INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

CARLOS HENRIQUE SANTOS BARROS

CLASSIFICAÇÃO DA UTILIDADE DE RESPOSTAS EM SAÚDE COM BASE EM VOTOS COMUNITÁRIOS: UMA ABORDAGEM AUTOMATIZADA PARA SISTEMAS DE QA

PICOS, PIAUÍ
2025

CARLOS HENRIQUE SANTOS BARROS

CLASSIFICAÇÃO DA UTILIDADE DE RESPOSTAS EM SAÚDE COM BASE EM VOTOS COMUNITÁRIOS: UMA ABORDAGEM AUTOMATIZADA PARA SISTEMAS DE QA

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. Dr. Rogério Figueredo de Sousa

PICOS, PIAUÍ
2025

CARLOS HENRIQUE SANTOS BARROS

CLASSIFICAÇÃO DA UTILIDADE DE RESPOSTAS EM SAÚDE COM BASE EM VOTOS COMUNITÁRIOS: UMA ABORDAGEM AUTOMATIZADA PARA SISTEMAS DE QA

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

BANCA EXAMINADORA:

Prof. Dr. Rogério Figueredo de Sousa(Orientador)
Instituto Federal do Piauí (IFPI)

Prof. Dr. Rafael Torres Anchiêta
Instituto Federal do Maranhão (IFMA)

Prof. Me. Manoel Messias Pereira Medeiros
Instituto Federal do Piauí (IFPI)

PICOS, PIAUÍ
2025

CLASSIFICAÇÃO DA UTILIDADE DE RESPOSTAS EM SAÚDE COM BASE EM VOTOS COMUNITÁRIOS: UMA ABORDAGEM AUTOMATIZADA PARA SISTEMAS DE QA

Carlos Henrique Santos Barros[1]
Rogério Figueredo de Sousa[2]

RESUMO

Este trabalho investiga a classificação da utilidade de respostas em serviços de Perguntas e Respostas na área da saúde. A pesquisa propõe um método supervisionado que combina técnicas de Processamento de Linguagem Natural e sinais de interação comunitária para priorizar respostas úteis. Foi construído o córpus SaudeBR-QA a partir do CatálogoMed, do qual se selecionaram 3.010 pares pergunta–resposta rotulados com base em votos/curtidas. Sete classificadores foram otimizados por validação cruzada estratificada. O *Random Forest* apresentou o melhor equilíbrio entre acurácia e F1 no conjunto de teste (acurácia de 79,07% e $F1_{macro}$ de 79%). A avaliação com usuários, por meio de um formulário, coletou julgamentos de público geral e profissionais de saúde, evidenciando alinhamento entre o modelo e o consenso humano e eventuais divergências. Os resultados demonstram a viabilidade de modelos leves para apoiar o ranqueamento de respostas e qualificar o acesso a informações médicas, além de disponibilizar um córpus público e um protocolo reprodutível para pesquisas futuras.

Palavras-chave: utilidade de respostas; saúde; processamento de linguagem natural; aprendizado de máquina; classificação supervisionada; Perguntas e Respostas.

ABSTRACT

This work investigates the classification of answer usefulness in health-related Question and Answer services. We propose a supervised method that combines Natural Language Processing techniques with community interaction signals to prioritize useful answers. We constructed the SaudeBR-QA corpus from CatálogoMed, from which 3,010 question–answer pairs were selected and labeled based on votes/likes. Seven classifiers were optimized via stratified cross-validation. Random Forest achieved the best balance between accuracy and F1 on the test set (79.07%) accuracy and 79% $F1_{macro}$). A user study, conducted through an online form, collected judgments from both the general public and health professionals, revealing alignment between the model and human consensus as well as occasional divergences. The results demonstrate the feasibility of lightweight models to support answer ranking and improve access to medical information, and they also provide a public corpus and a reproducible protocol for future research.

Keywords: answer helpfulness; health; natural language processing; machine learning; supervised classification; Health Question Answering.

