Análise e Curadoria de Dados Textuais para o Projeto CardioIA: Uma Abordagem Profissional em Processamento de Linguagem Natural

1. A Fundamentação do Problema de Dados em Saúde: Do Texto Livre ao Conhecimento Estruturado

A entrada de um paciente em um ambiente de saúde, seja para uma triagem inicial ou um acompanhamento de longo prazo, gera uma vasta quantidade de informações. No entanto, o desafio central para a aplicação de sistemas de inteligência artificial reside na natureza desses dados: embora existam campos estruturados para informações como idade e sexo, grande parte do conhecimento crítico está contida em dados narrativos não estruturados. Prontuários eletrônicos, relatórios médicos, notas de evolução de enfermagem e até mesmo publicações científicas são majoritariamente compostos por texto livre. A falta de padronização, o uso de acrônimos, a evolução da terminologia ao longo do tempo (como o termo de medição cardíaca que mudou de QRS para QRSd em poucos anos) e a presença de erros gramaticais e de digitação tornam a análise desses registros uma tarefa complexa e, em muitos casos, inviável em sua forma bruta.[1]

A mineração desses dados não estruturados é uma etapa fundamental para qualquer projeto de IA na saúde. Sem um método eficaz para extrair e organizar o conhecimento contido nessas narrativas, as informações permanecem inacessíveis para análises quantitativas e para o treinamento de modelos preditivos. O sucesso de uma pesquisa ou de uma plataforma inteligente, como o CardioIA, depende diretamente da capacidade de superar essa barreira, transformando a "memória escrita da história das condições de saúde" de uma pessoa em um formato que a tecnologia possa processar e aprender.[1]

O processamento de linguagem natural (NLP) emerge como a solução estratégica para este problema. Essa disciplina da inteligência artificial atua como a chave para decifrar a riqueza de informações escondida em textos clínicos e literários. Ao aplicar técnicas de NLP, é

possível "ler" e interpretar as narrativas, extraindo informações relevantes e transformando o texto livre em unidades de dados discretas e analisáveis. Esta transformação é a base para o desenvolvimento de modelos de aprendizado de máquina, pois converte dados qualitativos em uma forma quantitativa e padronizada. O caso da Mercy em St. Louis ilustra essa abordagem, onde o NLP é utilizado para analisar notas de texto livre de cardiologistas, permitindo que a equipe obtenha uma imagem detalhada da progressão da insuficiência cardíaca e de como os sintomas evoluem ao longo do tempo.[2] A capacidade de identificar padrões em campos narrativos que seriam difíceis de discernir para um profissional humano ou para um sistema que depende apenas de dados estruturados, possibilita a geração de

insights clínicos valiosos, melhorando a precisão do diagnóstico e a eficácia do tratamento.[4]

2. Estratégia de Pesquisa e Curadoria de Dados Textuais: Fontes, Coleta e Governança

A coleta de dados textuais para o projeto CardioIA deve ser deliberada e estratégica, buscando não apenas preencher a cota de dois textos, mas também estabelecer uma base de conhecimento diversificada e relevante para as fases futuras do projeto. A escolha das fontes reflete uma abordagem profissional que entende a importância de dados de alta qualidade e com diferentes propósitos clínicos e analíticos.

2.1. Fontes de Dados Sugeridas e Justificativa de Escolha

Para a atividade, foram selecionadas fontes que oferecem um espectro completo de informações essenciais para um ecossistema de cardiologia inteligente: desde o conhecimento científico e as diretrizes de saúde pública até a perspectiva humana e histórica da medicina.

- **Fontes Científicas e Oficiais (Acesso Aberto):**
    - **SciELO (Scientific Electronic Library Online):** Esta biblioteca digital é uma fonte inestimável de artigos revisados por pares, garantindo a validade e a relevância clínica do conteúdo. Publicações como os Arquivos Brasileiros de Cardiologia e a Revista Portuguesa de Cardiologia oferecem dados factuais, análises epidemiológicas e estudos de caso que são a base para os módulos de previsão médica e diagnóstico do CardioIA.[5] A pesquisa de artigos sobre temas específicos, como o impacto do COVID-19 em doenças cardiovasculares ou estudos sobre

