5. EXTRAÇÃO E ANÁLISE DE DADOS REGISTRADOS EM TEXTO LIVRE DE PRONTUÁRIO ELETRÔNICO DO PACIENTE POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL

Amanda Damasceno de Souza
Eduardo Ribeiro Felipe
Fernanda Farinelli

5.1 INTRODUÇÃO

Na área de saúde, a informação de qualidade para tomada de decisão é essencial por prestar melhor assistência ao paciente. Com isso, as Tecnologias de Informação (TI) impactaram consideravelmente a prática clínica ao se integrarem como peças fundamentais às rotinas dos profissionais da área de saúde (Shortliffe, 2014). A informação de assistência em saúde é registrada no Prontuário Eletrônico de Paciente (PEP), conhecido por vários outros termos: prontuário médico, prontuário nosológico do paciente, prontuário médico do paciente etc., e por termos que dizem respeito à sua documentação: laudo médico, relatório médico, exame médico e registro de saúde (Blobel, 2018; Conselho Regional de Medicina do Distrito Federal, 2006).

Nesse contexto, as terminologias clínicas padronizadas são importantes por realizarem a interface dos dados clínicos com os sistemas de atenção à saúde, entre eles o PEP (Rogers, 2005). Além disso, as terminologias padronizadas são recursos valorosos para possibilitar a interoperabilidade no PEP, ao colaborar na realização de auditoria, pesquisa, *benchmarking* e gerenciamento de resultados para o hospital (Miñarro-Giménez *et al.*, 2019). Schulz *et al.* (2017) citam três tipos de terminologias em saúde e propõem uma metodologia para realizar conexão entre elas: Terminologias de Interface (texto clínico do prontuário ou jargão médico), Terminologias de Referência (vocabulários controlados e/ou ontologias) e Terminologias

de Agregação (CID, SNOMED-CT)[48, 49]. Uma lacuna indicada por Schulz e*t al.* (2017) é encontrar uma forma de promover a conexão entre os dados clínicos presentes nos textos clínicos do PEP e as terminologias clínicas padronizadas. As terminologias clínicas padronizadas podem ser usadas para identificar o significado semântico dos conceitos lexicais em um texto clínico dos prontuários. Além disso, terminologias são importantes ao expandir o conceito, com um ou mais sinônimos, na identificação de abreviações e acrônimos em texto clínico e também para mapear conceitos para terminologias de outros sistemas (Dalianis, 2018a, p. 35). Para possibilitar a interoperabilidade entre as terminologias clínicas é preciso analisar os termos e dados clínicos do PEP e, com isso, buscar uma conexão destes termos para outros artefatos terminológicos[50] da área da saúde. Para solucionar este problema semântico, a Ciência da Informação (CI) e a Ciência da Computação (CC)[51] precisam caminhar juntas.

A CI tem um papel importante na área de saúde, por sua contribuição nas áreas de recuperação, representação e organização da informação. A CI lida com a representação, a armazenagem e a recuperação da informação, com o tratamento da informação na tarefa de "[...] construir interfaces entre os acervos de documentos e informações e seus usuários" (Alvarenga, 2003, p. 25). O bibliotecário que atua na saúde precisa desenvolver habilidades para manipular diferentes suportes de busca e recuperação da informação e, com isso, contribuir para a melhoria do tratamento da informação e o cuidado com a saúde. Cada vez mais a CI necessita caminhar junto com a TI na Recuperação da Informação (RI) de registros médicos presentes no PEP, a fim de possibilitar ao usuário o acesso à informação (Souza, 2021; Campos, 2001).

Os registros médicos clínicos são ricos em informação, e em grande parte são concebidos em formato de texto livre (dados clínicos). Os meios para extrair informação estruturada desses registros em texto livre é um significativo esforço de pesquisa (Zhou *et al.*, 2006). Nesse contexto, o volume

48 CID: Classificação Internacional de Doenças.

49 *SNOMED-CT: Systematized Nomenclature of Medical - Clinical Terms*.

50 Conjunto estruturado de termos por meio de uma metodologia no estabelecimento de relações hierárquicas ou semânticas.

51 Também representada pela TI neste contexto.

de informação produzido na pesquisa acadêmica e na prática clínica há muito exige tratamento computacional. Uma importante fonte de dados de pacientes, relevante para a pesquisa, além de essencial para a gestão das unidades de saúde, é o PEP. Dessa forma, técnicas de processamento de linguagem natural (PLN) são alternativas importantes para lidar com os campos de texto livre dessa fonte dinâmica, onde constantemente se registram novos dados. O PLN envolve processamento inteligente de texto, no qual o computador busca interpretar o que foi escrito em linguagem natural, valendo-se de métodos computacionais linguísticos. Essa abordagem de PLN visa à extração de informação específica de documentos ou coleções de documentos, a fim de serem aplicadas em campos de texto livre dos PEPs (Dalianis, 2018b).

A Mineração de Texto (*Text Mining*) consiste na aplicação de técnicas de garimpagem de dados para obtenção de informações importantes. É um processo que utiliza algoritmos capazes de analisar coleções de documentos escritos em linguagem natural, com o objetivo de extrair conhecimento e identificar padrões. Dentre as técnicas utilizadas, destaca-se o PLN (Dalianis, 2018b).

O processamento de textos clínicos e biomédicos, no âmbito da informática médica, envolve a utilização de métodos baseados em PLN. O *Text Mining* (TM) objetiva encontrar, previamente, fatos desconhecidos no texto ou em coleções de textos, assim como criar hipóteses para serem provadas. O PLN se refere ao processamento inteligente de texto, no qual o computador busca interpretar o que foi escrito em linguagem natural, utilizando, para isso, métodos computacionais linguísticos. Essas duas abordagens, de TM e PLN, se referem à Extração de Informação (EI), que busca encontrar informação específica em um documento ou em coleções de documentos (Dalianis, 2018b). Na análise de informação lexical (textos) do domínio biomédico, Bodenreider (2006) sugere utilizar soluções de *Machine Learning Technique* (ML). Para Dalianis (2018b), o Machine Learning (ML), ou seja, o aprendizado de máquina, é um conjunto de técnicas que usa dois padrões de comportamento, a saber: os supervisionados e os não supervisionados.

Como exemplos de métodos supervisionados, citamos o *Conditional Random Field* (CRF)/Campo Aleatório Condicional (CAC) e o *Support Vector Machine* (SVM)/ Máquina de Vetor de Suporte (MVS). Como métodos não

supervisionados, citamos o *Latent Semantic Analysis* (LSA)[52], o *Latent Semantic Indexing* (LSI)[53], o *Random Indexing*[54] e text clustering[55].

A recuperação do conhecimento, presente nos textos em linguagem natural, é uma tarefa árdua, que envolve técnicas como o PLN. O PLN diz respeito ao processamento inteligente de texto em linguagem natural, com utilização de métodos computacionais linguísticos (Manning; Schütze, 1999).

Com o desenvolvimento crescente das TI, uma equipe médica produz, hoje, uma quantidade de informação maior do que em qualquer outro momento da história. Grande parte dessa informação está em formato texto e digital. A sobrecarga de informação resultante de tanto material disponível impacta na tomada de decisão, sendo necessário utilizar recursos tecnológicos para recuperar o conteúdo relevante. Diante disso, são necessárias ferramentas para extração, análises e organização dos dados e informações (Blake, 2011). Assim, as Ciências Sociais Aplicadas com áreas como a CI assumem papel importante, ao considerar a informação como força construtiva na sociedade, por sua intensa e crescente aplicação em áreas computacionais em Saúde.