1 INTRODUÇÃO

[1] http://lattes.cnpq.br/2475000984926109. IFPI. capic.2023118tads0035@aluno.ifpi.edu.br.
[2] http://lattes.cnpq.br/9346694618502913. IFPI. rogerio.sousa@ifpi.edu.br.

Nos últimos anos, o acesso a informações médicas na internet tem crescido rapidamente, transformando a forma como as pessoas buscam esclarecimentos sobre sintomas, diagnósticos e possíveis tratamentos (OECD, 2023). Nesse contexto, os conteúdos gerados por usuários (*User Generated Content* – UGC) tornaram-se uma das principais fontes de informação na web, permitindo a participação ativa dos indivíduos na troca e avaliação de conhecimento. Em plataformas voltadas para a área médica (*Health Question-Answering* – HQA), esses conteúdos ampliam o acesso à informação e possibilitam análises que ajudam a melhorar a qualidade e a confiabilidade das respostas fornecidas (VERMA *et al.*, 2021). No Brasil, serviços como CatálogoMed[3] e Tua Saúde[4] exemplificam esse modelo, proporcionando ambientes interativos onde especialistas e a comunidade colaboram na produção e validação de conteúdos.

Apesar da conveniência desses sistemas, ainda há o desafio de garantir a qualidade das respostas disponibilizadas. Sem supervisão organizada e validação por profissionais especializados, podem surgir informações contraditórias, respostas incompletas ou excessivamente técnicas, fatores que podem confundir usuários e favorecer a disseminação de conteúdo equivocado.

Para mitigar esses desafios, muitos serviços de HQA adotam mecanismos de avaliação comunitária, como estrelas, votos ou curtidas. Contudo, a dependência exclusiva desse tipo de *feedback* pode levar a distorções na priorização das respostas, especialmente em plataformas com grande volume de interações. Além disso, Zoratto *et al.* (2023) alertam que a popularidade de uma resposta não garante sua precisão clínica, reforçando a necessidade de métodos que integrem múltiplos critérios de avaliação. Dessa forma, a adoção de abordagens automatizadas que combinem o *feedback* da comunidade com análises textuais surge como alternativa promissora para filtrar e ranquear conteúdos de qualidade (ROY *et al.*, 2023).

Assim, para este trabalho, foi implementado e validado um método automático para a classificação da utilidade de respostas médicas em língua portuguesa. A abordagem combinou técnicas de Processamento de Linguagem Natural (PLN) com características textuais (conteúdo e estrutura) e metadados de interação (*feedback* da comunidade). Para tal, foi construído um córpus próprio contendo pares pergunta–resposta do domínio médico e seus indicadores de utilidade correspondentes. Com esse material, diferentes algoritmos de classificação foram treinados e comparados, selecionando-se o modelo de melhor desempenho para ranquear respostas e subsidiar as análises. A validação foi realizada por meio da comparação das predições com avaliações de participantes e especialistas da área da saúde. Os resultados indicam que a solução proposta contribui para o aprimoramento do ranqueamento em serviços de HQA e para a ampliação da disponibilidade de informação, apoiando a tomada de

[3] <https://www.catalogo.med.br>
[4] <https://www.tuasaude.com>

decisão dos usuários.

2 TRABALHOS RELACIONADOS

A classificação de respostas úteis em plataformas de *Community Question Answering* (CQA) na área da saúde tornou-se um campo relevante de pesquisa, dada a necessidade de filtrar informações confiáveis em ambientes colaborativos. Esta seção revisa os trabalhos que contextualizam o presente estudo, abordando: mecanismos tradicionais de avaliação comunitária; modelos automáticos clássicos de aprendizado de máquina; e Modelos Supervisionados e Embeddings em CQA e Textos Clínicos.

2.1. MECANISMOS TRADICIONAIS DE AVALIAÇÃO COMUNITÁRIA

As abordagens iniciais em sistemas de Perguntas e Respostas (*QA*) voltados para a saúde baseavam-se em sinais comunitários — como quantidade de comentários, votos e reputação do autor — como indicador para a qualidade das respostas.

