, bem como o modelo de aprendizado gerado estão disponibilizados publicamente[1]. A disponibilização do código utilizado no trabalho permitirá que outros pesquisadores e profissionais usem, adaptem, aprimorem e/ou apliquem o presente trabalho em outras áreas do conhecimento e/ou outros tipos de documentos clínicos. O modelo de aprendizado disponibilizado já está pronto para uso e pode ser aplicado tanto em documentos clínicos quanto noutros tipos de documentos. Sua capacidade de desidentificar pode ser testada em diferentes tipos de documentos sem a necessidade de treinamento prévio. Da mesma maneira que o código pode ser aperfeiçoado, assim também é o modelo de aprendizado.

Por fim, uma contribuição mais do ponto de vista da discussão acadêmica são as possibilidades a serem exploradas na aplicação de PLN em documentos clínicos para uso secundário da informação. As unidades de saúde têm em mãos uma fonte de informação valiosa que está sendo desperdiçada nos bancos de dados. Tal informação tem potencial de produzir pesquisa de ponta e avanços científicos na área de saúde. É preciso propagar a mensagem de que o compartilhamento dessa informação é beneficial para a população. Há de se buscar motivação interna nessas unidades.

Acredita-se que a iniciativa só prosperará se o interesse partir internamente nas unidades de saúde. As dificuldades relatadas na seção 7.1 mostraram o quão difícil é trabalhar nessa ideia de maneira independente. Como dito no Capítulo 4, a desidentificação é uma tarefa que sofre com o dilema da causalidade. Dados não podem ser desidentificados sem que se tenha acesso a eles e não se pode ter acesso aos dados sem que eles estejam devidamente desidentificados. Se a iniciativa nasce internamente das instituições, o acesso primário aos dados é facilitado e a tarefa de coleta de dados, também. Com isso, seria possível construir uma base de treinamento melhor e construir algoritmos de aprendizado

[1] Disponível em https://github.com/guilhermenoronha/. Acessado em março de 2022

de máquina mais eficientes.

É preciso conciliar a pesquisa acadêmica com a iniciativa privada para que os resultados possam ser potencializados. Nesse ponto, acredita-se que a contribuição deste trabalho é afirmar ser possível produzir pesquisa de ponta no Brasil utilizando-se dados de saúde como uso secundário da informação. Agora é uma questão de engajamento.

7.3 Ideias para trabalhos futuros

O presente trabalho levantou mais dúvidas do que respondeu questões. Os resultados encontrados não foram suficientes para que o modelo de aprendizagem seja colocado num ambiente de produção com dados reais. A questão da privacidade é crítica e exige que o modelo tenha um desempenho macro melhor do que o apresentado. Para isso levantam-se hipóteses traduzindo em ideias de trabalhos futuros. Elas são apresentadas nas subseções seguintes.

7.3.1 Usar um conjunto de dados maior

As limitações de coleta de dados já citadas impactaram na análise. Ainda sim, o algoritmo conseguiu gerar um F-Score macro no valor esperado e um micro razoável pela quantidade de dados disponível. A hipótese a ser testada é se a metodologia apresentada neste trabalho traria resultados mais significantes num conjunto maior de dados. A palavra maior pode ser considerada em dois aspectos distintos neste caso: quanto à sua quantidade e quanto à sua densidade.

O método poderia ser testado numa quantidade maior de dados. Para este projeto, foram utilizados 194 documentos clínicos com mais de 60 000 *tokens*. A conclusão tirada na revisão de literatura sugeriria pelo menos 3254 documentos. Quais resultados seriam possíveis em bases com milhares ou milhões de documentos?

Outro jeito de avaliar esta hipótese é quanto à densidade de dados sensíveis presentes no *corpus*. Nela, o conjunto de dados poderá ser menor, desde que a frequência relativa de dados sensíveis seja maior. O presente projeto encontrou 3,19% de dados sensíveis, valor superior que o mínimo esperado pela metodologia, que é de 2,7%. Mas quando os dados sensíveis são separados média de cada categoria, o percentual médio de dados sensíveis cairia para 0,22% por categoria. A quantidade esperada de dados sensíveis num *corpus* de 2,7% seria por categoria ou geral? Treinar o algoritmo numa base com densas quantidades de dados sensíveis traria uma *performance* melhor para o algoritmo de aprendizado?

7.3.2 Anotar sentenças em vez de documentos

O presente trabalho optou por agrupar todos os dados referentes a um documento antes de anotá-lo. Isso criou documentos extensos com diferentes categorias de informações para serem anotados. Essa técnica foi escolhida porque facilita o trabalho do anotador e delimita o jeito de preencher um documento clínico por autor. Embora os autores de cada documento não sejam conhecidos, notam-se diferenças no preenchimento dos dados, pois cada profissional tem a sua própria metodologia para escrever os documentos.

Optar por anotar sentenças em vez de documentos faz com que grandes trechos dos documentos sejam descartados, pois não possuiriam nenhum tipo de dado sensível. Isso produziria um *corpus* menor e mais denso, uma das hipóteses levantadas na subseção 7.3.1. O *corpus* produzido por essa diretriz de anotação geraria resultados melhores no aprendizado?

7.3.3 Criar modelos de aprendizagem para dados distintos

Outra hipótese que surgiu durante a execução deste trabalho é a ideia de elaborar um modelo de aprendizado para cada tipo de dado disponível no conjunto de dados. Os dados obtidos para este trabalho consistem de 14 diferentes campos, conforme mostra a Tabela 8. Uma abordagem seria elaborar um modelo de aprendizado para cada um desses campos.

Nesse caso, a metodologia teria de ser adaptada em subtarefas. Ela pode ser adotada para cada campo individualmente. Ou seja, seria construído um *corpus* para cada campo específico anotado e preparado para ser consumido pelo modelo de aprendizado de máquina.

A ideia dessa hipótese é que cada campo possui suas particularidades de escrita. Com um vocabulário mais restrito, seria mais fácil para o modelo generalizar as classificações. A construção do *corpus* geral de documentos clínicos nada mais seria do que a junção de cada subtarefa agrupada pelo número do documento.

Presume-se que seja necessária a coleta de mais dados para a execução dessa hipótese. Acredita-se que muitos dados serão descartados, pois alguns campos como "MEDICAMENTOS EM USO" possuem pouquíssimos dados sensíveis.

7.3.4 Enriquecer os dados

Na revisão de literatura apontou-se que o treinamento de sistemas de REM possui uma série de características, como as mostradas na Tabela 7. A técnica escolhida para este trabalho utiliza *tokens* de predição e *Bag of words* como principais características. Uma das hipóteses levantadas é se o enriquecimento dos dados com outras características

produziria um modelo melhor.

Uma contrapartida dessa hipótese é o alto custo de anotação de dados e também de uma possível adaptação do tratamento a ser feito para que os dados possam ser consumidos corretamente pelo algoritmo de aprendizagem.

7.3.5 Adotar outros algoritmos de aprendizado

Embora menos comuns, o Capítulo 4 apresentou outros métodos de aprendizado com resultados interessantes. Mesmo que a literatura sugira as redes neurais como principal escolha, pensa-se que os demais métodos podem apresentar alguma vantagem num conjunto de dados limitado como o do presente trabalho. A hipótese seria testá-los para ver se obtêm *performance* superior à encontrada no Capítulo 6.

7.3.6 Adotar técnicas mais sofisticadas para sintonizar os hiperparâmetros

A técnica utilizada no presente trabalho foi a busca em *grid*. É uma técnica simples que cria uma série de combinações com diferentes parâmetros possíveis. A execução da busca em *grid*, no entanto, exige que esses parâmetros sejam escolhidos pelo autor de forma arbitrária. A limitação de escolha desses parâmetros, bem como os valores escolhidos podem não ter sido ótimos do ponto de vista de *performance*.

Uma hipótese a ser testada é se técnicas mais sofisticadas para sintonizar os hiperparâmetros poderiam produzir um resultado significativamente melhor. Burkov (2019) lista algumas técnicas que podem ser implementadas como a busca aleatória, otimização bayesiana de hiperparâmetros, técnicas baseadas em gradiente e técnicas de otimização evolucionária. Os algoritmos genéticos também podem ser utilizados para sintonizar melhor hiperparâmetros dos algoritmos de aprendizado (GORGOLIS et al., 2019).

