SaudeBR-QA: Um Córpus de Perguntas e Respostas para o Domínio da Saúde em Português Brasileiro

Carlos Henrique S. Barros, Gustavo F. Rodrigues de Sousa, Rogério F. de Sousa

[1] Laboratório de Inteligência Artificial, Robótica e Automação
Instituto Federal de Educação, Ciência e Tecnologia do Piauí – Picos, PI – Brasil

{carlosh3123, gustavofigueredo86}@gmail.com, rogerio.sousa@ifpi.edu.br

Abstract. *This paper presents SaudeBR-QA, a large-scale corpus composed of 23,382 question–answer pairs from the Brazilian Portuguese healthcare domain. The dataset was collected from a public health portal and contains real patient questions and verified answers from healthcare professionals. Each pair includes rich metadata such as medical specialty, quality metrics (scores from 0 to 5), and user likes. The corpus addresses the lack of domain-specific NLP resources for Brazilian Portuguese. SaudeBR-QA is expected to support research in text classification, automatic answer quality prediction, and other health-related text mining tasks.*

Resumo. *Este artigo apresenta o SaudeBR-QA, um córpus em larga escala com 23.382 pares de perguntas e respostas do domínio da saúde em português brasileiro. Os dados foram extraídos de um portal público, contendo perguntas reais de pacientes e respostas de profissionais verificados. Cada par possui metadados relevantes, incluindo especialidade médica, métricas de qualidade (0 a 5) e número de curtidas. O córpus oferece um recurso especializado para PLN em português, suprindo a escassez de bases anotadas para tarefas específicas. O SaudeBR-QA pretende apoiar pesquisas em classificação de textos, avaliação automática de respostas e outros cenários de mineração de textos em saúde.*

1. Introdução

O acesso a informações médicas na internet tem crescido de forma significativa, alterando a maneira como pacientes buscam esclarecimentos sobre sintomas, diagnósticos e possíveis tratamentos [OECD 2023]. Nesse cenário, conteúdos gerados por usuários (User Generated Content – UGC) tornaram-se fonte relevante de informação, especialmente em plataformas voltadas para Perguntas e Respostas (Health Question-Answering – HQA). Em ambientes desse tipo, há interação direta entre pacientes, comunidade e especialistas, ampliando a disseminação de conhecimento e a validação social das respostas [Verma et al. 2021].

Apesar dos avanços internacionais em Processamento de Linguagem Natural aplicado à saúde, ainda existe escassez de corpora modernos e de larga escala para o português brasileiro. Muitos recursos disponíveis restringem-se a textos formais, como artigos científicos ou bulas de medicamentos, e não contemplam interações espontâneas compostas por perguntas reais de pacientes. Nesse contexto, a construção de novos córpora abertos em português brasileiro é particularmente relevante. Recursos desse tipo podem subsidiar desde modelos clássicos de classificação até arquiteturas neurais de grande porte, permitindo investigar como diferentes abordagens lidam com linguagem informal, heterogeneidade

de vocabulário, descrições leigas de sintomas e diferentes especialidades médicas. Em particular, um córpus de Perguntas e Respostas em saúde pode ser empregado para treinar e avaliar modelos de linguagem de grande porte (LLMs) em português, ajustar sistemas de recuperação aumentada por geração (RAG), desenvolver métricas automáticas de qualidade de resposta e apoiar a criação de assistentes virtuais mais transparentes.

Para mitigar essa lacuna, este artigo apresenta o SaudeBR-QA: um córpus público [1], contemporâneo e composto por pares de perguntas e respostas provenientes de interações reais entre pacientes e profissionais de saúde. Além do conteúdo textual, o conjunto inclui avaliações numéricas, especialidades médicas e número de curtidas, permitindo sua aplicação em tarefas como classificação, predição de utilidade e geração de respostas explicativas. Os dados foram processados e anonimizados, garantindo conformidade ética e permitindo seu uso seguro pela comunidade científica.

2. Trabalhos Relacionados

Nesse contexto, a criação e análise de corpora de Perguntas e Respostas no domínio da saúde têm ganhado crescente atenção, impulsionadas pela importância desses dados para o desenvolvimento de sistemas capazes de compreender e gerar informações médicas de forma acessível e confiável. Diversos trabalhos exploram esse campo em diferentes idiomas, com distintos focos metodológicos e níveis de especialização. Trabalhos relacionados foram identificados a partir de uma busca bibliográfica em bases como Google Scholar[2] e PubMed[3], complementada pela análise de referências de artigos-chave na área. Foram utilizadas combinações de termos em inglês e português, tais como *"health question answering dataset"*, *"biomedical QA corpus"*, *"clinical corpus Portuguese"*.