O presente estudo se insere nesse contexto, e descreve tecnologias de informação e comunicação (TICs) utilizadas na extração e análise dos dados, nas etapas metodológicas de pesquisa acadêmica. O objetivo deste capítulo é demonstrar o uso de TICs, a saber, PLN e ferramentas de gestão de projeto, na metodologia de pesquisa *stricto sensu* (doutorado) no âmbito da CI aplicada em textos livres de PEP (intitulado Terminologia de Interface) no campo da saúde, especificamente na área de Ginecologia e Obstetrícia. A seguir, o tópico dois aborda o detalhamento da tecnologia que foi empregada na pesquisa, o tópico três aborda um exemplo prático da metodologia da pesquisa *stricto sensu*, ao abordar etapas de extração e análise de dados da Terminologia de Interface por meio de PLN no contexto da Ginecologia e Obstetrícia; o tópico quatro faz uma breve revisão

52 Em português: Análise Semântica Latente (ASL).

53 Em português: Indexação semântica latente.

54 Em português: indexação aleatória.

55 Em português: agrupamento de texto.

de literatura de pesquisas que utilizaram o PNL em terminologias clínicas. Por fim, o tópico cinco consiste nas considerações finais, seguidas das referências e dos anexos.

5.2 SOBRE A TECNOLOGIA EMPREGADA NA PESQUISA

A análise do problema, bem como do ambiente informacional, é a etapa inicial que fundamenta as escolhas e estratégias do modelo tecnológico. Com o foco na extração de dados em texto livre no contexto da saúde de Ginecologia e Obstetrícia, identificou-se que o modelo de registro das informações ultrapassa a complexidade da linguagem natural enquanto representação do conhecimento. A inserção de códigos médicos, abreviações, siglas e, por muitas vezes, o uso do jargão médico, demonstrou a dificuldade de análise do conjunto de dados. Alinhado a esse cenário, o cuidado com a exposição de dados sensíveis elevou o cuidado com a extração das informações de seu banco de dados original.

5.2.1 PROCESSO DE AQUISIÇÃO DOS DADOS

Optou-se, portanto, por uma estratégia de extração e tratamento dos dados autorizados pelo hospital. A Figura 1 permite visualizar graficamente o processo de Aquisição das informações para o experimento, onde:

1. A instituição de saúde avalia o experimento e, por meio do registro de processo administrativo, aprova o uso de dados reais em sua comissão de ética em pesquisa. Cabe ressaltar que pesquisas realizadas com dados de PEP necessitam de aprovação do Comitê de Ética em Pesquisa (CEP). Na área das Ciências da Saúde, as pesquisas, envolvendo seres humanos, são subordinadas à Resolução CNS nº 466, de 12 dezembro de 2012, e às especificidades da área de Ciências Humanas e Sociais, e contempladas pela Resolução nº 510, de 07 de abril de 2016 (Souza, 2022). Os dados utilizados neste estudo foram coletados de PEP de Hospital Privado, aprovados pelo CEP local e pelo número do CAAE: 03384418.0.0000.51259. A comissão instituiu regras a serem seguidas pelo departamento de TI da instituição.

2. Após a aprovação da comissão, a pesquisadora analisou o banco de dados da instituição a fim de identificar as informações passíveis de análise por meio de critérios alinhados ao objetivo da pesquisa.

3. Alinha-se com a gerência de tecnologia da informação uma estratégia de extração das informações do banco de dados oficial. Realizam-se processos de tratamento para alinhamento com a Lei Geral de Proteção de Dados[56] (LGPD) e a desidentificação de dados sensíveis, incapacitando a base de dados extraída na identificação de pacientes.

4. Este conjunto de dados extraído do banco de dados principal foi gravado em um banco denominado *PostgreSQL*, compactado para transferência e posterior uso no computador pessoal da pesquisadora.

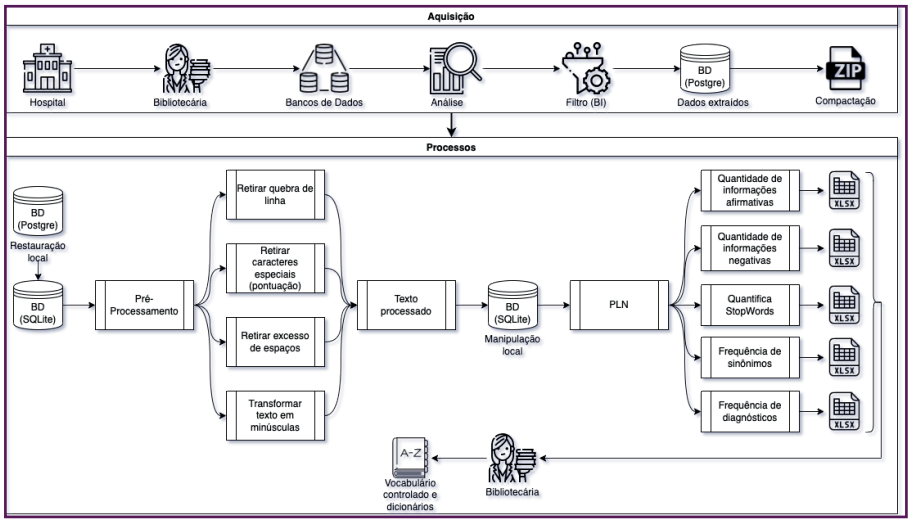
Pode-se ressaltar, nesse processo, a dinâmica tecnológica alinhada aos requisitos da pesquisa. Para a pesquisa no banco de dados principal da instituição foi usada a linguagem SQL[57]. Identificados os dados principais, foram desenvolvidas *queries*[58] no ambiente de *Business Intelligence* (BI) para conexão e seleção das informações. O conjunto de dados resultante (*dataset*) foi exportado para um banco de menor porte, como citado anteriormente no item IV.

56 Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709compilado.htm. Acesso em: 3 out. 2023.

57 Disponível em: https://www.ibm.com/docs/en/i/7.4?topic=concepts-structured-query-language. Acesso em: 3 out. 2023.

58 Formulações em linguagem computacional SQL para consultas ao banco de dados.

Figura 1 - Principais processos informacionais da metodologia



Fonte: Elaborado pelos autores (2023).

5.2.2 PROCESSOS TÉCNICOS

De posse das informações, definiu-se uma abordagem local para a pesquisa. A Figura 1 também demonstra os grupos "Aquisição" e "Processos", responsáveis pela manipulação dos dados. Como processos iniciais, destacam-se, de forma sintetizada: a) a extração da informação do banco de dados do Hospital (aquisição); b) a restauração dos dados em ambiente local; e c) a adequação do formato do banco de dados para manipulação.

5.2.2.1 AQUISIÇÃO

Os dados do Hospital estão armazenados em um sistema de grande porte. Após cuidadosa análise, para preservar as normas de sigilo de dados pessoais dos pacientes e exclusão de dados sensíveis, a realização da extração deu-se a partir de um recorte conceitual (anamnese e evolução). Os registros foram exportados para um banco de dados *PostgreSQL*[59]. Esse artefato (banco de

59 Disponível em: https://www.postgresql.org/. Acesso em: 3 out. 2023.

dados) foi o resultado do processo de identificação dos registros de interesse do projeto (anamneses e evoluções), constituindo o recorte a ser analisado. Esse arquivo foi enviado à pesquisadora em formato compactado, no formato ZIP, como fonte de dados principal para fomento da pesquisa. Tal etapa corresponde ao grupo "aquisição" da Figura 1.

5.2.2.2 RESTAURAÇÃO E ADEQUAÇÃO DO BANCO DE DADOS EM AMBIENTE LOCAL

Por se tratar de um banco de dados dependente de um *software* "servidor", optou-se por migrar os dados para um banco mais simples, denominado *SQLite*. O formato permitiu grande flexibilidade no acesso às informações por não precisar de um "serviço servidor" instalado no sistema operacional do computador pessoal da pesquisadora.