7.3.7 Discutir os benefícios da disponibilização de dados abertos com profissionais da saúde

Uma das hipóteses para a dificuldade de obtenção de dados é o desconhecimento dos benefícios que essa pesquisa traria por parte dos profissionais e instituições de saúde. Cabe refletir se as instituições e seus pesquisadores possuem conhecimento suficiente sobre os benefícios da disponibilização de dados. E, caso seja comprovado que há o conhecimento necessário, será que existe interesse das instituições na disponibilização desses dados? Uma pesquisa de campo poderia trazer uma luz à essa questão.

7.4 Comentários finais

A privacidade de dados no Brasil é um tema recente. A Lei Geral de Proteção de Dados foi aprovada em 2018 e vigora desde o final de 2020 (BRASIL, 2018). No momento em que esta tese está sendo finalizada, ainda surgem novos desdobramentos dessa lei. Em janeiro de 2022 foi aprovada uma lei que regulamenta a LGPD para agentes de tratamento de pequeno porte (BRASIL, 2022). Em fevereiro de 2022, o Congresso Nacional aprovou a proteção de dados como um direito constitucional (DAROS, 2022).

Ou seja, é um tema que ainda está sendo sedimentado no Brasil. Pensar na privacidade como um conceito, do inglês *privacy-by-design*, demandará tempo. No momento, as corporações estão preocupadas em proteger os dados para se adequar à lei e evitar transtornos maiores. Pensar na privacidade para adquirir e disseminar conhecimento é um passo além.

Assim como os conceitos de privacidade estão emergentes no Brasil, pode-se dizer o mesmo de outras áreas ligadas aos dados e à informação. Profissionais como cientistas, engenheiros e analistas de dados estão em alta no momento. Somente nos últimos anos é que a cultura de dados vem ganhando força no comércio e na indústria. As corporações estão entendendo que ter dados e saber usá-los é uma vantagem competitiva. Não fazer uso deles é ser subjugado fatidicamente pela concorrência feroz que é o mercado. A cultura de dados andará lado a lado com a privacidade. Percebe-se que ambas as áreas ainda têm muito a crescer no país.

Como discutido na seção 7.2, o presente trabalho apresenta à comunidade brasileira uma luz sobre questões envolvendo privacidade de dados e o PLN para o português do Brasil. Espera-se que novos trabalhos surjam e se complementem, seja na iniciativa de proteção à privacidade, na criação de novas técnicas de PLN ou na construção de novos repositórios para uso secundário da informação.

Referências

ABNT/CEE-78. *Informática em saúde — Sumário de alta de internação*. [S.l.], 2014. Citado 3 vezes nas páginas 48, 49 e 56.

ABNT/CEE-78). *Informática em saúde — Pseudonimização*. [S.l.], 2019. Citado na página 96.

AFZAL, Z. et al. Contextd: an algorithm to identify contextual properties of medical terms in a dutch clinical corpus. *BMC bioinformatics*, BioMed Central, v. 15, n. 1, p. 373, 2014. Citado na página 32.

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, v. 4, n. 3, p. 156–178, 2006. Citado na página 63.

ARAMAKI, E. et al. Automatic deidentification by using sentence features and label consistency. In: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*. [S.l.: s.n.], 2006. v. 2006, p. 10–11. Citado na página 73.

ARAMAKI, E. et al. Text2table: Medical text summarization system based on named entity recognition and modality identification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. [S.l.], 2009. p. 185–192. Citado na página 24.

BACIC, A. S. *Extração de informação e documentação de laudos médicos.* Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 34.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. [S.l.]: ACM Press Books, 2011. ISBN 978-0321416919. Citado na página 70.

BECKWITH, B. A. et al. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC medical informatics and decision making*, Springer, v. 6, n. 1, p. 12, 2006. Citado na página 79.

BERMAN, J. J. Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine*, Elsevier, v. 26, n. 1-2, p. 25–36, 2002. Citado na página 59.

BERMAN, J. J. Concept-match medical data scrubbing: how pathology text can be used in research. *Archives of pathology & laboratory medicine*, v. 127, n. 6, p. 680–686, 2003. Citado 2 vezes nas páginas 80 e 81.

BLAK, B. et al. Generalisability of the health improvement network (thin) database: demographics, chronic disease prevalence and mortality rates. *Journal of Innovation in Health Informatics*, v. 19, n. 4, p. 251–255, 2011. Citado na página 30.

BRASIL, R. F. d. *Constituição da República Federativa do Brasil de 1988*. 1988. Citado na página 23.

BRASIL, R. F. d. *LEI Nº 8.080, DE 19 DE SETEMBRO DE 1990.* 1990. Citado na página 46.

BRASIL, R. F. d. *TISS - Padrão para Troca de Informação de Saúde Suplementar.* 2016. Disponível em: <https://www.gov.br/ans/pt-br/assuntos/prestadores/padrao-para-troca-de-informacao-de-saude-suplementar-2013-tiss>. Citado na página 48.

BRASIL, R. F. d. *LEI Nº 13.709, DE 14 DE AGOSTO DE 2018.* 2018. Citado 3 vezes nas páginas 23, 37 e 128.

BRASIL, R. F. d. *RESOLUÇÃO CD/ANPD Nº 2, DE 27 DE JANEIRO DE 2022.* 2022. Citado na página 128.

BREZINA, V. *Statistics in corpus linguistics: A practical guide.* [S.l.]: Cambridge University Press, 2018. Citado na página 64.

BUI, D. D. A.; WYATT, M.; CIMINO, J. J. The uab informatics institute and 2016 cegs n-grid de-identification shared task challenge. *Journal of biomedical informatics*, Elsevier, v. 75, p. S54–S61, 2017. Citado na página 75.

BURKOV, A. *The hundred-page machine learning book.* [S.l.]: Andriy Burkov Quebec City, QC, Canada, 2019. Citado 2 vezes nas páginas 116 e 127.

CARRELL, D. et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 20, n. 2, p. 342–348, 2013. Citado na página 63.

CHAPMAN, C.; WANG, P.; STOLEE, K. T. Exploring regular expression comprehension. In: IEEE. *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE).* [S.l.], 2017. p. 405–416. Citado na página 78.

CHEN, J. H.; ASCH, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine*, NIH Public Access, v. 376, n. 26, p. 2507, 2017. Citado na página 22.

CHEN, T.; CULLEN, R. M.; GODWIN, M. Hidden markov model using dirichlet process for de-identification. *Journal of biomedical informatics*, Elsevier, v. 58, p. S60–S66, 2015. Citado na página 72.

CHEVRIER, R. et al. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 21, n. 5, p. e13484, 2019. Citado na página 23.

CHINCHOR, N. Muc-4 evaluation metrics. *Proceedings of the Fourth Message Understanding Conference*, p. 22–29, 1992. Citado na página 69.

CHINCHOR, N.; SUNDHEIM, B. M. Muc-5 evaluation metrics. In: *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993.* [S.l.: s.n.], 1993. Citado 2 vezes nas páginas 69 e 70.

Conselho Federal de Medicina. *RESOLUÇÃO CFM Nº 1.638/2002.* 2002. Citado na página 23.

Conselho Nacional da Saúde. *RESOLUÇÃO Nº 196, DE 10 DE OUTUBRO DE 1996*. 1996. Citado na página 59.

Conselho Nacional da Saúde. *RESOLUÇÃO Nº 466, DE 12 DE DEZEMBRO DE 2012*. 2012. Citado 2 vezes nas páginas 23 e 60.

Conselho Nacional da Saúde. *RESOLUÇÃO Nº 510, DE 7 DE ABRIL DE 2016*. 2016. Citado na página 60.

CRESWELL, J. W.; CRESWELL, J. D. *Research design: Qualitative, quantitative, and mixed methods approaches*. [S.l.]: Sage publications, 2017. Citado na página 86.

DALIANIS, H. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. [s.n.], 2018. ISBN 978-3-319-78502-8. Disponível em: <http://link.springer.com-/10.1007/978-3-319-78503-5>. Citado 7 vezes nas páginas 24, 26, 34, 58, 63, 68 e 70.

DALIANIS, H. et al. Stockholm epr corpus: A clinical database used to improve health care. In: *Swedish Language Technology Conference*. [S.l.: s.n.], 2012. p. 17–18. Citado na página 24.

DALIANIS, H.; HASSEL, M.; VELUPILLAI, S. The stockholm epr corpus-characteristics and some initial findings. *Proceedings of ISHIMR*, p. 243–249, 2009. Citado na página 31.