Um dos principais nessa área é o MedQuAD (*Medical Question Answering Dataset*) [Ben Abacha and Demner-Fushman 2019], que reúne mais de 47 mil pares de perguntas e respostas provenientes de fontes oficiais de saúde dos Estados Unidos, como os Institutos Nacionais de Saúde (NIH). O conjunto é amplamente utilizado em *benchmarks* de PR biomédico, embora restrito à língua inglesa e baseado em perguntas padronizadas, não em interações espontâneas entre pacientes e profissionais.

Outro recurso relevante é o emrQA [Pampari et al. 2018], que contém perguntas e respostas derivadas de registros médicos eletrônicos, voltadas à avaliação de modelos de PR com foco em raciocínio clínico. Apesar de sua importância, o emrQA não é de acesso público devido à natureza sensível dos dados clínicos, o que limita sua replicabilidade fora do contexto hospitalar norte-americano.

Para o português, o HAREM [Santos and Cardoso 2006] foi um grande esforço de avaliação em Reconhecimento de Entidades Nomeadas (NER) para o português, incluindo algumas entidades da área biomédica. Mais recentemente, o córpus SemClinBr [Ferreira et al. 2023] constitui um recurso clínico nacional, composto por anotações linguísticas e terminológicas em notas médicas reais, anonimizadas e revisadas. Apesar disso, o português brasileiro ainda carece de corpora públicos de larga escala estruturados no formato de pergunta e resposta.

[1] https://github.com/CarlHenry670/SaudeBR-QA

[2] Google Scholar. Disponível em: `https://scholar.google.com/`

[3] PubMed. Disponível em: `https://pubmed.ncbi.nlm.nih.gov/`

Diante desse panorama, o SaudeBR-QA distingue-se não apenas por ser um córpus em português brasileiro, mas também pelo conjunto de características que o diferenciam. Trata-se de um recurso de larga escala, com mais de 23 mil pares de perguntas e respostas. O córpus incorpora um conjunto rico de metadados, incluindo mais de 30 especialidades médicas, o que amplia a possibilidade de estudos em diferentes áreas, além de avaliações numéricas atribuídas pelos usuários a cada par de pergunta e resposta. Em conjunto, essas características tornam o SaudeBR-QA particularmente adequado para tarefas de predição de qualidade, classificação de intenções e modelagem de sistemas de diálogo no contexto da saúde, configurando uma contribuição essencial para o avanço do PLN biomédico em língua portuguesa.

3. Metodologia

Esta seção descreve o processo de construção do *dataset* SaudeBR-QA, desde a seleção da fonte de dados a organização dos registros.

3.1. Seleção da Fonte

O portal selecionado para a coleta foi o CatálogoMed[4], uma plataforma pública que reúne perguntas enviadas por usuários e respostas elaboradas por profissionais de saúde verificados, devidamente organizadas por especialidade médica. Esse portal oferece contexto clínico explícito, pergunta–resposta em português do Brasil, e sinais de qualidade/engajamento úteis para tarefas de avaliação de qualidade de respostas. Além disso, o conteúdo está publicamente acessível, as páginas são paginadas e mantêm uma marcação HTML estável, o que viabiliza extração reprodutível. Essa combinação de sinal e estrutura torna o portal particularmente adequado ao objetivo de pesquisa do SaudeBR-QA, que é oferecer um córpus para o estudo e o desenvolvimento de aplicações de PLN no domínio da saúde em português.

3.2. Tecnologias e Ferramentas

A coleta dos dados foi realizada por meio de técnicas de *web scraping*, a partir da navegação automatizada na plataforma CatálogoMed. A implementação foi realizada em linguagem de programação *Python* [Python 2025], combinando *Selenium WebDriver* [Selenium 2025] para renderização/automação de páginas dinâmicas e *BeautifulSoup* [Richardson 2025] para extração e limpeza de elementos HTML. Para manipulação tabular e exportação do *dataset* foi utilizado o *pandas* [McKinney 2010].