arritmia cardíaca, que descrevem sintomas como palpitações, fadiga e dispneia, fornece o conhecimento fundamental para a extração de entidades e a construção de modelos preditivos.[7]
    - **BVS (Biblioteca Virtual em Saúde) e o Site Oficial do SUS:** Essas plataformas são essenciais para contextualizar o projeto dentro da realidade do sistema de saúde pública brasileiro. A Estratégia de Saúde Cardiovascular (ECV) do Ministério da Saúde, por exemplo, detalha os principais fatores de risco para doenças cardiovasculares, como hipertensão arterial, diabetes e tabagismo, e define objetivos estratégicos para prevenção e controle.[10] Esses textos fornecem o arcabouço para entender a governança de dados e os objetivos de saúde a nível sistêmico, o que é crucial para garantir que o projeto seja ético, responsável e clinicamente relevante.
- **Fontes de Literatura Clássica (Domínio Público):**
    - **Projeto Gutenberg:** A inclusão de obras de domínio público, como o romance "St. Bernard's: The Romance of a Medical Student", amplia a visão do que constitui "dados" relevantes para a saúde.[11] Embora não seja um texto clínico, a narrativa oferece uma perspectiva histórica e humana sobre a medicina. Em um projeto que visa simular um ecossistema de cardiologia moderno, que vai além do diagnóstico físico para abordar a experiência holística do paciente, a análise dessas obras é valiosa. O processamento de linguagem natural pode ser aplicado a esses textos para análises de sentimento e temáticas, revelando a dimensão emocional e psicossocial da saúde, que é um fator de risco e um preditor de sucesso no tratamento. Esta abordagem integra a visão de um sistema que se preocupa com o bem-estar do paciente em sua totalidade, alinhando-se com a filosofia da medicina integral.

A tabela a seguir resume as fontes sugeridas e sua relevância estratégica para o projeto, demonstrando a complementaridade entre os tipos de dados e o pensamento crítico na sua seleção.

Fonte	URL de Exemplo	Tipo de Conteúdo	Relevância para o Projeto CardioIA
SciELO	https://www.scielo.br/j/abc/ [6]	Artigos e Pesquisas Científicas	Dados factuais e estatísticos para previsões médicas e a base de conhecimento para o diagnóstico.
BVS e Site do SUS	https://www.gov.br/	Relatórios de	Contexto da saúde

	saude/.../ecv [10]	Saúde Pública	pública brasileira, crucial para a governança de dados e definição de objetivos.
Projeto Gutenberg	http://www.gutenberg.org/ebooks/46431 [11]	Literatura Clássica em Saúde	Narrativas e descrições humanas, essenciais para análise de sentimento e assistência remota.

2.2. Diretrizes de Coleta e Organização Técnica

A organização dos dados é tão importante quanto a sua coleta. Para garantir que os dados textuais possam ser utilizados de forma eficaz nas fases subsequentes do projeto, é essencial seguir as melhores práticas de gerenciamento de repositórios. Os arquivos de texto (.txt) devem ser armazenados em uma subpasta dedicada no repositório do GitHub (e.g., assets ou docs), permitindo uma organização clara e modular. No arquivo principal README.md, deve-se incluir uma descrição detalhada de cada parte da atividade, os objetivos do projeto e, mais importante, a justificativa para a escolha das fontes e a relevância das análises de NLP. Esta documentação é o mapa que orientará os futuros colaboradores do projeto e demonstra o rigor técnico e a visão de longo prazo do trabalho.

3. Aplicações de NLP em Saúde Cardiovascular: Teoria e Casos de Uso

A aplicação de técnicas de NLP aos textos coletados não é uma mera formalidade, mas uma etapa crítica para transformar informação bruta em conhecimento acionável. Diferentes técnicas de NLP permitem extrair diferentes tipos de valor dos dados textuais, cada uma com uma aplicação específica no ecossistema do CardioIA.

3.1. Extração de Entidades Clínicas (NER): O Núcleo do Diagnóstico e da Triagem

O Reconhecimento de Entidades Nomeadas (NER) é uma técnica fundamental que permite identificar e classificar entidades específicas em um texto, como nomes de pessoas, locais ou organizações. No contexto médico, o NER é adaptado para extrair entidades clínicas, tais como sintomas, diagnósticos, medicamentos, condições e exames.[1] A aplicação do NER transforma as descrições narrativas de um paciente — como as presentes em prontuários ou em artigos científicos sobre arritmias (

palpitações no coração, hipotensão, fadiga, dispneia) — em dados estruturados que podem ser processados por um algoritmo.[7]

A dissertação de mestrado da UNESP serve como um exemplo prático e inspirador dessa aplicação. O estudo desenvolveu uma rede neural utilizando a biblioteca spaCy em Python para extrair informações de 30.000 prontuários de pacientes, demonstrando a viabilidade de transformar dados de campos narrativos em informações estruturadas. Os resultados obtidos foram significativos, com uma precisão de 72,7% para o modelo geral e um impressionante 90,3% para a classe Condição do Paciente. A capacidade de extrair e organizar informações como medicamentos prescritos, por exemplo, permite análises de agrupamento para traçar perfis de pacientes e entender seus comportamentos, revelando informações até então desconhecidas.[1]

A extração dessas entidades é diretamente aplicável aos módulos do CardioIA:

- **Triagem:** Um algoritmo de NER pode processar a descrição inicial de um paciente ("Sinto tontura e falta de ar") e extrair tontura e falta de ar como sintomas, o que auxilia na priorização e no direcionamento do atendimento.
- **Diagnóstico e Monitoramento:** A análise de textos clínicos pode extrair diagnósticos confirmados ou a lista de medicamentos prescritos, permitindo o rastreamento da evolução da doença. O estudo da Mercy em St. Louis utilizou o NLP com sucesso para rastrear a progressão de insuficiência cardíaca e a mudança de sintomas ao longo do tempo a partir de notas de texto livre, o que resultou em melhorias no tratamento e no design de dispositivos cardíacos.[2]

3.2. Análise de Sentimento: A Dimensão Emocional e Psicossocial da Saúde

A análise de sentimento é uma técnica de mineração de texto que determina a polaridade emocional ou a opinião expressa em um texto (positiva, negativa, neutra).[12] Embora o material de pesquisa disponível aplique essa técnica ao contexto da saúde mental, identificando sentimentos como

tristeza, solidão e medo, sua relevância para o projeto CardioIA transcende essa área.[13] A medicina moderna reconhece que a saúde cardiovascular está intrinsecamente ligada à saúde mental e ao bem-estar do paciente.

A aplicação da análise de sentimento em textos narrativos do paciente, como aqueles que seriam gerados por um sistema de assistência remota (e.g., um diário eletrônico do paciente), pode fornecer informações críticas sobre seu estado psicológico. Por exemplo, a detecção de sentimentos de desespero ou frustração pode indicar baixa adesão ao tratamento, depressão ou outros fatores que impactam negativamente a recuperação. A inclusão de fontes de literatura clássica, como a proposta na seção anterior, também permite a exploração dessa dimensão humana, preparando o terreno para um sistema que não trata apenas a doença, mas o paciente em sua totalidade.

3.3. Classificação e Sumarização de Tópicos: Organizando o Conhecimento em Escala

Com a vasta quantidade de informações disponíveis em bases de dados como SciELO e BVS, a classificação e a sumarização de tópicos são essenciais. A **classificação de tópicos** categoriza documentos em temas predefinidos (ex: hipertensão, arritmia, tratamento), enquanto a **sumarização** cria resumos concisos e informativos.

Para um sistema como o CardioIA, que depende de informações atualizadas para os módulos de previsões médicas e diagnóstico, a capacidade de processar rapidamente novos artigos científicos, como a seleção anual dos "melhores artigos em cardiologia", é uma vantagem competitiva e clínica. Algoritmos de NLP podem processar esses textos, classificando-os por tema (e.g., fatores de risco dietéticos ou taxas de mortalidade) e gerando resumos, agilizando o acesso do sistema e dos profissionais às descobertas mais recentes.[5]

4. O Valor Estratégico das Análises de NLP para o

Projeto CardioIA

A verdadeira relevância do processamento de linguagem natural no projeto CardioIA reside na sua capacidade de conectar dados dispersos e transformá-los na espinha dorsal dos módulos inteligentes. A tabela abaixo sintetiza como cada técnica de NLP se alinha diretamente com os objetivos do projeto, criando um ecossistema integrado e funcional.

Técnica de NLP	Caso de Uso no Projeto CardioIA	Relevância Clínica	Referência nos Dados de Pesquisa
Reconhecimento de Entidades Nomeadas	Extração de sintomas para triagem, rastreamento de medicamentos para monitoramento	Agiliza o diagnóstico inicial, permite o rastreamento da evolução da doença, facilita a pesquisa.	[1], [1]
Análise de Sentimento	Avaliação do bem-estar emocional do paciente para assistência remota	Promove uma abordagem de saúde holística, detecta fatores de risco psicossociais.	[13], [13]
Classificação e Sumarização de Tópicos	Categorização de literatura científica para previsões médicas e diagnóstico	Mantém o sistema atualizado com as últimas descobertas, otimiza o acesso ao conhecimento.	[5], [5]

4.1. O Papel do NLP na Governança de Dados e Mitigação de Vieses

A importância do NLP vai além da simples extração de informações. A disciplina também desempenha um papel fundamental na governança de dados, um conceito central na atividade. A governança de dados garante a qualidade, a segurança e a usabilidade das

informações. Em ambientes de saúde, onde a variabilidade e a subjetividade são comuns, o NLP age como um mecanismo de padronização.

Ao estruturar dados que antes existiam apenas em campos narrativos, o NLP força a consistência na extração de informações. Isso é vital para mitigar vieses que podem surgir de diferentes estilos de anotação de profissionais ou de mudanças na terminologia ao longo do tempo, como observado no estudo da Mercy em St. Louis.[2] Um modelo de NER treinado para reconhecer

hipertensão ou diabetes em qualquer contexto textual, independentemente do estilo do autor, assegura que esses dados sejam consistentemente representados, tornando-os mais transparentes e comparáveis. Essa padronização é o alicerce para uma boa governança, permitindo a auditoria de registros e a identificação de anomalias ou vieses nos padrões de registro dos profissionais de saúde, contribuindo para a construção de um sistema de IA mais justo e equitativo.