5.2.2.3 PRÉ-PROCESSAMENTO

De posse da estrutura, iniciou-se o tratamento dos dados por meio do desenvolvimento de algoritmos na linguagem de programação *Python*. Destaca-se o uso de bibliotecas para tarefas específicas, como: i) *openpyxl*[60] para criação/manipulação de planilhas eletrônicas; ii) *wordcloud*[61] para geração de nuvem de palavras e apresentação gráfica de informações textuais em destaque terminológico; e iii) *nltk*[62], uma reconhecida biblioteca para diversas funcionalidades em PLN.

Além disso, destaca-se o desenvolvimento dos algoritmos para tratamento e extração de dados específicos à pesquisa, que utilizaram as técnicas de PLN. Visto que os dados foram exportados em seu formato real, sem tratamento, alguns processos foram fundamentais para adequar um padrão textual e permitir o estabelecimento de regras para que os algoritmos pudessem se comportar de maneira esperada. Ou seja, a linguagem

60 Disponível em: https://openpyxl.readthedocs.io/. Acesso em: 3 out. 2023.

61 Disponível em: https://github.com/amueller/word_cloud. Acesso em: 3 out. 2023.

62 Disponível em: https://www.nltk.org/. Acesso em: 3 out. 2023.

natural, usada na descrição médica, proporciona um texto despadronizado de forma e sintaxe, o qual necessita de uma intervenção antes de realizar a análise, identificação e extração dos dados.

Existem diversas iniciativas para o mapeamento de caracteres e símbolos, como o *American Standard Code for Information Interchange* (ASCII), o *Unicode Transformation Format* (UTF-8), o *American National Standards Institute* (ANSI), Windows-1252, e ISO-8859-1, os quais são padrões que viabilizam a correta exibição de fontes (letras e símbolos) para textos eletrônicos. A maneira como o texto é codificado – denominada *charset* – é importante para proporcionar uma versão final adequada, compreensível e compatível com os mesmos caracteres ou símbolos, codificados pelas terminologias clínicas. O padrão de codificação define a forma como o texto resultante da conversão deve ser representado, além de permitir a leitura e o entendimento humano. Considerando o idioma do projeto, onde tanto a base de dados como as listas terminológicas foram codificadas em português, não houve necessidade de conversão da codificação. Elencam-se, no Quadro 1, os principais processos necessários para adequação dos dados antes da aplicação dos algoritmos de extração.

Quadro 1 - Processos de tratamento textual

Técnica	Objetivo	Referência
stopWords	Retirar termos de pouca ou nenhuma carga semântica à pesquisa	https://github.com/stopwords-iso/stopwords-pt
acentuação	retirar todos os caracteres acentuados e substituir pelo respectivo caracter não acentuado	"á", "à", "ã", "é", "í", "ô", "ç"

Técnica	Objetivo	Referência
excesso de espaços	retirar excesso de espaços em branco a fim de padronizar a tokenização[63] dos termos	expressões regulares
caracteres especiais	retirar caracteres da constante "punctuation" em Python	""!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"""
quebra de linha	Os textos originais foram formatados com quebras de linhas para facilitar o entendimento humano, mas não são necessários ao processamento computacional. Os caracteres para esta formatação "/n" foram excluídos, tornando o texto uma sequência de caracteres (string) para padronizar o parse[64] do texto pelas funções	método replace para cada \n
minúsculas / maiúsculas	padronizar todo o texto em minúscula para permitir matching[65] perfeitos	método lower() - case-folding[66]

Fonte: Elaborado pelos autores (2023).

63 Sequência de caracteres em grupos separados por espaço.

64 Processo de análise sintática pelo *algoritmo* a fim de identificar estruturas textuais.

65 Processo de encontrar combinações de textos (*strings*) idênticas.

66 Disponível em: https://nlp.stanford.edu/IR-book/html/htmledition/capitalizationcase-folding-1.html. Acesso em: 3 out. 2023.

No processo de PLN, porém, destaca-se a *tokenização*. Trata-se de uma técnica de análise textual que permite a quebra de texto, composto por muitas palavras (ou *strings*), em palavras e símbolos separados em *substrings*, resultando em uma estrutura de dados em forma de lista (*array*), a qual permite uma análise individualizada. O resultado dessa etapa é um texto armazenado em nova coluna no banco de dados. A coluna será usada para a próxima etapa, de extração, processamento e análise de informação. O Quadro 2 exibe um exemplo desse tratamento em relação ao registro original.

Quadro 2 – Tratamento dos textos para processamento

Texto original	Texto tratado com pré-processamento
## GINECOLOGIA ## PACIENTE SUBMETIDO A RESSECÇÃO DE TUMOR DE PAREDE ABDOMINAL COM EXERESE DE SEGMENTO DE APONEUROSE , SOB RAQUE ANESTESIA. ATO CIRURGICO SEM INTERCORRENCIAS ENVIADO MATERIAL PARA EXAME ANATOMO-PATOLOGICO	ginecologia paciente submetido a ressecção de tumor de parede abdominal com exerese de segmento de aponeurose sob raque anestesia ato cirurgico sem intercorrencias enviado material para exame anatomo patológico

Texto original	Texto tratado com pré-processamento
PACIENTE 46 ANOS , G1P1A0 , APRESENTANDO METRORRAGIA REFRATARIO AO TRATAMENTO CONSERVADOR. AO ULTRASSOM TRANSVAGINAL PRESENÇA DE VOLUMOSO MIOMA SUBMUCOSO. APRESENTOU ANEMIA AGUDA COM HB- 7,6G/DL PRE-OP—ASA 1 PRE-ANESTESICO—ASA1 PA- 120X80MMHG MUCOSAS HIPOCORADAS E HIDRATADA AFEBRIL BCNRNF SONS RESP NORMAIS ABDOME LIVRE , SEM SINAIS DE IRRITAÇÃO PERITONEAL UTERO AUMENTADO DE VOLUME COM SANGRAMENTO PERSISTENTE CD- HISTERECTOMIA TOTAL COM ANEXECTOMIA UNILATERAL	paciente 46 anos g1p1a0 apresentando metrorragia refratario ao tratamento conservador ao ultrassom transvaginal presença de volumoso mioma submucoso apresentou anemia aguda com hb 7 6g dl pre op asa 1 pre anestesico asa1 pa 120x80mmhg mucosas hipocoradas e hidratada afebril bcnrnf sons resp normais abdome livre sem sinais de irritação peritoneal utero aumentado de volume com sangramento persistente cd histerectomia total com anexectomia unilateral

Fonte: Dados da pesquisa (2021).

Como padrão de saída, para facilitar a análise especialista, os dados foram gravados em planilhas eletrônicas independentes (Figura 1).

O código-fonte do projeto está gravado no repositório digital *GitHub* para fins de fomento à pesquisa e discussões com outros pesquisadores. As

ferramentas utilizadas no desenvolvimento técnico da pesquisa foram relacionadas a seguir (Quadro 3):

Quadro 3 – Ferramentas utilizadas na condução técnica da metodologia da pesquisa

Ferramenta	Utilização	Acesso
GitHub	Repositório digital para salvar os algoritmos e o conjunto de resultados da extração de dados utilizando as técnicas de PLN foram salvas no repositório digital	https://github.com/amandadsouza/RiLN
PostgreSQL	Servidor de banco de dados para restaurar os dados	https://www.postgresql.org/
Visual Studio Code	Ambiente de desenvolvimento para desenvolvimento de software	https://code.visualstudio.com/
DBeaver	Interface para interação com os diferentes formatos de bancos de dados	https://dbeaver.com/

Fonte: Elaborado pelos autores (2023).

Esses processos relacionados ao tratamento textual constituem o pré-processamento dos dados recebidos pelo hospital.