DALIANIS, H.; SKEPPSTEDT, M. Creating and evaluating a consensus for negated and speculative words in a swedish clinical corpus. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*. [S.l.], 2010. p. 5–13. Citado na página 22.

DAROS, G. *Proteção a dados pessoais vira direito constitucional; entenda o que muda*. 2022. `https://www.uol.com.br/tilt/noticias/redacao/2022/02/11/protecao-a-dados-pessoais-agora-e-direito-constitucional-o-que-muda.htm`. Citado na página 128.

DEAN, S. M. et al. Design and hospitalwide implementation of a standardized discharge summary in an electronic health record. *The Joint Commission Journal on Quality and Patient Safety*, Elsevier, v. 42, n. 12, p. 555–AP11, 2016. Citado 2 vezes nas páginas 23 e 46.

DEHGHAN, A. et al. Combining knowledge-and data-driven methods for de-identification of clinical narratives. *Journal of biomedical informatics*, Elsevier, v. 58, p. S53–S59, 2015. Citado 2 vezes nas páginas 74 e 83.

DEHGHAN, A. et al. Learning to identify protected health information by integrating knowledge-and data-driven algorithms: A case study on psychiatric evaluation notes. *Journal of biomedical informatics*, Elsevier, v. 75, p. S28–S33, 2017. Citado 3 vezes nas páginas 73, 74 e 83.

DELEGER, L. et al. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of biomedical informatics*, Elsevier, v. 50, p. 173–183, 2014. Citado 3 vezes nas páginas 23, 97 e 98.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, Elsevier, v. 42, n. 5, p. 760–772, 2009. Citado na página 43.

Department of Health and Human Services. *Standards for Privacy of Individually Identifiable Health Information*. [S.l.], 2000. 82461-82829 p. Disponível em: <https://www.federalregister.gov/documents/2000/12/28/00-32678/standards-for-privacy-of-individually-identifiable-health-information>. Citado na página 61.

DERNONCOURT, F.; LEE, J. Y.; SZOLOVITS, P. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*, 2017. Citado na página 77.

DERNONCOURT, F. et al. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, Oxford University Press, v. 24, n. 3, p. 596–606, 2017. Citado na página 77.

DORR, D. A. et al. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of information in medicine*, Schattauer GmbH, v. 45, n. 03, p. 246–252, 2006. Citado na página 63.

DOUGLASS, M. et al. Computer-assisted de-identification of free text in the mimic ii database. In: IEEE. *Computers in Cardiology, 2004*. [S.l.], 2004. p. 341–344. Citado na página 63.

EGUALE, T.; BARTLETT, G.; TAMBLYN, R. Rare visible disorders/diseases as individually identifiable health information. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA... Annual Symposium proceedings. AMIA Symposium*. [S.l.], 2005. v. 2005, p. 947–947. Citado na página 65.

EMAM, K. E. et al. Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, JMIR Publications Inc., Toronto, Canada, v. 8, n. 4, p. e28, 2006. Citado na página 65.

FIELSTEIN, E. M.; BROWN, S. H.; SPEROFF, T. Algorithmic de-identification of va medical exam text for hipaa privacy compliance: preliminary findings. *Medinfo*, v. 1590, 2004. Citado 2 vezes nas páginas 79 e 81.

FRIEDLIN, F. J.; MCDONALD, C. J. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 15, n. 5, p. 601–610, 2008. Citado 2 vezes nas páginas 79 e 81.

GALVÃO, M. C. B.; RICARTE, I. L. M. O prontuário eletrônico do paciente no século xxi: contribuições necessárias da ciência da informação. *InCID: Revista de Ciência da Informação e Documentação*, v. 2, n. 2, p. 77–100, 2011. Citado na página 60.

GAO, Y. et al. Constructing a chinese electronic medical record corpus for named entity recognition on resident admit notes. *BMC medical informatics and decision making*, BioMed Central, v. 19, n. 2, p. 56, 2019. Citado na página 33.

GARDNER, J.; XIONG, L. Hide: an integrated system for health information de-identification. In: IEEE. *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. [S.l.], 2008. p. 254–259. Citado 2 vezes nas páginas 73 e 74.

GARFINKEL, S. L. De-identification of personal information. *National institute of standards and technology*, 2015. Citado na página 58.

GAUDET-BLAVIGNAC, C. et al. De-identification of french medical narratives. *Swiss Medical Informatics*, EMH Media, v. 34, n. 00, 2018. Citado 2 vezes nas páginas 80 e 81.

GIL, A. C. *Fundamentos de Metodologia Científica*. 7. ed. [S.l.]: Atlas, 2019. Citado na página 86.

GKOULALAS-DIVANIS, A.; LOUKIDES, G.; SUN, J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, Elsevier, v. 50, p. 4–19, 2014. Citado na página 65.

GOLDBERG, Y. *Neural Network Methods in Natural Language Processing*. [S.l.: s.n.], 2017. 1–309 p. ISSN 10636919. ISBN 9781627052986. Citado na página 76.

GORGOLIS, N. et al. Hyperparameter optimization of lstm network models through genetic algorithm. In: IEEE. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. [S.l.], 2019. p. 1–4. Citado na página 127.

GROUIN, C.; NÉVÉOL, A. De-identification of clinical notes in french: towards a protocol for reference corpus development. *Journal of biomedical informatics*, Elsevier, v. 50, p. 151–161, 2014. Citado na página 63.

GUILLEN, R. et al. Automated de-identification and categorization of medical records. In: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*. [S.l.: s.n.], 2006. v. 116. Citado na página 79.

GUO, Y. et al. Identifying personal health information using support vector machines. In: CITESEER. *i2b2 workshop on challenges in natural language processing for clinical data*. [S.l.], 2006. p. 10–11. Citado na página 75.

GUPTA, D.; SAUL, M.; GILBERTSON, J. Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American journal of clinical pathology*, Oxford University Press Oxford, UK, v. 121, n. 2, p. 176–186, 2004. Citado 2 vezes nas páginas 80 e 81.

HANAUER, D. et al. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International journal of medical informatics*, Elsevier, v. 82, n. 9, p. 821–831, 2013. Citado 2 vezes nas páginas 63 e 64.

HARA, K. et al. Applying a svm based chunker and a text classifier to the deid challenge. In: *i2b2 Workshop on challenges in natural language processing for clinical data*. [S.l.: s.n.], 2006. p. 10–11. Citado na página 75.

HAVERINEN, K. et al. Parsing clinical finnish: Experiments with rule-based and statistical dependency parsers. In: *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*. [S.l.: s.n.], 2009. p. 65–72. Citado na página 33.

HAVERINEN, K. et al. Dependency-based propbanking of clinical finnish. In: *Proceedings of the Fourth Linguistic Annotation Workshop*. [S.l.: s.n.], 2010. p. 137–141. Citado na página 33.

HE, B. et al. Crfs based de-identification of medical records. *Journal of biomedical informatics*, Elsevier, v. 58, p. S39–S46, 2015. Citado 2 vezes nas páginas 73 e 74.

HENRIKSSON, A.; KVIST, M.; DALIANIS, H. Detecting protected health information in heterogeneous clinical notes. *Studies in health technology and informatics*, v. 245, p. 393, 2017. Citado na página 64.

HERRETT, E. et al. Data resource profile: clinical practice research datalink (cprd). *International journal of epidemiology*, Oxford University Press, v. 44, n. 3, p. 827–836, 2015. Citado na página 30.

HUANG, Z.; XU, W.; YU, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. Citado na página 98.

III, S. J. D.; WEAVER, A. C.; HUGHES, K. K. Health insurance portability and accountability act. *Security Issues in the Digital Medical Enterprise*, v. 72, n. 2, p. 9–18, 2004. Citado na página 61.

ISO. *Health informatics — Pseudonymization*. Geneva, Switzerland, 2017. Citado na página 58.

IV, W. F. S. et al. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 2, p. 143–154, 2014. Citado na página 29.

JANOWICZ, K.; KESSLER, C. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, Taylor & Francis, v. 22, n. 10, p. 1129–1157, 2008. Citado na página 68.

JENSEN, K. et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, Nature Publishing Group, v. 7, p. 46226, 2017. Citado na página 22.

JIANG, Z. et al. De-identification of medical records using conditional random fields and long short-term memory networks. *Journal of biomedical informatics*, Elsevier, v. 75, p. S43–S53, 2017. Citado na página 77.

JOHNSON, A. et al. Mimic-iv-ed. 2021. Citado na página 28.