Wongchaisuwat, Klabjan & Jonnalagadda (2016) desenvolveram um algoritmo semi-supervisionado para automatizar a seleção da melhor resposta em plataformas de *QA* sobre saúde. Utilizando técnicas de recuperação de informação e aprendizado semi-supervisionado, a abordagem foi avaliada em um córpus do *Yahoo! Answers*, com foco em perguntas sobre alcoolismo. Os resultados mostraram que o modelo atingiu uma acurácia de **86,2%**. Além disso, o uso de *features* do *Unified Medical Language System (UMLS)* aumentou a performance em cerca de 8 pontos percentuais, confirmando o valor de incorporar terminologia biomédica ao processo. Entre os atributos mais relevantes para distinguir respostas válidas das inválidas destacam-se: (I) comprimento do texto; (II) quantidade de *stop words* na pergunta; (III) distância semântica entre a pergunta testada e outras perguntas do córpus; e (IV) número de termos relacionados à saúde sobrepostos entre pergunta e resposta.

O desempenho elevado em um contexto desafiador evidencia a eficácia de abordagens híbridas que combinam métricas sociais com processamento de linguagem e conhecimento médico estruturado.

2.2. MODELOS AUTOMÁTICOS CLÁSSICOS E APRENDIZADO DE MÁQUINA

As abordagens baseadas em *machine learning (ML)* passaram a oferecer alternativas mais robustas aos métodos puramente baseados em regras. Yadav, Yadav *et al.* (2024) apresentaram uma revisão abrangente sobre os sistemas de *Question Answering (QA)* médico, cobrindo desde técnicas clássicas de recuperação de informação até os avanços com modelos de *Deep Learning (DL)* e *Transformers*. O estudo destacou que a incorporação de modelos pré-treinados, como *BioBERT* e *ClinicalBERT*, tem

superado representações como *bag-of-words*, alcançando melhorias consistentes em *benchmarks* biomédicos — por exemplo, *BioBERT* apresentou desempenho superior em relação a modelos genéricos no *BioASQ 2022* e *ClinicalBERT* obteve ganhos relevantes em tarefas clínicas no *SQuAD* e *PubMed*.

Além disso, o sistema *MedTalk*, desenvolvido pelos autores, alcançou **85%** de *F1-score* em testes de desempenho, confirmando a eficácia de arquiteturas baseadas em aprendizado profundo. A revisão também evidenciou que a integração de bases biomédicas estruturadas, como ontologias e *knowledge graphs*, tem potencial para aumentar a confiabilidade e a explicabilidade das respostas em contextos clínicos críticos, sendo particularmente útil para consultas complexas.

Enquanto isso, An, Li & Zhou (2023) realizaram uma análise sistemática sobre as aplicações de *machine learning* na saúde, avaliando um total de 93 publicações. Os autores propuseram uma taxonomia para organizar as principais áreas de aplicação, incluindo diagnóstico, prognóstico, monitoramento remoto, suporte à decisão clínica e descoberta de fármacos. O estudo evidenciou que modelos supervisionados, como *Support Vector Machines (SVMs)*, *Random Forests* e *Gradient Boosting*, têm sido amplamente empregados em tarefas de classificação de dados clínicos, apresentando desempenho consistente. Já as técnicas não supervisionadas, como *clustering* e *autoencoders*, demonstraram potencial promissor para a detecção de anomalias e a identificação de padrões ocultos em grandes conjuntos de dados, embora os autores ressaltem a necessidade de validação mais robusta em cenários clínicos reais.

Esses achados reforçam que a introdução de técnicas de *ML* tem ampliado a acurácia e a eficiência de diferentes aplicações na saúde, contribuindo para sistemas capazes de apoiar decisões clínicas em contextos complexos.