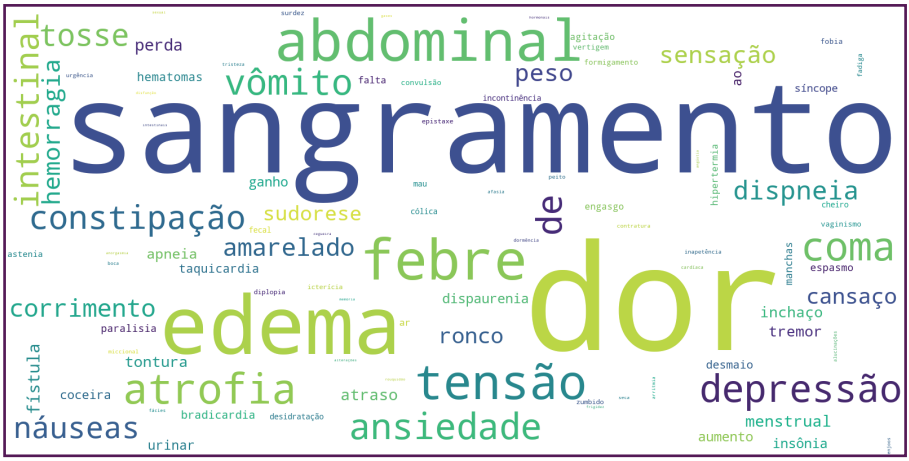
Após a realização deste ajuste textual, procedeu-se com a análise dos dados e o desenvolvimento de algoritmos específicos para o processamento das informações. Na medida que os dados foram manipulados, dois modelos de saída para análise foram produzidos.

5.2.3 PROCESSOS DE SAÍDA (APRESENTAÇÃO)

O modelo de saída para análise se deu por meio de apresentação analítica com gravação de dados tabulares em planilha eletrônica. As análises quantitativas e modelos de frequência foram analisados de forma individual (por registro no modelo de banco de dados) e por agrupamento, permitindo ao especialista um modo de rastreio e percepção de validação do algoritmo que, por vezes, foi aperfeiçoado.

Outro modelo de saída foi a representação das frequências terminológicas no formato nuvem de palavras, conforme a Figura 2. Este modo de visualização permite destacar os termos mais frequentes em determinados contextos, além de facilitar o entendimento do usuário final. Neste processo foi usada a biblioteca *Python word_cloud*[67].

Figura 2 – Nuvem de Palavras para representação dos termos da Terminologia de Interface



Fonte: Dados da pesquisa de doutorado de Souza (2021).

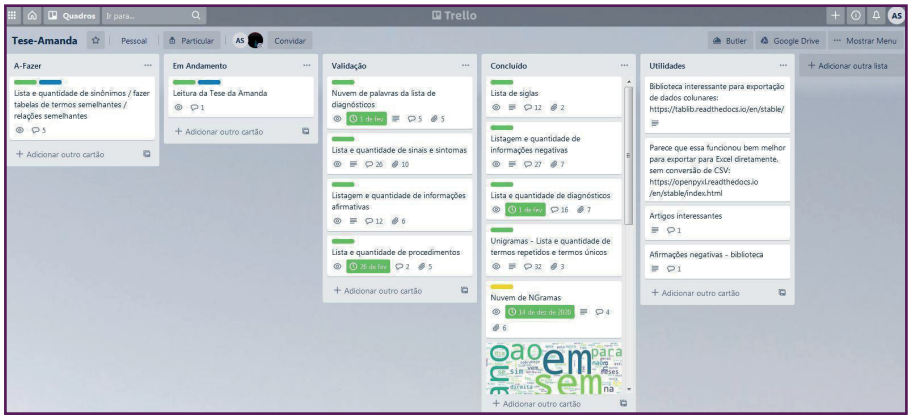
67 Disponível em: https://github.com/amueller/word_cloud. Acesso em: 3 out. 2023.

5.2.4 PROCESSOS ADMINISTRATIVOS

Na condução da pesquisa, diferentes especialistas se uniram a fim de contribuir para o propósito do trabalho. Nesse aspecto, outras ferramentas tecnológicas foram usadas para a gestão de tarefas e documentos.

Para o controle de tarefas em equipe, utilizou-se o *software Trello*[68]. Este programa permitiu estabelecer grupo de tarefas, sua cronologia e *status* para fins de alinhamento de tarefas e responsabilidades. Por meio do método Kanban[69] é possível identificar facilmente tarefas em andamento, bem como seu status e responsável.

Figura 3 - Divisão de tarefas da metodologia pesquisa no *Trello*



Fonte: Captura de tela do Trello utilizada para a pesquisa de doutorado de Souza (2021).

68 Disponível em: https://trello.com/pt-BR. Acesso em: 3 out. 2023.

69 Disponível em: https://blog.trello.com/br/metodo-kanban. Acesso em: 3 out. 2023.

Para o controle dos diversos arquivos gerados (excetuando o código fonte), foi usado a ferramenta *Google Drive* (Quadro 4).

Quadro 4 – Ferramentas utilizadas nos processos administrativos

Ferramenta	Utilização	Acesso
Trello	Software para o controle de tarefas em equipe necessárias na condução da metodologia da pesquisa	https://trello.com/b/ur2kJ7Vm/tese-amanda
Google Drive	Software para centralizar arquivos "em nuvem" para compartilhamento de dados e backup	disponível para usuários do gmail

Fonte: Elaborado pelos autores (2023).

Os processos mencionados nesta seção configuram os aspectos técnicos e operacionais para manipulação dos dados em face de uma necessidade intelectual. A próxima seção abordará os problemas e motivações na adoção das tecnologias para a solução dos problemas específicos no ambiente da pesquisa.

5.3 A TECNOLOGIA E A PESQUISA

Nesta seção, apresentam-se as etapas metodológicas aplicadas de extração e análise de dados por meio de PLN no contexto da Ginecologia e Obstetrícia na pesquisa *stricto sensu*, sendo elas: 1) Definir lista preliminar de termos para delimitar algoritmo; 2) Extrair de dados a partir de ferramenta e técnicas automáticas de PLN: Algoritmos desenvolvido utilizando linguagem *Python*; 3) Analisar os termos extraídos: frequências absolutas e relativas; e 4) Apresentação dos termos por meio de nuvem de palavras e tabelas.

1. Definição das listas preliminares de termos para delimitar algoritmo do PLN

Para delimitar os algoritmos na busca por dados, foram criadas listas preliminares de termos da Ginecologia e Obstetrícia em relação a: sinais e sintomas e diagnóstico, entre outros tipos de termos. As listas de termos preliminares foram criadas junto à equipe do Núcleo Integrado de Pesquisa e Tratamento da Endometriose (NIPTE)[70] e com auxílio de membros da Clínica de Ginecologia e Obstetrícia do local da pesquisa. As listas foram desenvolvidas também com base em terminologias utilizadas em formulários criados pelo NIPTE, formulários da clínica de Ginecologia e Obstetrícia de coleta de dados no REDCap[71], do sistema MV-PEP, anamnese, evolução, com validação e correção de médicos da Ginecologia e Obstetrícia. A listagem de termos do contexto específico da Obstetrícia, foram utilizadas as terminologias de protocolos clínicos e manuais, por exemplo: Secretaria de Estado de Saúde de Minas Gerais e Associação de Ginecologistas e Obstetras de Minas Gerais (2013), Peixoto (2014), Brasil (2004), Brasil (2016), Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (2016) e Matos *et al.* (2017).

Em relação a doenças/diagnóstico, a CID-10 nas letras N, Q e Z, classifica as doenças relacionadas à especialidade ginecológica, assim na lista preliminar de diagnósticos foi utilizada a CID-10 e indicação de termos pelos ginecologistas do hospital (Organização Mundial da Saúde, 1994).