JOHNSON, K. W. et al. Evaluation of patient re-identification using laboratory test orders and mitigation via latent space variables. In: WORLD SCIENTIFIC. *PSB*. [S.l.], 2019. p. 415–426. Citado na página 64.

Joint Commission International. *Joint Commission International Accreditation Standards for Hospitals: Including Standards for Academic Medical Center Hospitals*. [S.l.]: Joint Commission Resources, 2020. Citado 2 vezes nas páginas 48 e 49.

JURAFSKY, D.; MARTIN, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd. ed. [S.l.]: Prentice Hall, 2008. 1038 p. ISSN 1098-6596. ISBN 9780131873216. Citado 5 vezes nas páginas 66, 67, 68, 71 e 72.

KARA, E. et al. A domain-adapted dependency parser for german clinical text. In: *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*. [S.l.: s.n.], 2018. Citado na página 33.

KEMENY, J. G.; SNELL, J. L. *Markov chains*. [S.l.]: Springer-Verlag, New York, 1976. Citado na página 71.

KIND, A. J.; SMITH, M. A. Documentation of mandated discharge summary components in transitions from acute to subacute care. *Advances in patient safety: new directions and alternative approaches (Vol. 2: culture and redesign)*, Agency for Healthcare Research and Quality (US), 2008. Citado na página 47.

KOKKINAKIS, D. Collection, encoding and linguistic processing of a swedish medical corpus–the medlex experience. *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, p. 1200–1205, 2006. Citado na página 22.

KOKKINAKIS, D.; THURIN, A. Anonymisation of swedish clinical data. In: SPRINGER. *Conference on Artificial Intelligence in Medicine in Europe*. [S.l.], 2007. p. 237–241. Citado na página 63.

KVIST, M.; VELUPILLAI, S. Scan: A swedish clinical abbreviation normalizer. In: SPRINGER. *International Conference of the Cross-Language Evaluation Forum for European Languages*. [S.l.], 2014. p. 62–73. Citado na página 22.

LEE, H.-J. et al. A hybrid approach to automatic de-identification of psychiatric notes. *Journal of biomedical informatics*, Elsevier, v. 75, p. S19–S27, 2017. Citado na página 73.

LEE, J. Y. et al. Feature-augmented neural networks for patient note de-identification. *arXiv preprint arXiv:1610.09704*, 2016. Citado na página 77.

LEECH, G. *Developing linguistic corpora: a guide to good practice adding linguistic annotation*. [S.l.]: Lancaster University, 2004. ISSN 1463 5194. Citado na página 41.

LI, X.-B.; QIN, J. Anonymizing and sharing medical text records. *Information Systems Research*, INFORMS, v. 28, n. 2, p. 332–352, 2017. Citado 2 vezes nas páginas 64 e 65.

LIU, Z. et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics*, Elsevier, v. 58, p. S47–S52, 2015. Citado na página 74.

LIU, Z. et al. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, Elsevier, v. 75, p. S34–S42, 2017. Citado na página 77.

LOPES, M. V. d. O.; ARAÚJO, T. L. d. Processo de informatização em saúde: temas abordados em artigos publicados no período de 1978 a 1998. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 36, p. 25–32, 2002. Citado na página 21.

LOURENÇÃO, L. G.; JUNIOR, C. d. J. F. Implantação do prontuário eletrônico do paciente no brasil. *Enfermagem Brasil*, v. 15, n. 1, p. 44–53, 2016. Citado na página 21.

MACLEOD, H. et al. Identifying rare diseases from behavioural data: a machine learning approach. In: IEEE. *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on*. [S.l.], 2016. p. 130–139. Citado 2 vezes nas páginas 22 e 43.

MARCONI, M. d. A.; LAKATOS, E. M. *Fundamentos de Metodologia Científica*. 8. ed. [S.l.]: Atlas São Paulo, 2017. Citado na página 86.

MARIMON, M. et al. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS. org), Bilbao, Spain (Sep 2019), TBA*. [S.l.: s.n.], 2019. Citado na página 34.

MAYO, M.; YOGARAJAN, V. A nearest neighbour-based analysis to identify patients from continuous glucose monitor data. In: SPRINGER. *Asian Conference on Intelligent Information and Database Systems*. [S.l.], 2019. p. 349–360. Citado na página 65.

MEYSTRE, S. et al. Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics*, Georg Thieme Verlag KG, v. 26, n. 01, p. 38–52, 2017. Citado na página 60.

MEYSTRE, S. M. et al. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, BioMed Central, v. 10, n. 1, p. 70, 2010. Citado 6 vezes nas páginas 62, 68, 69, 70, 81 e 95.

MEYSTRE, S. M. et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, Georg Thieme Verlag KG, v. 17, n. 01, p. 128–144, 2008. Citado na página 22.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 69.

Ministério da Saúde. *Manual técnico do Sistema de Informação Hospitalar*. [S.l.]: Ministério da Saúde Brasília, 2007. Citado na página 48.

Ministério da Saúde. Portaria nº 2.073, de 31 de agosto de 2011. regulamenta o uso de padrões de interoperabilidade e informação em saúde para sistemas de informação em saúde no âmbito do sistema único de saúde, nos níveis municipal, distrital, estadual e federal, e para os sistemas privados e do setor de saúde suplementar. *Diário Oficial da União*, 2011. Citado 2 vezes nas páginas 22 e 56.

Ministério da Saúde. *Modelo de Informação Sumário de Alta*. 2017. Citado 9 vezes nas páginas 13, 46, 47, 50, 51, 52, 53, 54 e 55.

MITCHELL, T. M. Machine learning and data mining. *Communications of the ACM*, ACM New York, NY, USA, v. 42, n. 11, p. 30–36, 1999. Citado na página 70.

MORRISON, F. P. et al. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 16, n. 1, p. 37–39, 2009. Citado 2 vezes nas páginas 80 e 81.

MÜLLER, A. C.; GUIDO, S. et al. *Introduction to machine learning with Python: a guide for data scientists*. [S.l.]: "O'Reilly Media, Inc.", 2016. Citado 2 vezes nas páginas 75 e 76.

MURPHY, S. N. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, BMJ Group, v. 17, n. 2, p. 124–130, 2010. Citado na página 26.

NASRABADI, N. M. Pattern recognition and machine learning. *Journal of electronic imaging*, International Society for Optics and Photonics, v. 16, n. 4, p. 049901, 2007. Citado na página 69.

National e-Health Transition Authority. *Implementing electronic discharge summaries: the JMO perspective*. 2010. Citado na página 48.

National Science Board. *Science and Engineering Indicators 2018*. [S.l.]: National Science Foundation, 2018. Citado 2 vezes nas páginas 37 e 42.

NEAMATULLAH, I. et al. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, Springer, v. 8, n. 1, p. 32, 2008. Citado 2 vezes nas páginas 63 e 80.

NÉVÉOL, A. et al. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, v. 9, n. 1, p. 1–13, 2018. ISSN 20411480. Citado 2 vezes nas páginas 24 e 42.

NEVES, M. L.; JIMENO-YEPES, A.; NÉVÉOL, A. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In: *LREC*. [S.l.: s.n.], 2016. Citado na página 43.

NIZAMUDDIN, U.; DALIANIS, H. Detection of spelling errors in swedish clinical text. In: *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014), SLTC 2014, Uppsala*. [S.l.: s.n.], 2014. Citado na página 22.

OLIVEIRA, L. E. S. e et al. Learning portuguese clinical word embeddings: a multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task. 2019. Citado na página 35.

OLIVEIRA, L. E. S. e et al. A rule-based method for continuity of care identification in discharge summaries. In: *MedInfo*. [S.l.: s.n.], 2013. p. 1221. Citado na página 35.

PACHECO, E. J. Morphomap: mapeamento automático de narrativas clínicas para uma terminologia médica. Universidade Tecnológica Federal do Paraná, 2009. Citado na página 34.

PAKHOMOV, S.; PEDERSEN, T.; CHUTE, C. G. Abbreviation and acronym disambiguation in clinical discourse. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2005. v. 2005, p. 589. Citado na página 22.

PANTAZOS, K.; LAUESEN, S.; LIPPERT, S. Preserving medical correctness, readability and consistency in de-identified health records. *Health informatics journal*, SAGE Publications Sage UK: London, England, v. 23, n. 4, p. 291–303, 2017. Citado 2 vezes nas páginas 63 e 65.

PEREZ, A. et al. Semi-supervised medical entity recognition: A study on spanish and swedish clinical corpora. *Journal of biomedical informatics*, Elsevier, v. 71, p. 16–30, 2017. Citado na página 24.