2.3. REPRESENTAÇÕES VETORIAIS E EMBEDDINGS EM CQA E SAÚDE

O estudo de Tian, Zhang & Li (2013) buscou enfrentar o problema da ausência de respostas aceitas em plataformas*(CQA)*, como o *Stack Overflow*. Para isso, os autores analisaram um subconjunto com 103.793 perguntas e 196.145 respostas, representadas numericamente por *TF- IDF (Term Frequency-Inverse Document Frequency)* e descritas por 16 atributos divididos em três grupos: (i) qualidade do conteúdo da resposta (tamanho, legibilidade, presença de código ou links), (ii) relação pergunta–resposta (similaridade textual e tempo de resposta) e (iii) contexto entre respostas (número de competidores, ordem de publicação e similaridade entre respostas). Utilizando o classificador *Random Forest*, obtiveram uma acurácia média de 72,27%, destacando que o contexto entre respostas foi o fator mais determinante. Respostas aceitas tendiam a ser postadas mais cedo, eram mais detalhadas, diferentes das demais e recebiam maior número de comentários.

De forma complementar, Santin *et al.* (2024) aplicaram Processamento de Lingua-

gem Natural para classificar laudos radiológicos de doenças benignas da vesícula biliar com indicação cirúrgica. Um dos pontos do trabalho foi o uso de *Embeddings*, especificamente o modelo *Word2Vec*, que transformou palavras em vetores de 300 e 1000 dimensões, permitindo capturar relações semânticas no texto clínico. Dois modelos de *deep learning* foram testados: *CNN* e *BiLSTM*. Ambos alcançaram alto desempenho, com *F1-score* superior a 0,95, sendo a BiLSTM a mais eficiente (0,96732 em ambas as dimensões). Os resultados mostraram que, devido ao vocabulário técnico restrito dos laudos radiológicos, o aumento da dimensionalidade do *Word2Vec* não gerou ganhos expressivos. Assim, *embeddings* menores se mostraram tão eficazes quanto representações mais complexas, oferecendo boa performance com menor custo computacional.

Em conclusão, os estudos evidenciam a eficácia dos diferentes tipos de representações no contexto, destacando como elas contribuem significativamente para a melhoria dos resultados em trabalhos e pesquisas voltados para a área. Essas representações facilitam a compreensão de dados, o que, por sua vez, favorece a tomada de decisões mais informadas e eficazes no campo trabalhado.

3 REVISÃO TEÓRICA

A tarefa de modelagem e predição da utilidade de opiniões insere-se no campo da Mineração de Opiniões, uma área da ciência da computação que visa analisar a qualidade e relevância de conteúdos gerados por usuários. Os primeiros trabalhos preocupados com a utilidade das opiniões emitidas surgiram por volta de 2006 (KIM *et al.*, 2006; ZHANG; VARADARAJAN, 2006; GHOSE; IPEIROTIS, 2007). Eles perceberam o aumento da quantidade de conteúdo opinativo gerado pela comunidade na Web e que esse conteúdo variava grandemente em qualidade e, portanto, era necessário determinar quais comentários os usuários considerariam úteis (KIM *et al.*, 2006).

A utilidade de uma opinião ou comentário pode ser entendida como a relevância e a eficácia de uma resposta ou avaliação em relação às necessidades e expectativas dos usuários. Kim *et al.* (2006) mencionam alguns problemas que justificam a relevância da tarefa. Na época, os principais sites de comércio eletrônico (*Amazon.com*, *Overstock.com*, *Epinions.com*) já estavam preocupados em destacar os comentários mais úteis e já permitiam aos usuários votarem nos comentários que acreditavam terem sido úteis para eles. Entretanto, essa abordagem possui limitações evidentes. Comentários recentes ou pouco expostos tendem a receber menos votos, itens de baixa popularidade não acumulam evidências suficientes. Além disso, existe uma crescente preocupação com a manipulação por *spammers* (LIU *et al.*, 2007; PANG; LEE, 2008). Pessoas mal intencionadas que, desonestamente, fazem avaliações falsas para que comentários forjados sejam mais bem ranqueados.