Em relação aos sinais e sintomas, Souza e Teixeira (2017) abordam que a consulta ginecológica está relacionada a três queixas principais: sangramentos anormais, corrimentos patológicos e dores pélvicas. Assim, os sinais e sintomas referentes a essas queixas deverão ser notificados na anamnese do prontuário. Para a lista de sinais e sintomas também foi utilizado: *National Library of Medicine* (NLM) *Classification 2020 Summer Edition* (Willis, 2019), *Wikipédia*[72], Falcão Júnior *et al.* (2017) e a CID-10 (Organização Mundial da Saúde, 1994) (ANEXO A). Para a pré-lista de

70 Disponível em: https://www.feliciorocho.org.br/servicos/endometriose. Acesso em: 3 out. 2023.

71 Disponível em: https://redcap.feliciorocho.org.br/redcap/index.php. Acesso em: 3 out. 2023.

72 SINAL MÉDICO. *In*: WIKIPÉDIA: a enciclopédia livre. [*S. l.*]: Wikimedia Foundation, 17 ago. 2018. Disponível em: https://pt.wikipedia.org/wiki/Sinal_m%C3%A9dico. Acesso em: 12 out. 2020.

sinais e sintomas é importante incluir sobre os sistemas: circulatório e respiratório; digestivo e abdômen; pele e tecido subcutâneo; nervoso e músculo esquelético; e urinário. Incluir também termos sobre: cognição, percepção, estado emocional e comportamento; fala e voz; sinais e sintomas gerais. Esta pré-lista deve ser verificada por um ginecologista, ou seja, especialista de domínio.

Após projetar os algoritmos, foram realizadas diferentes interações, para extração de termos em relação aos dois tipos documentos da Terminologia de Interface (anamnese e evolução). Foram detectados: a) a presença de sinais e sintomas; b) diagnósticos; e c) termos mais frequentes e termo únicos.

2. Extrair de dados a partir de ferramenta e técnicas automáticas de PLN: Algoritmos desenvolvido utilizando linguagem Python

Passo 1 – Frequência de diagnósticos: Para extrair quais eram diagnósticos e sua quantidade nos documentos eletrônicos (PEPs), foi criada uma lista dos diagnósticos em arquivo texto, que por sua vez foram lidos pelo algoritmo, a fim de criar uma lista (*array*). A leitura dos documentos no banco de dados foi segmentada por tipo de análise ("Anamnese" e "Evolução"). Percorre-se, portanto, cada lista do banco de dados e para cada documento, verifica-se se os diagnósticos (já disponíveis na memória) estão presentes. Uma estrutura de dados organizada por chave: valor, denominada dicionário, em linguagem de programação *Python*, permite armazenar diagnóstico (chave) e sua quantidade (valor) encontrada. Esta estrutura foi usada para posterior gravação, em arquivo de formato planilha eletrônica.

Passo 2 – Frequência de sinais e sintomas: de forma semelhante ao processo de diagnóstico, uma lista de termos sinais e sintomas ginecológicos foi desenvolvida pela pesquisadora. A lista de sinais e sintomas permitiu identificar estes termos nos documentos e levantar sua quantidade para armazenar o resultado em arquivo.

Passo 3 – Quantidade de termos repetidos/únicos: A estratégia para identificar os termos repetidos e únicos deu-se pela estratificação de cada documento em tokens e, a partir dessa grande lista, indicar para cada item da lista, sua repetição ou participação única.

Após a realização dos passos 1, 2 e 3 foi necessário analisar os termos extraídos por meio do PLN.

3. Análise dos termos extraídos: frequências absolutas e relativas

A etapa de extração de termos, descrita anteriormente permitiu extrair: a) frequência de diagnósticos, (para essa tarefa foi construída uma lista de termos para delimitar o *algoritmo*); b) frequência de sinais e sintomas, (para essa tarefa foi construída uma lista de termos para delimitar o *algoritmo*); c) frequência de termos únicos e repetidos, (para essa tarefa não foi necessária uma lista de termos para delimitar o *algoritmo*).

Posteriormente ao PLN foi realizada estatística dos termos extraídos pelos algoritmos, para fins de análise de frequência absoluta (f_a quantidade de vezes que cada termo aparece) e frequência relativa (f_r é o percentual de vezes que cada f_a aparece em relação ao total da amostra n). Foram avaliadas as seguintes variáveis:

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos sobre sinais e sintomas (total) na denominada Terminologia de Interface, ou seja, texto clínico do prontuário ou jargão médico (Schulz *et al.*, 2017);

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos sobre diagnósticos (total) na Terminologia de Interface;

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos repetidos na Terminologia de Interface;

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos únicos na Terminologia de Interface;

- Listar frequência relativa dos termos únicos, termos repetidos, sinais e sintomas, e diagnóstico, em formato de tabelas e nuvem de palavras.

Os cálculos de frequência absoluta e frequência relativa foram realizados utilizando a fórmula (Figura 5).

Figura 5 – Fórmula para cálculo da frequência

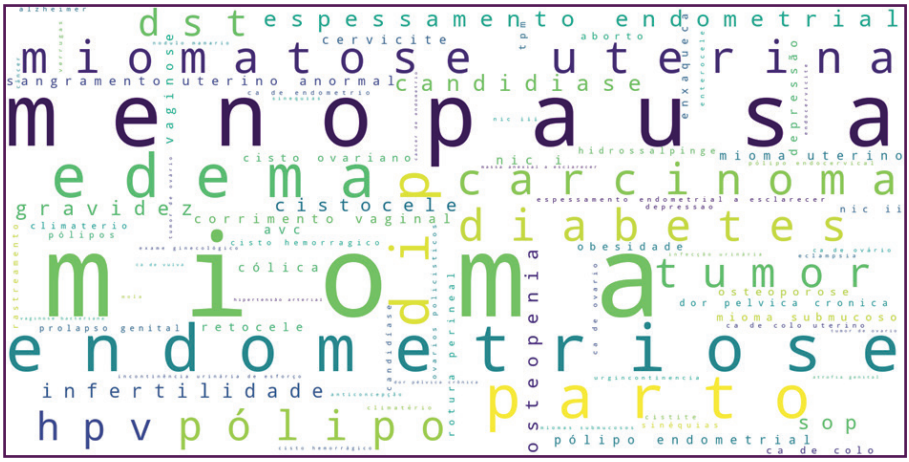
$$x = \frac{fa * 100}{n}$$

Fonte: Elaborado pelos autores (2023).

Apresentação dos termos por meio de nuvem de palavras e tabelas

Após a análise das frequências dos termos (Terminologia de Interface), esses os resultados foram ilustrados conforme exemplos da Figura 6 e Tabela 1.

Figura 6 – Nuvem de palavras de diagnóstico da Terminologia de Interface



Fonte: Souza (2021, p. 158).

Esta representação em nuvem de palavras ilustra quais os termos mais representativos nas notificações em campos abertos do PEP, ou seja, na

Terminologia de Interface. Já a Tabela 1 representa, de forma quantitativa, os termos mais frequentes.

Tabela 1 – Exemplo da frequência absoluta de diagnóstico da Terminologia de Interface

Termo	*f*a	Termo	*f*a
mioma	1986	espessamento endometrial	274
menopausa	1199	candidiase	257
endometriose	948	gravidez	207
parto	806	sop	206
edema	800	infertilidade	200
dip	602	cistocele	180
miomatose uterina	419	osteopenia	178
diabetes	402	sangramento uterino anormal	169
carcinoma	390	corrimento vaginal	168
pólipo	351	NIC I	158
tumor	349	vaginose	151
hpv	306	avc	150
dst	277	cólica	150

Fonte: Souza (2021).

A listagem de termos únicos foi importante para identificar os termos não recuperados pelo algoritmo. A Tabela 2 descreve as frequências absolutas e relativas da Terminologia de Interface.