PESTIAN, J. P. et al. A shared task involving multi-label classification of clinical free text. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. [S.l.], 2007. p. 97–104. Citado 2 vezes nas páginas 28 e 29.

PESTIAN, J. P. et al. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, SAGE Publications Sage UK: London, England, v. 5, p. BII–S9042, 2012. Citado na página 28.

PHILLIPS, D. C.; PHILLIPS, D. C.; BURBULES, N. C. *Postpositivism and educational research*. [S.l.]: Rowman & Littlefield, 2000. Citado na página 86.

PHUONG, N. D.; CHAU, V. T. N. Automatic de-identification of medical records with a multilevel hybrid semi-supervised learning approach. In: IEEE. *2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*. [S.l.], 2016. p. 43–48. Citado na página 74.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020. Citado na página 115.

PUSTEJOVSKY, J.; STUBBS, A. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. [S.l.]: "O'Reilly Media, Inc.", 2013. Citado 3 vezes nas páginas 94, 95 e 97.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Ieee, v. 77, n. 2, p. 257–286, 1989. Citado na página 72.

REIS, Z. S. N. et al. Análise do conteúdo do sumário de alta obstétrica em maternidade de referência uma oportunidade para repensar a estratégia da continuidade do cuidado materno e neonatal. *Rev méd Minas Gerais*, v. 25, n. 4, 2015. Citado na página 46.

RICHTER-PECHANSKI, P. et al. Deep learning approaches outperform conventional strategies in de-identification of german medical reports. *Studies in health technology and informatics*, v. 267, p. 101–109, 2019. Citado na página 77.

ROTHSTEIN, M. A. Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics*, Taylor & Francis, v. 10, n. 9, p. 3–11, 2010. Citado na página 64.

RUCH, P. et al. Medical document anonymization with a semantic lexicon. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Symposium*. [S.l.], 2000. p. 729. Citado 2 vezes nas páginas 80 e 81.

SALUJA, B. et al. Anonymization of sensitive information in medical health records. 2019. Citado na página 77.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, n. 3, p. 210–229, 1959. Citado na página 70.

SANTOS, T. O. dos; PEREIRA, L. P.; SILVEIRA, D. T. Implantação de sistemas informatizados na saúde: uma revisão sistemática. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, v. 11, n. 3, 2017. Citado na página 21.

SASAKI, Y. et al. The truth of the f-measure. *Teach Tutor mater*, v. 1, n. 5, p. 1–5, 2007. Citado 2 vezes nas páginas 69 e 70.

SCHILLER, A. et al. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. [S.l.], 1999. Citado na página 33.

SEIFFE, L. *Linguistic Modeling for Text Analytic Tasks for German Clinical Texts*. Dissertação (Mestrado) — Technische Universität Berlin, 2018. Citado na página 33.

SHIN, H.; MARKEY, M. K. A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *Journal of Biomedical Informatics*, Elsevier, v. 39, n. 2, p. 227–248, 2006. Citado na página 22.

SPAT, S. et al. Enhanced information retrieval from narrative german-language clinical text documents using automated document classification. *Studies in health technology and informatics*, IOS Press; 1999, v. 136, p. 473, 2008. Citado na página 24.

SRIVASTAVA, A. et al. A recurrent neural network architecture for de-identifying clinical records. In: *Proceedings of the 13th international conference on natural language processing*. [S.l.: s.n.], 2016. p. 188–197. Citado na página 78.

STUBBS, A.; FILANNINO, M.; UZUNER, Ö. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, Elsevier, v. 75, p. S4–S18, 2017. Citado na página 70.

STUBBS, A.; KOTFILA, C.; UZUNER, Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, Elsevier, v. 58, p. S11–S19, 2015. Citado na página 70.

STUBBS, A.; UZUNER, Ö. De-identification of medical records through annotation. In: *Handbook of Linguistic Annotation*. [S.l.]: Springer, 2017. p. 1433–1459. Citado 3 vezes nas páginas 95, 96 e 98.

STUBBS, A. et al. Challenges in synthesizing surrogate phi in narrative emrs. In: *Medical Data Privacy Handbook*. [S.l.]: Springer, 2015. p. 717–735. Citado na página 59.

SWEENEY, L. Replacing personally-identifying information in medical records, the scrub system. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA annual fall symposium*. [S.l.], 1996. p. 333. Citado 2 vezes nas páginas 79 e 81.

SWEENEY, L. Simple demographics often identify people uniquely. *Health (San Francisco)*, v. 671, p. 1–34, 2000. Citado na página 64.

SZARVAS, G.; FARKAS, R.; BUSA-FEKETE, R. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 14, n. 5, p. 574–580, 2007. Citado na página 76.

SÁ, T. Q. V. D. *PROPOSTA PARA O COMPARTILHAMENTO DE INFORMAÇÕES SOBRE O CUIDADO OBSTÉTRICO ENTRE A REDE DE ATENÇÃO BÁSICA E A MATERNIDADE*. Dissertação (Mestrado) — UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2018. Citado na página 92.

TAIRA, R. K.; BUI, A. A.; KANGARLOO, H. Identification of patient name references within medical documents using semantic selectional restrictions. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Symposium*. [S.l.], 2002. p. 757. Citado na página 73.

The Health and Social Care Information Centre. *Standards for the clinical structure and content of patient records*. 2013. Citado na página 48.

THOMAS, S. M. et al. A successful technique for removing names in pathology reports using an augmented search and replace method. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Symposium*. [S.l.], 2002. p. 777. Citado 2 vezes nas páginas 79 e 81.

TRIENES, J. et al. Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. *arXiv preprint arXiv:2001.05714*, 2020. Citado na página 77.

UZUNER, Ö. et al. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 15, n. 1, p. 14–24, 2008. Citado na página 27.

UZUNER, Ö.; LUO, Y.; SZOLOVITS, P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 14, n. 5, p. 550–563, 2007. Citado 6 vezes nas páginas 27, 61, 68, 70, 81 e 82.

UZUNER, Ö. et al. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, Elsevier, v. 42, n. 1, p. 13–35, 2008. Citado na página 63.

VELUPILLAI, S. Automatic classification of factuality levels: A case study on swedish diagnoses and the impact of local context. In: *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*. [S.l.: s.n.], 2011. Citado na página 22.

VELUPILLAI, S. et al. Developing a standard for de-identifying electronic patient records written in swedish: precision, recall and f-measure in a manual and computerized annotation trial. *International journal of medical informatics*, Elsevier, v. 78, n. 12, p. e19–e26, 2009. Citado na página 63.

VELUPILLAI, S.; KVIST, M. Fine-grained certainty level annotations used for coarser-grained e-health scenarios. In: SPRINGER. *International Conference on Intelligent Text Processing and Computational Linguistics*. [S.l.], 2012. p. 450–461. Citado na página 22.

VERYARD, R. *Information modelling: practical guidance*. [S.l.]: Prentice-Hall, Inc., 1992. Citado na página 49.

VIANA, V.; TAGNIN, S. E. *Corpora no ensino de línguas estrangeiras*. [S.l.]: Hub Editorial, 2011. Citado 2 vezes nas páginas 39 e 40.

VINCZE, V. et al. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, BioMed Central, v. 9, n. 11, p. S9, 2008. Citado 3 vezes nas páginas 22, 24 e 29.

WELLNER, B. et al. Rapidly retargetable approaches to de-identification in medical records. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 14, n. 5, p. 564–573, 2007. Citado 2 vezes nas páginas 72 e 73.

World Health Organization. *Electronic health records: manual for developing countries*. [S.l.]: Manila: WHO Regional Office for the Western Pacific, 2006. Citado na página 58.

YADAV, S. et al. Patient data de-identification: a conditional random-field-based supervised approach. In: *Handbook of Research on Applied Cybernetics and Systems Science*. [S.l.]: IGI Global, 2017. p. 234–253. Citado 4 vezes nas páginas 73, 74, 78 e 83.

YANG, H.; GARIBALDI, J. M. Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, Elsevier, v. 58, p. S30–S38, 2015. Citado na página 74.

YANG, H. et al. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 16, n. 4, p. 596–600, 2009. Citado na página 43.

YOGARAJAN, V.; MAYO, M.; PFAHRINGER, B. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765*, 2018. Citado 2 vezes nas páginas 68 e 70.