3.3. Visão geral do coletor

O *scraping*[5] foi projetado para percorrer todo o portal, organizado por especialidade médica. A lista de especialidades disponível no site foi utilizada como ponto de partida, de modo que cada especialidade correspondesse a um bloco independente de navegação e extração. Algumas especialidades foram deliberadamente ignoradas por apresentarem quantidade extremamente reduzida de pares de pergunta e resposta, o que as tornava pouco informativas para compor o córpus final.

O processo foi implementado da seguinte maneira:

[4] https://www.catalogo.med.br/

[5] https://github.com/CarlHenry670/Crawler-QA

1. **Inicialização**: o *Main.py* configura o navegador (*Selenium/Chrome*) e carrega as dependências de *parsing* e escrita de dados.
2. **Leitura de controle**: são verificados os arquivos auxiliares. Carrega-se o *progresso.json* (para retomar de onde parou) e a lista de URLs com as especialidades (*especialidades.txt*). Se já existir o CSV final, ele é aberto em modo de acréscimo.
3. **Filtragem e verificação de status por especialidade**: antes de iniciar a coleta, o sistema checa se a especialidade já foi concluída em execução anterior. Se a especialidade estiver elegível e não concluída, inicia-se a coleta; caso contrário, passa-se para a próxima.
4. **Navegação paginada**: o coletor percorre as páginas da especialidade (página inicial e subsequentes), realizando pequenas pausas aleatórias para evitar sobrecarga no servidor.
5. **Renderização e extração**: cada página é carregada pelo Selenium e, em seguida, o HTML resultante é processado com *BeautifulSoup* para extrair perguntas, respostas e metadados associados.
6. **Organização dos registros**: as respostas são normalizadas (por exemplo, junção de parágrafos), deduplicadas quando necessário e armazenadas *uma por linha*, preservando o ID lógico da pergunta.
7. **Persistência incremental**: após cada página, os novos registros são adicionados ao CSV e o *progresso.json* é atualizado com a página atual e o status da especialidade, permitindo retomada segura em caso de interrupção.
8. **Condição de parada**: caso ocorram páginas sucessivas sem novos itens, a especialidade é marcada como concluída. Em cenários de indisponibilidade temporária do site, o sistema salva o que já foi coletado e finaliza, possibilitando a retomada posterior.

A etapa de coleta pode ser visualizada de maneira sucinta na Figura 1.

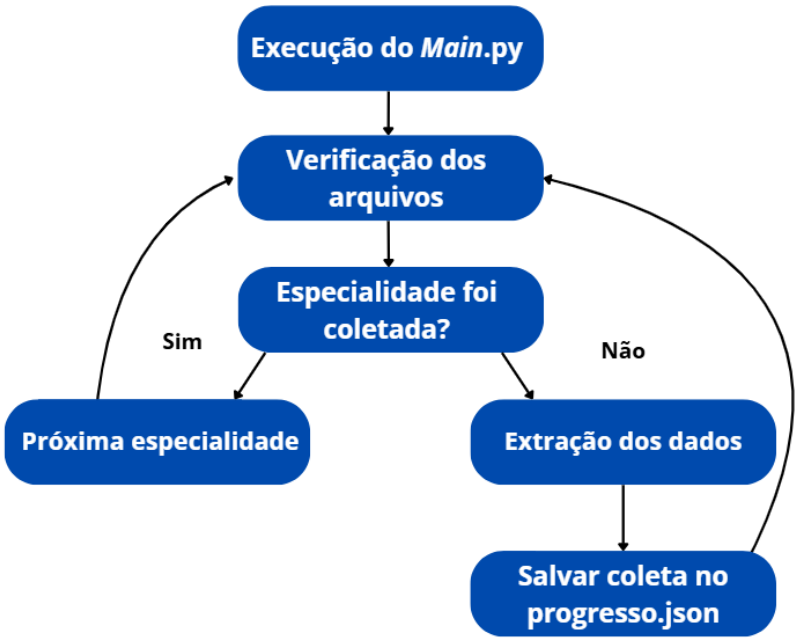

Figura 1. Fluxo de trabalho para extração dos dados.

4. Córpus e Análise da Estrutura

O córpus final SaudeBR-QA contém **23.382** pares de pergunta-resposta em português brasileiro. Foram removidos campos textuais com dados pessoais sensíveis, dados duplicados e sobreposições de registros. O recurso está estruturado como um único arquivo CSV para facilitar o uso, onde cada linha representa uma interação única.