Tabela 2 – Frequência absoluta (n) e frequência relativa (%) de termos únicos e repetidos na Terminologia de Interface

Variáveis	Terminologia de Interface	
	n	%
Termos únicos	16.248	1,19
Termos repetidos	1.348.116	**98,81**
Total	**1.364.364**	**100**

Fonte: Souza (2021).

O tópico quatro a seguir aborda estudos semelhantes na literatura.

5.4 ANÁLISE TERMINOLÓGICA DE TEXTO CLÍNICO POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL: UM BREVE RELATO DA LITERATURA

Pesquisas envolvendo o processamento de texto clínico (*Text Mining*) para conectar terminologias de referência e agregação já foram realizadas. Soderland *et al.* (1995) cita um trabalho da *The National Center for Intelligent Information Retrieval* (CIIR) da *University of Massachusetts* na cidade de Amherst, na mineração de textos de PEP para realizar uma classificação automatizada com a utilização da terminologia de agregação, CID-9. A pesquisa buscou automatizar os códigos CID-9 para sumários de alta hospitalar, na qual foi realizada a tarefa de automatizar a "compreensão do conteúdo dos sumários para rotular frases que contenham informação relativa a (1) diagnóstico e (2) sinais ou sintomas de doença" (Soderland *et al.*, 1995, p. 1, tradução nossa). Pesquisas que envolvem extração de informação de texto clínico de PEP foram realizadas por Kim *et al.* (2017), Wang *et al.* (2012), Zhou *et al.* (2006), Meystre *et al.* (2010), entre outros. No estudo de Kim *et al.* (2017), utilizou-se o TM para extrair informação sobre intervenção coronária percutânea. Wang *et al.* (2012) utilizaram técnica de ML por meio do *Support Vector Machine* (SVM), para extrair resultados

de diagnóstico de textos clínicos do PEP sobre angiografia coronariana e câncer de ovário.

Um estudo de Zhou *et al.* (2006) descreveu um sistema de extração de Informação Médica (MedIE), que extraiu uma variedade de informações e registros clínicos de textos clínicos do paciente sobre queixas de doenças da mama. Meystre *et al.* (2010) fizeram uma revisão da literatura sobre pesquisas recentes em desidentificação de documentos de texto clínico narrativo em PEP. Estudos como esses demonstram a importância de se analisar a necessidade da conexão de dados entre Sistemas de Informação em Saúde (SISs).

Outro trabalho relevante que retrata um conjunto de ações direcionado à aplicação de PLN em textos biomédicos e mineração de textos, está em (Kafkas; Toonsi; Alsaedi, 2023). Destaca-se o uso dos corpus derivados do algoritmo da *Google Bidirectional Encoder Representations from Transformers*[73] (BERT) e os processos de transformação, tratamento, *tokenização*, reconhecimento de entidades, normalização e identificação das relações no texto em linguagem natural.

Por fim o estudo de Souza *et al.* (2022) abordou o trabalho do Bibliotecário da Saúde, junto à equipe médica, em pesquisas sobre as terminologias clínicas em prontuários na área de saúde, no qual este profissional precisa saber recuperar dados a partir da PLN e técnicas de inteligência artificial com a finalidade de contribuir para a melhoria do tratamento da informação, no cuidado em saúde. Na pesquisa foram identificados por meio de vocabulários terminológicos com apoio de algoritmos em *Python* termos relacionados que envolviam: presença de sinais, sintomas, expressões negativas e afirmativas, termos únicos e mais frequentes, siglas e abreviaturas, entre outros.

73 Disponível em: https://arxiv.org/pdf/1810.04805.pdf. Acesso em: 3 out. 2023.

5.5 CONSIDERAÇÕES FINAIS

Este estudo relatou uma metodologia de PLN e ferramentas de gestão de projeto, aplicada a pesquisa *stricto sensu* (doutorado) no âmbito da CI e no campo da saúde, especificamente na área de Ginecologia e Obstetrícia. Para isso foi demonstrado as etapas para se realizar análise terminológica de textos clínicos de campo aberto do PEP, denominada Terminologia de Interface, com dados reais de hospital privado.

A análise de textos clínicos é uma área de grande importância no contexto da Medicina. Neste trabalho, observou-se que a análise das informações registradas no PEP permitiu traçar as características da terminologia clínica, que retrataram as principais características (diagnósticos, sinais e sintomas) em um público específico da área de saúde (Ginecologia e Obstetrícia). Essa informação foi disponibilizada à instituição hospitalar a fim de embasar decisões futuras que impactem em investimentos estruturais, na contratação de especialistas, na seleção de pacientes para estudos clínicos, entre outros. Os resultados também permitiram a análise dos artefatos terminológicos (Terminologia de Interface), evidenciando a necessidade de atualização e aproximação da realidade de representação do diagnóstico, sinais e sintomas em terminologia clínicas como a CID-10 e em linguagem natural (jargão médico). Ou seja, os artefatos terminológicos precisam evoluir na medida que os profissionais da área necessitam representar seu conhecimento e informações na descrição do diagnóstico, sinais e sintomas e evolução do paciente.

Reconhece-se, contudo, que as limitações no processamento e análise da linguagem natural ainda são grandes no contexto da sintática e semântica. Problemas como erros ortográficos, abreviações, mnemônicos, pontuações, quebras de linha, entre outros citados no trabalho original, foram evidências de desafios que precisam ser tratados para uma análise confiável. A principal dificuldade em analisar o jargão médico utilizado no PEP, referiu-se aos seus aspectos epistemológicos que dependem fortemente do contexto médico. A intenção desta análise terminológica não foi criar terminologias ou guidelines para as mesmas, mas sim entender suas características, possibilidades de extração e de análises.

Uma das principais contribuições da pesquisa, foi indicar formas de delimitar o algoritmo no domínio da Ginecologia e Obstetrícia, na língua

portuguesa, e a partir da extração de termos da Terminologia de Interface, possibilitar futuramente o enriquecimento de artefatos terminológicos como as Ontologias Biomédicas e Vocabulários Controlados.

REFERÊNCIAS

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**, Florianópolis, v. 8, n. 15, p. 18-40, 1. sem. 2003. DOI: https://doi.org/10.5007/1518-2924.2003v8n15p18. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18. Acesso em: 10 out 2019.

BLAKE, Catherine. Text mining. **Annual Review of Information Science and Technology**, [*s. l.*], v. 45, n. 1, p. 121-155, jan. 2011. DOI: https://doi-org.ez106.periodicos.capes.gov.br/10.1002/aris.2011.1440450110. Disponível em: https://asistdl-onlinelibrary-wiley.ez106.periodicos.capes.gov.br/doi/10.1002/aris.2011.1440450110. Acesso em: 3 out. 2023.

BLOBEL, Bernd. Interoperable EHR Systems: challenges, standards and solutions. **European Journal for Biomedical Informatics**, [*s. l.*], v. 14, n. 2, p. 10-19, 2018. DOI: https://doi.org/10.24105/ejbi.2018.14.2.3. Disponível em: https://www.ejbi.org/scholarly-articles/interoperable-ehr-systems--challenges-standards-and-solutions.pdf. Acesso em: 2 out. 2023.

BODENREIDER, Olivier. Lexical, terminological and ontological resources for biological text mining. *In*: ANANIDOU, Sophia; MCNAUGHT, John (ed.). **Text mining for biology and biomedicine**. London, UK: Artech House, 2006. Chapter 3, p. 43-66. Disponível em: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2006007.pdf. Acesso em: 6 out. 2023.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas. **Política nacional de atenção integral à saúde da mulher**: princípios e diretrizes. Brasília: Ministério da Saúde, 2004. 82 p. (Série C. Projetos, Programas e Relatórios). Disponível em : http://bvsms.saude.gov.br/bvs/publicacoes/politica_nac_atencao_mulher.pdf. Acesso em: 8 jan. 2020.