YOGARAJAN, V.; PFAHRINGER, B.; MAYO, M. Automatic end-to-end de-identification: Is high accuracy the only metric? *arXiv preprint arXiv:1901.10583*, 2019. Citado 3 vezes nas páginas 64, 81 e 82.

ZENG, Q. T. et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, BioMed Central, v. 6, n. 1, p. 30, 2006. Citado na página 22.

ZHAO, Y.-S. et al. Leveraging text skeleton for de-identification of electronic medical records. *BMC medical informatics and decision making*, Springer, v. 18, n. 1, p. 18, 2018. Citado 2 vezes nas páginas 77 e 78.

Anexos

ANEXO A – Parecer consubstanciado do Hospital das Clínicas

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: Tratamento da informação médica para atendimento à necessidade de privacidade: desidentificação textual de prontuários eletrônicos na língua portuguesa do Brasil

Pesquisador: Mauricio Barcellos Almeida
Área Temática:
Versão: 4
CAAE: 02429618.1.0000.5149
Instituição Proponente: Universidade Federal de Minas Gerais
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.384.141

Apresentação do Projeto:
O projeto Tratamento da informação médica para atendimento à necessidade de privacidade: desidentificação textual de prontuários eletrônicos na língua portuguesa do Brasil – CAAE 02429618.1.0000.5149, obteve sua aprovação junto ao CEP na data de 24 de Março de 2019, sob parecer de n. 3.230.537. O objetivo primário da pesquisa era construir e disponibilizar gratuitamente um corpus, na língua portuguesa e desidentificado, de prontuários de pacientes. O protocolo aprovado propõe a metodologia da pesquisa em duas etapas práticas: criação de corpus e desenvolvimento de algoritmo de desidentificação de prontuários de pacientes. Para isso é necessário a obtenção de prontuários eletrônicos de pacientes devidamente solicitados ao comitê de ética e pesquisa da UFMG. Ele será disponibilizado publicamente com uma parceria junto ao programa de pós-graduação em estudos linguísticos da UFMG e o programa de pós-graduação em gestão e organização do conhecimento após a sua desidentificação que usará aprendizado de máquina como recurso.
O projeto encontra-se em andamento e a etapa atual é a coleta de dados do Hospital das Clínicas. No entanto, há uma dificuldade na obtenção desses dados que justifica a solicitação dessa emenda. Esta emenda tem como objetivo coletar mais dados para a realizar a pesquisa em desidentificação. Foram solicitados dados de ginecologia do Hospital Felício Rocho referentes ao ano de 2018.

Endereço: Av. Presidente Antônio Carlos,6627 2º Ad Sl 2005
Bairro: Unidade Administrativa II **CEP:** 31.270-901
UF: MG **Município:** BELO HORIZONTE
Telefone: (31)3409-4592 **E-mail:** coep@prpq.ufmg.br

Os riscos para esta etapa permanecem idênticos aos da etapa anterior e, portanto, não é necessário tomar novas medidas para atenuá-los. Os riscos da etapa anterior estão associados com o vazamento parcial ou total das informações pessoais dos pacientes. O vazamento dos dados poderia colocar em risco a integridade física, moral, pessoal, religiosa, cultural e ética dos pacientes.
Informa-se que a instituição coparticipante, que abriga os pacientes que se pretende incluir consentiu com a pesquisa, formalizando o aceite por meio da assinatura do termo de Anuência, documento este, que foi apresentado como anexo a esta emenda. Da mesma forma esclarece-se que os demais documentos solicitados para essa emenda foram anexados a esta plataforma ao qual solicita-se apreciação.

Objetivo da Pesquisa:
Objetivo Primário:
O objetivo geral é construir, para fins de pesquisa, um corpus de prontuários de pacientes na língua portuguesa e desidentificado.
Objetivo Secundário:
Os objetivos específicos são:
Levantar técnicas de desidentificação de textos médicos para a língua portuguesa;
Criar programas de computador que automatizem o processo de desidentificação usando aprendizado de máquina;
Criar um corpus de prontuários com pacientes pseudônimos, para que seja protegida a identidade das pessoas, o qual será usado como treinamento de algoritmos de desidentificação;
Criar outro corpus de prontuários de pacientes, com dados já desidentificados, o qual possibilite a continuidade de pesquisas em medicina e informática médica;
Estudar a legislação e as implicações legais que envolvem o trabalho, considerando que o corpus ainda pertence ao HC e sujeito a sua aprovação para disseminação.

Avaliação dos Riscos e Benefícios:
• Riscos:
Como não há mudança na metodologia e nos critérios de inclusão e alteração, os riscos estão inalterados em relação à versão atual do projeto aprovado.
• Benefícios:
Como não há mudança na metodologia e nos critérios de inclusão e alteração, os benefícios estão

inalterados em relação à versão atual do projeto aprovado.

Comentários e Considerações sobre a Pesquisa:

Inalterados em relação ao projeto mais recente aprovado.

Considerações sobre os Termos de apresentação obrigatória:

. Foram apresentados: carta da emenda, carta de anuência da nova instituição assinada, novo TCUD assinado, declaração atualizada de comprometimento do pesquisador responsável, declaração atualizada de dispensa do TCLE para os novos dados, sendo todos os documentos conformes às versões anteriores já aprovadas quando do projeto original.

. Os demais documentos permanecem válidos.

. A nova instituição coparticipante foi devidamente adicionada no formulário da Plataforma Brasil.

Recomendações:

. Quando da apresentação de relatórios (parciais e final), será necessário apresentar a folha de rosto atual devidamente datada, assinada e carimbada pela diretoria da unidade da UFMG responsável pela pesquisa.

Conclusões ou Pendências e Lista de Inadequações:

Confiante de que a recomendação será seguida, sou, SMJ, favorável à aprovação da emenda.

Considerações Finais a critério do CEP:

Tendo em vista a legislação vigente (Resolução CNS 466/12), o CEP-UFMG recomenda aos Pesquisadores: comunicar toda e qualquer alteração do projeto e do termo de consentimento via emenda na Plataforma Brasil, informar imediatamente qualquer evento adverso ocorrido durante o desenvolvimento da pesquisa (via documental encaminhada em papel), apresentar na forma de notificação relatórios parciais do andamento do mesmo a cada 06 (seis) meses e ao término da pesquisa encaminhar a este Comitê um sumário dos resultados do projeto (relatório final).

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_1602431_E1.pdf	26/08/2020 15:17:30		Aceito
Outros	Carta_de_Emenda.pdf	26/08/2020 15:16:59	GUILHERME FRANCIS DE	Aceito

Outros	Carta_de_Emenda.pdf	26/08/2020 15:16:59	NORONHA	Aceito
Folha de Rosto	folhaDeRosto.pdf	25/08/2020 14:24:04	GUILHERME FRANCIS DE NORONHA	Aceito
Projeto Detalhado / Brochura Investigador	Pr__Projeto_COEPE.pdf	25/08/2020 14:19:29	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Guilherme_Francis_de_Noronha.pdf	25/08/2020 14:15:06	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Zilma_Silveira_Nogueira_Reis.pdf	25/08/2020 14:13:31	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Heliana_Ribeiro_de_Mello.pdf	25/08/2020 14:13:00	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_do_Sistema_de_Currculos_Lattes_Maurcio_Barcellos_Almeida.pdf	24/08/2020 16:42:41	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Curriculo_Lattes_Amanda_Damasceno_de_Souza.pdf	13/08/2020 16:04:12	Amanda Damasceno de Souza	Aceito
Declaração de Instituição e Infraestrutura	Declaracao_Anuencia_Infraestrutura_instituicao_Hospital_Felicio_Rocho.pdf	13/08/2020 16:01:11	Amanda Damasceno de Souza	Aceito
Outros	TERMO_DE_COMPROMISSO_PARA_UTILIZAO_DE_DADOS_DE_ARQUIVO.pdf	28/07/2020 14:54:04	GUILHERME FRANCIS DE NORONHA	Aceito
Declaração de Pesquisadores	DECLARAO_DE_COMPROMETIMENTO_DO_PESQUISADOR.pdf	28/07/2020 14:51:54	GUILHERME FRANCIS DE NORONHA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	DECLARAO_DE_DISPENSA_DE_TERMO_DE_CONSENTIMENTO_LIVRE_E_ESCLARECIDO.pdf	28/07/2020 14:40:20	GUILHERME FRANCIS DE NORONHA	Aceito
Declaração de Instituição e Infraestrutura	modelo_HCUFMGEBSERH_aprovado_GEP.pdf	24/03/2019 18:33:13	Mauricio Barcellos Almeida	Aceito
Declaração de Pesquisadores	termos_pesquisador.pdf	24/03/2019 18:32:57	Mauricio Barcellos Almeida	Aceito
Declaração de Instituição e Infraestrutura	parecer_unidade_funcional.pdf	24/03/2019 18:32:28	Mauricio Barcellos Almeida	Aceito
Outros	resposta_240319.pdf	24/03/2019 18:21:55	Mauricio Barcellos Almeida	Aceito
Outros	tecnologiadeinformacao.pdf	19/02/2019 10:25:48	EDILAINE APARECIDA DE	Aceito