5. Conclusão e Recomendações

A pesquisa e a curadoria dos dados textuais para o projeto CardioIA constituem a base sólida sobre a qual todo o ecossistema de cardiologia inteligente será construído. A tarefa vai muito além da simples coleta de arquivos; ela exige uma compreensão profunda da natureza dos dados de saúde e do papel transformador do Processamento de Linguagem Natural. O sucesso do projeto depende da capacidade de converter o conhecimento contido em narrativas não estruturadas — sejam elas de artigos científicos, relatórios oficiais ou literatura clássica — em dados limpos e utilizáveis.

A presente análise demonstra que cada fonte de dados e cada técnica de NLP tem um propósito estratégico específico. A extração de entidades clínicas, a análise de sentimento e a classificação de tópicos não são apenas ferramentas isoladas, mas componentes interligados que, juntos, capacitam os módulos de triagem, diagnóstico, monitoramento, assistência remota e previsões médicas do CardioIA. A coleta de textos de diferentes naturezas (científica e literária) é um reflexo do compromisso do projeto com uma abordagem holística, que considera tanto os dados clínicos quanto a dimensão humana da experiência de saúde.

Para as fases subsequentes, recomenda-se que o trabalho de curadoria de dados textuais seja um processo contínuo. A documentação detalhada no arquivo README.md é crucial para o sucesso da colaboração em grupo. Além disso, a validação do modelo de NLP é um passo fundamental. Como destacado em estudos de campo, é importante que o algoritmo seja treinado e refinado constantemente com novas informações para que seja continuamente

aprimorado, além de retroalimentado com a perspectiva de outros profissionais da saúde para elevar sua qualidade e reduzir a parcialidade dos resultados.[1] A aplicação de análises estatísticas multivariadas nos dados extraídos, como a de agrupamento, pode revelar novos perfis de pacientes e comportamentos, reforçando a importância da extração de informações de texto livre para avanços na saúde.[1]

Works cited

1. Uso de Redes Neurais Artificiais para Extração de Dados de ..., accessed August 27, 2025, https://repositorio.unesp.br/bitstreams/2b67e26f-040f-4f28-b4da-18d246959eb9/download
2. Tracking the Progression of Heart Failure Using NLP | Epic, accessed August 27, 2025, https://www.epic.com/epic/post/tracking-progression-heart-failure-using-nlp/
3. Natural Language Processing for Cardiovascular Applications | Request PDF, accessed August 27, 2025, https://www.researchgate.net/publication/360109084_Natural_Language_Processing_for_Cardiovascular_Applications
4. INTELIGÊNCIA ARTIFICIAL NA MEDICINA DIAGNÓSTICA - Brazilian Journal of Implantology and Health Sciences, accessed August 27, 2025, https://bjihs.emnuvens.com.br/bjihs/article/download/4235/4316/9407
5. Os Melhores Artigos de 2022 nos Arquivos Brasileiros de Cardiologia e na Revista Portuguesa de Cardiologia - SciELO, accessed August 27, 2025, https://www.scielo.br/j/abc/a/nRRHpYYCBXmYzTDTJhT3wKQ/?lang=pt
6. SciELO Arquivos Brasileiros de Cardiologia, accessed August 27, 2025, https://www.scielo.br/j/abc/
7. Arritmias supraventriculares: Uma revisão de literatura - Portal de Revistas da UNIDEP, accessed August 27, 2025, https://periodicos.unidep.edu.br/sante/article/view/148
8. Artigo – Arritmia Cardíaca - ICTDF, accessed August 27, 2025, https://ictdf.org.br/artigo-arritmia-cardiaca/
9. APA - SciELO Preprints, accessed August 27, 2025, https://preprints.scielo.org/index.php/scielo/citationstylelanguage/get/apa?submissionId=627&publicationId=649
10. Estratégia de Saúde Cardiovascular — Ministério da Saúde, accessed August 27, 2025, https://www.gov.br/saude/pt-br/composicao/saps/ecv
11. St. Bernard's: The Romance of a Medical Student by Edward Berdoe ..., accessed August 27, 2025, http://www.gutenberg.org/ebooks/46431
12. a aplicação da técnica de análise de sentimento em mídias sociais como instrumento para as práticas da gestão social em nível governamental - SciELO, accessed August 27, 2025, https://www.scielo.br/j/rap/a/GD3F8HdkQKGSHy8zzV8w9Ys/
13. Uma avaliação da capacidade de Modelos de Linguagem para ..., accessed August 27, 2025, https://sol.sbc.org.br/index.php/sbcas/article/view/35504