BRASIL. Ministério da Saúde; INSTITUTO SÍRIO-LIBANÊS DE ENSINO E PESQUISA. **Protocolos da Atenção Básica**: saúde das mulheres. Brasília: Ministério da Saúde; Instituto Sírio-Libanês de Ensino e Pesquisa, 2016. 230 p. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/protocolos_atencao_basica_saude_mulheres.pdf. Acesso em: 23 set 2023.

CAMPOS, Maria Luiza de Almeida. **Linguagem documentária**: teorias que fundamentam sua elaboração. Niterói, RJ: EdUFF, 2001. 133p.

COMISSÃO NACIONAL DE INCORPORAÇÃO DE TECNOLOGIAS (Brasil). **Diretrizes Nacionais de Assistência ao Parto Normal**. Brasília: Ministério da Saúde, maio 2016. 399 p. (Relatório de recomendação, nº 211).

CONSELHO REGIONAL DE MEDICINA DO DISTRITO FEDERAL. **Prontuário médico do paciente**: guia para uso prático. Brasília: CRM-DF, 2006. Disponível em: https://crmdf.org.br/wp-content/uploads/2021/05/prontuario-medico-do-paciente-1.pdf. Acesso em: 2 out. 2023.

DALIANIS, Hercules. Characteristics of patient records and clinical corpora. *In*: DALIANIS, Hercules. **Clinical text mining**: secondary use of electronic patient records. [*S. l.*]: Springer Cham, 2018b. Chapter 4, p. 21-34. DOI: https://doi.org/10.1007/978-3-319-78503-5_4. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-78503-5_4. Acesso em: 2 jan. 2019.

DALIANIS, Hercules. Medical classifications and terminologies. *In*: DALIANIS, Hercules. **Clinical text mining**: secondary use of electronic patient records. [*S. l.*]: Springer Cham, 2018a. Chapter 5, p. 35-43. DOI: https://doi.org/10.1007/978-3-319-78503-5_5. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-78503-5_5. Acesso em: 2 jan. 2019.

FALCÃO JÚNIOR, João Oscar Almeida *et al*. **Ginecologia e obstetrícia**: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017.

KAFKAS, Senay; TOONSI, Sumyyah; ALSAEDI, Sakhaa. T1: Natural Language Processing Tutorial for Biomedical Text Mining (Half-day). *In*: INTERNATIONAL CONFERENCE ON BIOMEDICAL ONTOLOGY, 14.; SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL JOINT CONFERENCE, 16., 2023, Brasília. [Tutorial]. Brasília, DF: Faculdade de Ciência da Informação, UnB,

2023. Disponível em: https://github.com/stoonsi/ICBO-NLP-for-Biomedical--Text-Mining-tutorial/tree/main. Acesso em: 19 set. 2023.

KIM, Yoon Seob *et al.* Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. **PLoS One**, San Francisco, v. 12, n. 8, e0182889, Aug. 2017. DOI: https://doi.org/10.1371/journal.pone.0182889. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182889. Acesso em: 5 out. 2023.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge, Massachusetts: MIT Press, 1999. 620 p.

MATOS, Margarida Santos *et al.* **Manual de ginecologia**. Salvador: EBMSP, 2017.

MEYSTRE, Stephane M. *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. **BMC Medical Research Methodology**, [London], v. 10, article number 70, 2010.

MINAS GERAIS. Secretaria de Estado de Saúde; ASSOCIAÇÃO DE GINECOLOGISTAS E OBSTETRAS DE MINAS GERAIS. **Atenção à saúde da gestante**: novos critérios para estratificação de risco e acompanhamento da gestante: Programa Viva Vida: Projeto Mães de Minas. Belo Horizonte: SES-MG, maio 2013. [Nota Técnica Conjunta]. Disponível em: https://www.conass.org.br/liacc/wp-content/uploads/2015/02/Oficina--3-Estratificacao-de-Risco-GESTANTE.pdf. Acesso em: 5 out. 2023.

MIÑARRO-GIMÉNEZ, Jose A. *et al*. Quantitative analysis of manual annotation of clinical text samples. **International Journal of Medical Informatics**, [*s. l.*], v. 123, p. 37-48, Mar. 2019. DOI: https://doi.org/10.1016/j.ijmedinf.2018.12.011. Disponível em: https://www.sciencedirect.com/science/article/pii/S1386505618305446. Acesso em: 2 out. 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **CID-10**: Classificação Estatística Internacional de Doenças. v. 1. [*S. l.*]: Edusp, 1994. Disponível em: https://www.medicinanet.com.br/cid10.htm. Acesso em: 19 set. 2023.

PEIXOTO, Sérgio. **Manual de assistência pré-natal**. 2. ed. São Paulo: FEBRASGO, 2014. Disponível em: https://www.abenforj.com.br/site/arquivos/manuais/304_Manual_Pre_natal_25SET.pdf. Acesso em: 5 out. 2023.

ROGERS, Jeremy. **Using medical terminologies**. 2005. Disponível em: http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/termi nologyusing. Htm. Acesso em: 5 mar. 2019.

SCHULZ, Stefan *et al*. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. *In*: GUNDLAPALLI, Adi V.; JAULENT, Marie-Christine (ed.). **MEDINFO 2017**: precision healthcare through informatics. Proceeding of the 16th World Congress on Medical and Health Informatics. Amsterdam: IOS Press; International Medical Informatics Associations, c2017. p. 940-944. (Studies in Health Technology and Informatics, v. 245). DOI: https://doi.org/10.3233/978-1-61499-830-3-940. Disponível em: https://ebooks.iospress.nl/publication/48291. Acesso em: 2 out. 2023.

SHORTLIFFE, Edward H. Biomedical informatics: the science and the pragmatics. *In*: SHORTLIFFE, Edward H.; CIMINO, James J. (ed.). **Biomedical informatics**: computer applications in health care and biomedicine. 4th ed. London: Springer-Verlag, 2014. Cap. 1, p. 3-37. DOI: https://doi.org/10.1007/978-1-4471-4474-8_1.

SODERLAND, Stephen *et al*. **Machine learning of text analysis rules for clinical records**. [*S. l.: s. n.*], 1995. Disponível em : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.1340. Acesso em: 5 mar. 2019.

SOUZA, Amanda Damasceno de. Exigências éticas da pesquisa. *In*: CASTELLANO, Elisabete Gabriela; ASSIS, Orly Zucatto Mantovani de (org.). **Metodologia do trabalho e da pesquisa científica**. São Carlos: Diagrama, 2022. p. 487-507.

SOUZA, Amanda Damasceno de *et al.* O bibliotecário e a pesquisa terminológica em prontuários na área de saúde. *In*: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 29., 26 a 30 de setembro de 2022, [evento *online*]. **Anais** [...]. São Paulo: FEBAB, 2022. v. 1. n. 1. [Eixo 4 - Ciência da Informação: diálogos e conexões].

Disponível em: https://portal.febab.org.br/cbbd2022/article/view/2550. Acesso em: 23 set. 2023.

SOUZA, Amanda Damasceno de. **O discurso na prática clínica e as terminologias de padronização**: investigando a conexão. 2021. Tese (Doutorado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: http://hdl.handle.net/1843/38044. Acesso em: 23 set. 2023.

SOUZA, José Helvécio Kalil de; TEIXEIRA, Ivana Vilela. Anamnese e exame físico em ginecologia: propedêutica em ginecologia: aspectos atuais. *In*: FALCÃO JÚNIOR, João Oscar de Almeida *et al*. **Ginecologia e obstetrícia**: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017. cap. 18, p. 249-256.

WANG, Zhuoran *et al*. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. **PLoS One**, San Francisco, v. 7, n. 1, e30412, Jan. 2012. DOI: https://doi.org/10.1371/journal.pone.0030412. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030412. Acesso em: 5 out. 2023.