Outros	tecnologiadeinformacao.pdf	19/02/2019 10:25:48	SOUZA	Aceito
Declaração de Instituição e Infraestrutura	Parecer_GOB16_2018.jpg	08/01/2019 19:31:05	Mauricio Barcellos Almeida	Aceito
Recurso Anexado pelo Pesquisador	resposta_080119.pdf	08/01/2019 18:56:39	Mauricio Barcellos Almeida	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	just.pdf	06/11/2018 13:38:01	Mauricio Barcellos Almeida	Aceito
Parecer Anterior	2.pdf	05/11/2018 17:35:55	Mauricio Barcellos Almeida	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

BELO HORIZONTE, 07 de Novembro de 2020

Assinado por:
Críssia Carem Paiva Fontainha
(Coordenador(a))

ANEXO B – Parecer consubstanciado do Felício Rocho

HOSPITAL FELÍCIO ROCHO/MG

PARECER CONSUBSTANCIADO DO CEP

Elaborado pela Instituição Coparticipante

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Tratamento da informação médica para atendimento à necessidade de privacidade: desidentificação textual de prontuários eletrônicos na língua portuguesa do Brasil

Pesquisador: Mauricio Barcellos Almeida
Área Temática:
Versão: 2
CAAE: 02429618.1.3002.5125
Instituição Proponente: Hospital Felício Rocho/MG
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.529.223

Apresentação do Projeto:
A pesquisa médica que utiliza corpora textuais como objeto de pesquisa enfrenta limitações de reprodutibilidade devido à dificuldade de acesso aos dados de prontuários de pacientes. Esses dados devem ser desidentificados para assegurar aos cidadãos, de acordo com as leis brasileiras, o direito à privacidade. Muitos dados não podem ser usados em pesquisas médicas ou nas áreas de informática em função das
limitações de acesso à informação. Ainda há a barreira linguística que impede que pesquisas feitas em outros idiomas, possam ser aproveitadas para a língua portuguesa do Brasil.Problema: o formato e a forma de uso dos prontuários de pacientes na língua portuguesa, assim como em outros idiomas, dificultam o uso de técnicas de mineração de dados. Essas técnicas são essenciais para fazer frente ao volume de dados produzido hoje na área médica, o qual não é mais tratável sem o uso de recursos computacionais. Desambiguação de termos, identificação de abreviações, acrônimos e limitações de interoperabilidade semântica, entre diferentes instituições e campos médicos, são alguns dos problemas enfrentados. Apenas a criação de um corpus textual de prontuários de pacientes pode impulsionar a produção de pesquisa científica ao auxiliar a resolver problemas de manipulação de grandes volumes de textos médicos produzidos diariamente. A disponibilização desse corpus depende da desidentificação dos textos, de forma que, os dados sigilosos dos pacientes estejam legal e moralmente protegidos. A revisão de literatura não

Endereço: Rua Uberaba, n° 500, 5° andar, Núcleo de Ciências da Saúde Felício Rocho
Bairro: Barro Preto **CEP:** 30.180-082
UF: MG **Município:** BELO HORIZONTE
Telefone: (31)3514-7626 **Fax:** (31)3514-7626 **E-mail:** cep@feliciorocho.org.br

identificou ferramentas e iniciativas de
desidentificação de textos médicos para a língua portuguesa do Brasil.Metodologia: serão estudadas técnicas de aprendizado de máquina para desidentificação automática de prontuários de pacientes. Dois corpora serão criados utilizando dados em parceria com o Hospital Felício Rocho da UFMG. O primeiro corpus será usado no treinamento de algoritmos computacionais e o segundo conterá os dados resultantes do processo de desidentificação. Os corpora e as ferramentas criadas ao longo do projeto
serão de guarda do Hospital Felício Rocho, e por este, disponibilizados de acordo com sua conveniência e legislação pertinente, para fins de pesquisa e consulta de profissionais da saúde, Ciência da Computação, Ciência da Informação, dentre outros possíveis interessados. Ao final do projeto, o pesquisador solicitará formalmente ao Hospital Felício Rocho que os dados supracitados, anonimizados e sob guarda daquele hospital sejam tornados públicos, respeitadas as limitações da legislação, uma vez que um dos objetivos do projeto é fomentar a pesquisa médica. Isso possibilitaria que toda a comunidade científica brasileira tivesse acesso rápido a um conjunto de dados ainda inexistente no país para pesquisa, o qual já não carrega mais qualquer possibilidade de rastreabilidade. Resultados esperados: ao final da realização do trabalho é esperado que o corpus criado contribua em diferentes áreas médicas facilitando a obtenção de dados, fomentando a proteção aos pacientes e impulsionando as pesquisas no país. Em última instância pretende-se promover o acesso à informação médica no Brasil em benefício da sociedade.

Objetivo da Pesquisa:

Objetivo Primário:

O objetivo geral é construir, para fins de pesquisa, um corpus de prontuários de pacientes na língua portuguesa e desidentificado.

Objetivos Secundários:

Os objetivos específicos são:

-levantar técnicas de desidentificação de textos médicos para a língua portuguesa;

-criar programas de computador que automatizem o processo de desidentificação usando aprendizado de máquina ;

-criar um corpus de prontuários com pacientes pseudônimos, para que seja protegida a identidade das pessoas, o qual será usado como treinamento de algoritmos de desidentificação;

-criar outro corpus de prontuários de pacientes, com dados já desidentificados, o qual possibilite a continuidade de pesquisas em medicina e

informática médica;
-estudar a legislação e as implicações legais que envolvem o trabalho, considerando que o corpus ainda pertence ao HC e sujeito a sua aprovação
para disseminação.

Avaliação dos Riscos e Benefícios:
Riscos:
Os riscos dessa pesquisa está associada com o vazamento parcial ou total das informações pessoais dos pacientes. O vazamento dos dados poderia colocar em risco a integridade física, moral, pessoal, religiosa, cultural e ética dos pacientes.
Benefícios:
Possibilidades de vazamento e exposição de dados sensíveis de pacientes serão eliminados, uma vez que os dados serão completamente anonimizados e desidentificados sem qualquer possibilidade de identificação por parte de outros usuários. Isso desvincula totalmente a relação entre os diagnósticos médicos e os pacientes.
Dessa forma, a dificuldade ao acesso de prontuários de pacientes em pesquisa científica reduzirá sensivelmente. Segundo a lei brasileira, um corpus desidentificado não é mais informação do paciente. Dessa forma serão eliminados todos os aspectos de ética e segurança dos pacientes, referente à condução de pesquisa usando especificamente os dados disponibilizados, serão eliminados. Pesquisadores, médicos, desenvolvedores de aplicações, dentre outros profissionais podem ter acesso à informação sem necessidade de submissão aos comitês de ética e pesquisa.
O corpus disponibilizado publicamente se apresentará como uma fonte confiável e fidedigna de informação médica. Outros médicos poderão consultar procedimentos e tratamentos realizados pelo Hospital Felício Rocho sem comprometer os profissionais