4.1. Estrutura de Dados

O conjunto de dados apresenta seis colunas:

- **ID:** Um identificador exclusivo para a entrada.
- **Especialidade:** A especialidade médica (por exemplo, alergia+e+imunologia).
- **Pergunta:** A Pergunta do paciente.
- **Resposta:** A resposta do profissional de saúde.
- **Avaliação:** Uma classificação numérica de 0 a 5 atribuída pelos pacientes.
- **Curtidas:** O número de ”curtidas” que a resposta recebeu.

Um exemplo de entrada do córpus é mostrado na Tabela 1.

Tabela 1. Exemplos de entradas do córpus SaudeBR-QA (texto truncado).

ID	Especialidade	Pergunta (snippet)	Resposta (snippet)	Avaliação	Curtidas
2	alergia+e+imunologia	Olá, boa noite. Tenho urticária...	Pode ser urticária crônica...	4	1
1	alergia+e+imunologia	Minha filha tem uma alergia...	A chamada imunoterapia...	0	0

4.2. Análise Estatística

Tabela 2. Análise estatística do *SaudeBR-QA* (resumo geral, comprimentos e distribuição de avaliações).

Resumo do Córpus		**Comprimento Médio (caracteres)**	
Linhas (pares Q&A)	23.382	Perguntas	141,03
Colunas	6	Respostas	211,39
Especialidades únicas	72		
IDs únicos de perguntas	17.769		
Média de respostas por pergunta	1,32		
Máximo de respostas por pergunta	3		

Distribuição de *Avaliação* (0–5)

Classe	**Frequência**	**Percentual (%)**
0	21.628	92,50
1	115	0,49
2	50	0,21
3	117	0,50
4	209	0,89
5	1.263	5,40
Total	**23.382**	**100,00**
Média de *Avaliação* ($\bar{x}$)		0,33
Desvio-padrão (σ)		1,20
Mediana		0

A Tabela 2 apresenta um panorama quantitativo e descritivo do córpus SaudeBR-QA, consolidando informações estruturais e métricas básicas de distribuição. O conjunto contém 23.382 pares de pergunta e resposta, totalizando 17.769 identificadores únicos de perguntas e abrangendo 72 especialidades médicas distintas, o que evidencia sua diversidade temática dentro do domínio da saúde. Em média, cada pergunta recebeu aproximadamente 1,32 respostas, com um máximo de três, o que reflete tanto a participação colaborativa dos profissionais de saúde quanto a heterogeneidade no engajamento das discussões.

Em termos de comprimento textual, observa-se que as respostas são, em geral, mais extensas que as perguntas — apresentando comprimento médio de 211,39 caracteres contra 141,03. Essa diferença é esperada em contextos de Pergunta-Resposta, uma vez que respostas médicas tendem a ser explicativas, incorporando justificativas clínicas ou instruções de cuidado. Esse padrão reforça a adequação do córpus para estudos de geração e avaliação de respostas longas.

A distribuição de avaliações numéricas (escala de 0 a 5) revela uma predominância marcante da classe 0, responsável por 92,5% das instâncias. Essa concentração sugere que grande parte das respostas não recebeu avaliação explícita por parte dos usuários, um fenômeno comum em portais abertos, onde a interação avaliativa é opcional. Apesar disso, o córpus ainda contém 1.263 exemplos com nota máxima (5), representando 5,4% das amostras, que podem servir como subconjunto de referência para tarefas de previsão de qualidade de respostas. A média de avaliação ($\bar{x}$=0,33) e o desvio-padrão moderado ($\sigma = 1,20$) indicam distribuição assimétrica, fortemente enviesada para valores baixos.

De forma geral, esses resultados demonstram que o SaudeBR-QA é um córpus amplo, diversificado e com propriedades estatísticas que refletem o comportamento real de usuários e profissionais em ambientes de saúde online.

5. Experimento e Validação do Córpus

Com o objetivo de obter uma evidência de que o SaudeBR-QA contém sinal útil para tarefas supervisionadas, foi realizado um experimento de classificação binária de utilidade de respostas "útil"vs. "não útil/pouco útil"sobre um subconjunto do córpus. A intenção desse experimento é verificar, se representações textuais relativamente simples, combinadas a atributos, são capazes de explorar regularidades presentes nos dadoscoletados.