WILLIS, Sharon R. NLM Classification 2019 Summer Edition Now Available. Posted: 2019 Oct. 1. **NLM Technical Bulletin**, Bethesda, n. 430, e4, Sep-Oct 2019. Disponível em: https://www.nlm.nih.gov/pubs/techbull/so19/so19_nlm_classification_summer_2019.html. Acesso em: 21 jul. 2020.

ZHOU, Xiaohua *et al*. Approaches to text mining for clinical medical records. *In*: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 21st, 2006, Dijon, France, April 23-27. **Proceedings** [...]. New York: ACM, April, 2006. p. 235-239. DOI: https://doi.org/10.1145/1141277.1141330. Disponível em: http://www.ischool.drexel.edu/faculty/hhan/SAC2006_CAHC.pdf. Acesso em: 20 jun. 2019.

ANEXO A – EXEMPLO DE CLASSIFICAÇÃO DAS DOENÇAS GINECOLÓGICAS E DE SINAIS E SINTOMA DA CID-10 EM PORTUGUÊS

Fração da Tabela R – Sinais e Sintomas
R00 - Anormalidades do Batimento Cardíaco
R01 - Sopros e Outros Ruídos Cardíacos
R02 - Gangrena Não Classificada em Outra Parte
R03 - Valor Anormal da Pressão Arterial Sem Diagnóstico
R04 - Hemorragia Das Vias Respiratórias
R05 - Tosse
R06 - Anormalidades da Respiração
R07 - Dor de Garganta e no Peito
R09 - Outros Sintomas e Sinais Relativos Aos Aparelhos Circulatório e Respiratório
R10 - Dor Abdominal e Pélvica
R11 - Náusea e Vômitos
R12 - Pirose
R13 - Disfagia
R14 - Flatulência e Afecções Correlatas
R15 - Incontinência Fecal
R16 - Hepatomegalia e Esplenomegalia Não Classificadas em Outra Parte

Fração da Tabela R – Sinais e Sintomas
R17 - Icterícia Não Especificada
R18 - Ascite
R19 - Outros Sintomas e Sinais Relativos ao Aparelho Digestivo e ao Abdome
R20 - Distúrbios da Sensibilidade Cutânea
R21 - Eritema e Outras Erupções Cutâneas Não Especificadas
R22 - Tumefação, Massa ou Tumoração Localizadas da Pele e do Tecido Subcutâneo
R23 - Outras Alterações Cutâneas
R25 - Movimentos Involuntários Anormais
R26 - Anormalidades da Marcha e da Mobilidade
R27 - Outros Distúrbios da Coordenação
R29 - Outros Sintomas e Sinais Relativos Aos Sistemas Nervoso e Osteomuscular
R30 - Dor Associada à Micção
R31 - Hematúria Não Especificada
R32 - Incontinência Urinária Não Especificada
R33 - Retenção Urinária
R34 - Anúria e Oligúria

Fração da Tabela R – Sinais e Sintomas
R35 - Poliúria
R36 - Secreção Uretral
R39 - Outros Sintomas e Sinais Relativos ao Aparelho Urinário
R40 - Sonolência, Estupor e Coma
R41 - Outros Sintomas e Sinais Relativos à Função Cognitiva e à Consciência
R42 - Tontura e Instabilidade
R43 - Distúrbios do Olfato e do Paladar
R44 - Outros Sintomas e Sinais Relativos às Sensações e às Percepções Gerais
R45 - Sintomas e Sinais Relativos ao Estado Emocional
R46 - Sintomas e Sinais Relativos à Aparência e ao Comportamento
R47 - Distúrbios da Fala Não Classificados em Outra Parte
R48 - Dislexia e Outras Disfunções Simbólicas, Não Classificadas em Outra Parte
R49 - Distúrbios da Voz
[...]

Fonte: Tabela CID-10 de sinais e sintomas[74].

74 Disponível em: https://www.medicinanet.com.br/cid10/r.htm. Acesso em: 15 fev. 2020.

DADOS DOS AUTORES:

Amanda Damasceno de Souza

Amanda Damasceno de Souza é doutora em Gestão e Organização do Conhecimento e mestre em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG) e graduada em Biblioteconomia também pela UFMG.Atuou como Bibliotecária Clínica no Hospital Felício Rocho e na Oncologia do Hospital Belo Horizonte e da Santa Casa de Belo Horizonte. Atualmente é Coordenadora do Comitê de Ética em Pesquisa da Universidade FUMEC e é membro dos Grupos de Pesquisa Núcleo de Estudos e Pesquisas sobre Recursos, Serviços e Práxis Informacionais (NERSI), do grupo Representação do Conhecimento, Ontologias e Linguagem (ReCOL), do NCOR-BR, do Comitê ABNT CE 021:002.032 e da Comissão de Pesquisa e Iniciação Científica (CoPIC) da Universidade FUMEC. Atua como docente no Bacharelado em Estética, Bacharelado em Administração e no Programa de Pós-Graduação em Tecnologia da Informação Comunicação e Gestão do Conhecimento (PPGTICGC) da Universidade FUMEC e editora das Revistas Estética em Movimento e Código-31.

https://orcid.org/0000-0001-6859-4333
amanda.dsouza@fumec.br

Eduardo Ribeiro Felipe

Eduardo Felipe é professor Adjunto do curso de Engenharia de Computação na Universidade Federal de Itajubá. Doutor em Gestão e Organização do Conhecimento pela UFMG. Mestre em Ciência da Informação pela UFMG e Pós-graduado em Engenharia de Software pela PUC-Minas. Possui graduação como Tecnólogo em Processamento de Dados pelo Centro Universitário Newton Paiva. Membro do grupo de pesquisa Representação do Conhecimento, Ontologias e Linguagem (ReCOL). Membro do Grupo de pesquisa; Laboratório de Robótica, Sistemas Inteligentes e Complexos - RobSIC. Membro do Conselho Universitário CONSUNI, Membro do Núcleo Docente Estruturante NDE (Computação). Atua nas áreas de Linguagens de Programação, Desenvolvimento Web e Mobile, Recuperação da Informação e Ontologias.

https://orcid.org/0000-0003-1690-2044
eduardo.felipe@unifei.edu.br

Fernanda Farinelli

Fernanda Farinelli é Professora Adjunta na Faculdade de Ciência da Informação da UnB. Doutora em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação da UFMG pesquisando o tema ontologias formais realistas como solução de integração semântica de dados. Ontologista responsável pelo projeto da OntONeo (Ontologia do domínio obstétrico e neonatal). Pesquisadora visitante no Departamento de Filosofia e no Departamento de Informática Biomédica da Universidade Estadual de Nova York em Buffalo entre 05/2015 e 04/2017. Mestre em Administração de Empresas com ênfase em Gestão estratégica da informação (Fundação Pedro Leopoldo/MG). Especialista em Banco de Dados (UNI-BH). Bacharel em Ciência da Computação (PUC-MG). Possui mais de 15 anos de experiência em Gestão de Dados atuando com administração de banco de dados, arquitetura e administração de dados e implantação de governança de dados em grandes empresas como Unisys Brasil, Cedro Têxtil, Prodemge. Atua há cerca de 15 anos como docente em cursos de graduação e pós-graduação em renomadas instituições de ensino no estado de Minas Gerais como PUC-MG, IEC, Fundação Pedro Leopoldo, Universidade de Itaúna, Faculdade Cotemig, Unipac e IGTI. Possui as certificações CDMP, CBIP, CDP e OCP.

https://orcid.org/0000-0003-2338-8872
fernanda.farinelli@unb.br

Como referenciar o capítulo 5:

SOUZA, Amanda Damasceno de; FELIPE, Eduardo Ribeiro; FARINELLI, Fernanda. Extração e análise de dados registrados em texto livre de prontuário eletrônico do paciente por meio de processamento de linguagem natural. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 5. p. 103-138. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap5.