Comentários e Considerações sobre a Pesquisa:
O pesquisador principal atendeu todas as pendências definidas pelo CEP/HFR a saber:
CARTA DE RESPOSTA ÀS PENDÊNCIAS
Título do projeto: Tratamento da informação médica para atendimento à necessidade de
privacidade: desidentificação textual de prontuários eletrônicos na língua portuguesa do Brasil
Pesquisador responsável: Maurício Barcellos Almeida
Data: 04/01/2021
Em resposta às pendências informadas, seguem as informações necessárias:
Pendência 1: No resumo está descrito que “Dois corpora serão criados utilizando dados em

parceria
com o Hospital Felício Rocho da UFMG". Este CEP esclarece que a instituição de saúde Hospital Felício Rocho não apresenta relações com a Universidade Federal de Minas Gerais como pode ser implícito na frase acima. Portanto, solicita a correção do texto.
Resposta a pendência 1: Estou ciente de que o Hospital Felício Rocho não possui relações com a UFMG e que os dados apresentados no texto referente a isso constitui em erro de projeto. Dessamaneira, faz-se valer as considerações emitidas pela instituição co-participante.
Pendência 2: Nos objetivos secundários, solicita-se esclarecimento do texto "o corpus ainda pertenceao HC e sujeito a sua aprovação para disseminação".
Resposta a pendência 2: Gostaria de esclarecer que o corpus pertence ao Hospital Felício Rocho e não ao HC e que o texto apresenta erro de projeto. Quanto a aprovação para disseminação, tal hipótese está descartada conforme discutido na pendência 3.
Pendência 3: O colegiado deste CEP mostrou preocupação quanto ao fato do pesquisador tornar público os dados de prontuários médicos pertencentes a pacientes tratados no Hospital Felício Rocho. Mesmo que não identificados e em acordo com a atual legislação brasileira, há riscos relacionados a publicização dos dados. Apesar de tornar o paciente não identificável, algumas informações podem levar a identificação do tipo de tratamento utilizado, da equipe de saúde que prestou assistência, do desfecho ocorrido, entrou outras informações. Desta forma, solicitamos que o
pesquisador melhor esclareça: (1) os riscos de tornar públicos dados de prontuários dos pacientes,mesmo que não identificados; (2) como serão armazenados os dados desidentificados e (3) como pretende tornar público os dados desidentificados.
Resposta a pendência 3: Gostaria de afirmar que estou abandonando a hipótese de publicizar os dados coletados no Hospital Felício Rocho. Estou fazendo um direcionamento de pesquisa para desenvolver a metodologia de desidentificação para o sumário de alta brasileiro. Dessa maneira, os dados coletados no Hospital Felício Rocho serão usados apenas internamente para testar e validar a metodologia em desenvolvimento. Portanto, reafirmo, que os dados desidentificados não serão
disponibilizados ao público. Assim sendo, fica dispensado os esclarecimentos (1), (2) e (3) solicitados na pendência 3.
Pendência 4: Este colegiado solicita esclarecimentos sobre como o pesquisador terá acesso aosdados de prontuários de pacientes do Hospital Felício Rocho para, posteriormente, realizar a

desidentificação.
Resposta a pendência 4: A subinvestigadora da pesquisa e funcionária do Hospital Felício Rocho, Amanda Damasceno de Souza, realizará a solicitação dos dados junto a equipe de informática do HFR, após a aprovação da pesquisa pelo CEP. Para a coleta de dados no PEP do HFR será desenhada umaestratégia de Business Intelligence (BI), com a finalidade de recuperar somente os dados de interesse da pesquisa. Como resultado do processo de BI, um banco de dados relacional em PostgreSQL dos sumários de alta, será exportado e enviado por e-mail. Esta tarefa já foi anteriormente realizada com os dados de 2018 da Ginecologia para a pesquisa “Terminologias para padronização de sistemas em saúde e sua conexão com ontologias de mapeamento para fins de interoperabilidade de dados clínicos do prontuário eletrônico do paciente”.
Pendência 5: Este colegiado solicita o parecer do setor de informática do Hospital Felício Rocho a respeito do projeto de pesquisa para ter conhecimento sobre: (1) se é possível a liberação de dados dos prontuários à equipe de pesquisa, (2) se há concordância em realizar a liberação de dados desta instituição e (3) os processos sobre como será feita a eventual liberação de dados à equipe de pesquisa. Cabe ao investigador ou à equipe de pesquisa providenciar tal parecer para a apreciação
deste Comitê de Ética.
Resposta a pendência 5: O projeto anexado na Plataforma Brasil já contém em anexo o documento chamado “Declaração de Anuência e Infraestrutura da Instituição” onde o setor de informática do Hospital Felício Rocho diz estar ciente do projeto. É importante enfatizar, no entanto, que os pesquisadores do projeto assumem o compromisso de avaliação dos dados pelo CEP, mesmo com os dados desidentificados. Dessa maneira, exclui-se do projeto o item que afirma que o projeto não precisará de passar pelo Comitê de Ética em Pesquisa do HFR. Será anexada uma versão 2 do projeto
com as alterações solicitadas pelo Comitê de Ética em Pesquisa do Hospital Felício Rocho.
Cordialmente,
Guilherme Noronha – Pesquisador
Pelo exposto considerando um projeto de doutorado relevante,já aprovado na COEP UFMG e por atender as pendências do CEP/HFR somos SMJ do Colegiado do CEP/HFR pela aprovação do projeto.

Considerações sobre os Termos de apresentação obrigatória:
Os termos estão adequados e a documentação completa.

Endereço: Rua Uberaba, n° 500, 5° andar, Núcleo de Ciências da Saúde Felício Rocho
Bairro: Barro Preto **CEP:** 30.180-082
UF: MG **Município:** BELO HORIZONTE
Telefone: (31)3514-7626 **Fax:** (31)3514-7626 **E-mail:** cep@feliciorocho.org.br

Recomendações:

Enviar relatórios semestrais/final ao CEP/HFR.

Conclusões ou Pendências e Lista de Inadequações:

Não se aplica

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1660151.pdf	20/01/2021 12:40:03		Aceito
Outros	Carta_de_Resposta_a_Pendencia.pdf	20/01/2021 12:25:54	GUILHERME FRANCIS DE NORONHA	Aceito
Projeto Detalhado / Brochura Investigador	Pr__Projeto_COEPE_HFR.pdf	20/01/2021 12:08:46	GUILHERME FRANCIS DE NORONHA	Aceito
Folha de Rosto	folha_de_rosto.pdf	26/11/2020 15:33:26	GUILHERME FRANCIS DE NORONHA	Aceito
Declaração de Pesquisadores	declara_MBA.pdf	23/11/2020 16:01:31	Mauricio Barcellos Almeida	Aceito
Declaração de Instituição e Infraestrutura	declara_HFR_MBA.pdf	23/11/2020 15:58:44	Mauricio Barcellos Almeida	Aceito
Outros	Carta_de_Emenda.pdf	26/08/2020 15:16:59	GUILHERME FRANCIS DE NORONHA	Aceito
Projeto Detalhado / Brochura Investigador	Pr__Projeto_COEPE.pdf	25/08/2020 14:19:29	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Guilherme_Francis_de_Noronha.pdf	25/08/2020 14:15:06	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Zilma_Silveira_Nogueira_Reis.pdf	25/08/2020 14:13:31	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_Heliana_Ribeiro_de_Mello.pdf	25/08/2020 14:13:00	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	Currculo_do_Sistema_de_Currculos_Lattes_Maurcio_Barcellos_Almeida.pdf	24/08/2020 16:42:41	GUILHERME FRANCIS DE NORONHA	Aceito

Continuação do Parecer: 4.529.223

Outros	Curriculo_Lattes_Amanda_Damasceno_de_Souza.pdf	13/08/2020 16:04:12	Amanda Damasceno de Souza	Aceito
Outros	TERMO_DE_COMPROMISSO_PARA_UTILIZAO_DE_DADOS_DE_ARQUIVO.pdf	28/07/2020 14:54:04	GUILHERME FRANCIS DE NORONHA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	DECLARAO_DE_DISPENSA_DE_TERMO_DE_CONSENTIMENTO_LIVRE_E_ESCLARECIDO.pdf	28/07/2020 14:40:20	GUILHERME FRANCIS DE NORONHA	Aceito
Outros	resposta_240319.pdf	24/03/2019 18:21:55	Mauricio Barcellos Almeida	Aceito
Outros	tecnologiadeinformacao.pdf	19/02/2019 10:25:48	EDILAINE APARECIDA DE SOUZA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	just.pdf	06/11/2018 13:38:01	Mauricio Barcellos Almeida	Aceito
Parecer Anterior	2.pdf	05/11/2018 17:35:55	Mauricio Barcellos Almeida	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

BELO HORIZONTE, 08 de Fevereiro de 2021

Assinado por:
Daniel Mendes Pinto
(Coordenador(a))

Endereço: Rua Uberaba, n° 500, 5° andar, Núcleo de Ciências da Saúde Felício Rocho
Bairro: Barro Preto **CEP:** 30.180-082
UF: MG **Município:** BELO HORIZONTE
Telefone: (31)3514-7626 **Fax:** (31)3514-7626 **E-mail:** cep@feliciorocho.org.br