Foram considerados dois subconjuntos extraídos do SaudeBR-QA, ambos compostos por perguntas acompanhadas de múltiplas respostas no domínio da saúde. O primeiro, consiste em 1.978 pares de perguntas, em que cada pergunta está associada exatamente a duas respostas. O segundo contém 1.032 pares de perguntas, nos quais cada pergunta possui três respostas distintas. Em ambos os casos, a estrutura dos dados permite comparar respostas alternativas para uma mesma pergunta, condição essencial para a estratégia de rotulagem adotada. Para o experimento relatado nesta seção, os dois subconjuntos foram combinados, totalizando 3.010 pares pergunta–resposta.

O processo de rotulagem foi guiado por um critério heurístico baseado em sinais de interação da comunidade. Em cada pergunta considerada, havia exatamente uma resposta que recebeu ao menos uma curtida. Essa resposta foi automaticamente rotulada como "útil"enquanto as demais respostas associadas à mesma pergunta, que não receberam curtidas, foram rotuladas como "não útil/pouco útil".

Essa estratégia explora o fato de que, nesses subconjuntos, a curtida é um sinal de preferência relativa: dado um conjunto de respostas para a mesma pergunta, aquela que recebe curtida tende a ser percebida como mais adequada ou mais relevante pelos usuários. Assim, este experimento deve ser entendido a detectar aprendibilidade a partir do comportamento dos usuários.

Após a rotulagem, os dados foram particionados em treino (80%) e teste (20%), utilizando divisão estratificada em relação ao rótulo, de modo a preservar a proporção original entre as classes em ambos os conjuntos.

Para representar os pares pergunta–resposta, foram utilizadas três famílias de características, combinadas em um único vetor. Em primeiro lugar, foram empregados vetores TF–IDF com vocabulários independentes para pergunta e resposta [Manning et al. 2008]. Em seguida, foram imcorporados *embeddings* FastText em português, nos quais cada texto é representado pelo vetor médio das palavras ponderado pelos pesos de TF–IDF [Hartmann et al. 2017]. Por fim, foi aplicada uma engenharia de atributos, incluindo medidas como comprimento da resposta, comprimento relativo entre pergunta e resposta e similaridade semântica entre ambos os textos.

A ideia é que o comprimento e o comprimento relativo forneçam uma indicação do nível de detalhamento e da extensão da resposta, enquanto a similaridade pergunta–resposta captura o grau de alinhamento semântico entre a dúvida do usuário e o conteúdo da resposta. Ao combinar tais atributos com as representações textuais TF–IDF e *embeddings*, obtém-se um vetor de características mais completo, capaz de incorporar tanto informações lexicais quanto sinais estruturais e semânticos adicionais.

O vetor final de características, que alimenta o classificador, é construído pela concatenação de cinco blocos distintos, conforme a seguinte expressão:

$$\mathbf{x} = \big[\text{TF–IDF}(\text{pergunta}) \parallel \text{TF–IDF}(\text{resposta}) \parallel \mathbf{v_p}(\text{pergunta}) \parallel \mathbf{v_r}(\text{resposta}) \parallel \text{extras}\big].$$

em que $\mathbf{v_p}$ e $\mathbf{v_r}$ denotam os vetores médios FastText da pergunta e da resposta, respectivamente, e $\mathbf{extras}$ agrega os atributos derivados (comprimentos e similaridade, entre outros).

Os hiperparâmetros dos classificadores foram ajustados por meio de Busca em Grade (*GridSearchCV*) [Pedregosa et al. 2011] com Validação Cruzada Estratificada (*StratifiedKFold* de 5 partições) [Kohavi 1995], aplicada exclusivamente sobre o conjunto de treino. A Tabela 3 apresenta os resultados obtidos.

Tabela 3. Baselines para *útil* vs. *não útil*.

Modelo	Acc	Prec	Rec	F1 (macro)
LinearSVC	73,09	73,00	73,00	73,00
Random Forest	**79,07**	**79,00**	**79,00**	**79,00**
Gradient Boosting	77,08	77,00	77,00	77,00

Nota: valores aproximados.

Os resultados na Tabela 3 mostram que os modelos considerados conseguem explorar, em alguma medida, o sinal presente no subconjunto rotulado, obtendo métricas significativamente acima do acaso. Entre eles, o melhor desempenho global foi alcançado pelo *Random Forest*, com acurácia e F1 (macro) em torno de 79%.

Para investigar a origem desse sinal e responder de forma mais direta à questão sobre o papel de cada família de atributos, foi conduzido um estudo de ablação por blocos de *features* utilizando o *Random Forest*, modelo que apresentou o melhor desempenho na Tabela 3. Nesse experimento, avaliou-se sistematicamente o impacto de diferentes combinações de TF–IDF, *embeddings* e atributos derivados (*extras*), mantendo o mesmo protocolo de particionamento dos dados. A Tabela 4 resume os resultados.

Tabela 4. Ablação por blocos de *features* (Random Forest).

Conjunto de features	**#feat**	**F1 (macro) hold-out**	ΔF1 vs TF–IDF
TF–IDF + Embeddings + Extras	13,002	**79,23%**	+8,16 pp
Embeddings + Extras	603	77,57%	+6,50 pp
Embeddings apenas	600	76,24%	+5,17 pp
Extras apenas	3	75,91%	+4,84 pp
TF–IDF + Embeddings	12,999	75,58%	+4,51 pp
TF–IDF + Extras	12,402	71,19%	+0,12 pp
TF–IDF apenas	12,399	71,07%	+0,00 pp

Nota: ΔF1 calculado em relação a TF–IDF apenas. Valores em %.

Os resultados da ablação mostram que a configuração completa, que combina TF–IDF, *embeddings* e atributos derivados, atinge a melhor F1 (macro), com 79,23%, resultando em um ganho de aproximadamente 8 pontos percentuais em relação ao uso exclusivo de TF–IDF.

Em síntese, os resultados das Tabelas 3 e 4 reforçam a ideia de que o SaudeBR-QA é rico tanto em informação linguística quanto em padrões estruturais relacionados ao comportamento dos usuários, tornando-o adequado para estudos de predição de utilidade, classificação e ranqueamento de respostas em saúde.

6. Conclusão, Limitações e Trabalhos Futuros

Este artigo apresentou o SaudeBR-QA, um novo córpus em larga escala com 23.382 pares de perguntas e respostas em português brasileiro, provenientes de interações reais entre pacientes e profissionais de saúde. Conforme demonstrado, o recurso supre uma lacuna importante ao disponibilizar dados contemporâneos e orientados a tarefas de QA no domínio médico em português. Os experimentos confirmaram que o córpus possui sinal linguístico e contextual suficiente para tarefas supervisionadas.

Apesar de sua contribuição, é importante reconhecer algumas limitações. Primeiramente, o córpus é derivado de uma única fonte de dados, o que pode introduzir um viés de plataforma no estilo das perguntas e respostas. Soma-se a isso a ausência de mais metadados, como data e hora das postagens ou a vinculação institucional dos profissionais de saúde. A falta dessas informações restringe o potencial de análise, reduzindo o alcance

das investigações possíveis a partir do córpus e limitando as questões de pesquisa que podem ser exploradas.

Como direções futuras, pretende-se empregar o SaudeBR-QA na avaliação de modelos de Perguntas e Respostas. Para mitigar as limitações dos rótulos de avaliação e engajamento, um passo crucial será o desenvolvimento de um subconjunto anotado por especialistas da área médica, permitindo a validação da precisão clínica. Isso habilitará o córpus para tarefas mais rigorosas de predição de qualidade e geração de respostas.

Referências

Ben Abacha, A. and Demner-Fushman, D. (2019). Medquad: Medical question answering dataset containing 47,457 qa pairs from trusted sources. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 154–164. Association for Computational Linguistics.

Ferreira, D., Alva-Manchego, F., Luz, S., et al. (2023). Semclinbr: An annotated corpus for clinical semantic textual similarity in brazilian portuguese. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing (BioNLP 2023)*, pages 37–47. Association for Computational Linguistics.

Hartmann, N. S., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and text classification. *arXiv preprint arXiv:1708.06025*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1143.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 51–56.

OECD (2023). *Health at a Glance 2023: OECD Indicators*. OECD Publishing, Paris. Chapter: Digital health. Accessed: 2025-08-30.

Pampari, A., Raghavan, P., Liang, J., and Peng, J. (2018). emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2357–2368. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Python (2025). Python language reference, version 3.x. `https://www.python.org/`. Accessed: 2025-11-07.

Richardson, L. (2025). Beautiful soup documentation. `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`. Accessed: 2025-11-07.

Santos, D. and Cardoso, N. (2006). Harem: An evaluation contest for named entity recognition in portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1986–1991. European Language Resources Association (ELRA).

Selenium (2025). Selenium webdriver. `https://www.selenium.dev/documentation/webdriver/`. Accessed: 2025-11-07.

Verma, M., Thadani, K., and Mishra, S. (2021). Powering covid-19 community q&a with curated side information. *arXiv preprint arXiv:2101.11556*.