Diante disso, é relevante que a utilidade dos comentários seja avaliada automaticamente assim que eles sejam submetidos, permitindo acelerar a consolidação dos ranques de comentários e consequentemente o feedback aos usuários. A abordagem automatizada, utilizando modelos de aprendizado de máquina, surge como uma solução promissora. De acordo com Liu (2012), o objetivo é determinar automaticamente a utilidade de cada opinião. Esse tipo de modelagem permite que a avaliação da utilidade seja realizada de maneira mais objetiva e independente das flutuações de popularidade e do viés dos usuários. Mais atualmente, o estudo de Diaz & Ng (2018) define a tarefa mais amplamente da seguinte forma: "A modelagem e predição da utilidade de opiniões é uma tarefa que estuda os fatores que determinam a utilidade das opiniões e tenta corretamente predizê-la". Essa definição amplia a abordagem inicial ao incluir uma análise dos diversos fatores que influenciam a percepção de utilidade, como contexto semântico, relevância do conteúdo, e a interação com o público. A tarefa, portanto, envolve não apenas a identificação das características textuais e contextuais das opiniões, mas também a integração de variáveis como reputação do autor e feedback da comunidade, que são essenciais para uma predição precisa da utilidade das opiniões em plataformas colaborativas.

Apesar da tarefa ter se iniciado no domínio de opiniões (produtos, serviços) ela pode ser facilmente expandida para outros tipos de textos. Nos últimos anos, a tarefa de classificação da utilidade se expandiu para diversos domínios, como mercado automotivo (CAO; YANG, 2022), turismo (XIA, 2023), medicamentos (ZHENG *et al.*, 2021), e na área de saúde (SHAH; MUHAMMAD; LEE, 2022). E o mesmo acontece no domínio de perguntas e respostas sobre saúde. Trabalhos como Qiu *et al.* (2022), Hu *et al.* (2017) e Yadav, Yadav *et al.* (2024) demonstram como a comunidade está desenvolvendo técnicas para escolher perguntas e respostas de qualidade nos serviços de HQA. No estudo de (QIU *et al.*, 2022) utilizaram combinação de características densidade de palavras-chave e dados de comportamento do usuário, os autores conseguiram prever a qualidade das respostas. Modelos baseados em DL, como BERT e ClinicalBERT, que são treinados especificamente para dados biomédicos, têm se mostrado altamente eficazes no contexto da saúde, superando abordagens tradicionais em tarefas de classificação de respostas (YADAV; YADAV *et al.*, 2024).

Em síntese, a literatura converge para dois eixos principais no campo da classificação de utilidade:

1. Identificação de Features Informativas: São as características que descrevem e capturam o conteúdo e contexto de uma resposta. Estas incluem as propriedades textuais, como a qualidade do conteúdo, a relevância semântica e os metadados de interação, como o número de curtidas, a reputação do autor e o tempo de publicação.

2. Abordagens de Modelagem: As técnicas utilizadas para modelar a utilidade variam desde os métodos clássicos de aprendizado supervisionado (SVMs, árvores

de decisão e regressão logística) até arquiteturas neurais complexas (*deep learning, transformers* e redes neurais convolucionais). A integração dessas abordagens permite mitigar os vieses nos sinais comunitários, proporcionando uma classificação mais robusta e precisa das respostas.

4 METODOLOGIA

Neste capítulo, são apresentadas as etapas adotadas para a realização deste estudo, que visam classificar a utilidade de respostas em plataformas de Perguntas e Respostas voltadas à saúde. A metodologia é composta por um conjunto de procedimentos sistemáticos para coleta, organização e processamento dos dados, bem como para a construção e avaliação do modelo de aprendizado de máquina.

Quanto ao tipo e à abordagem, este estudo é: (i) aplicado, por buscar resolver o problema prático de classificar automaticamente a utilidade das respostas; (ii) experimental, por construir, treinar e testar modelos sob condições controladas; (iii) descritivo, por caracterizar o córpus e os padrões pergunta–resposta; abordagem mista(quanti qualitativa), pois combina análises quantitativas (métricas de avaliação e testes estatísticos) com análises qualitativas (inspeção de casos representativos e análise de erros), conforme a literatura metodológica(LAKATOS; MARCONI, 2017).

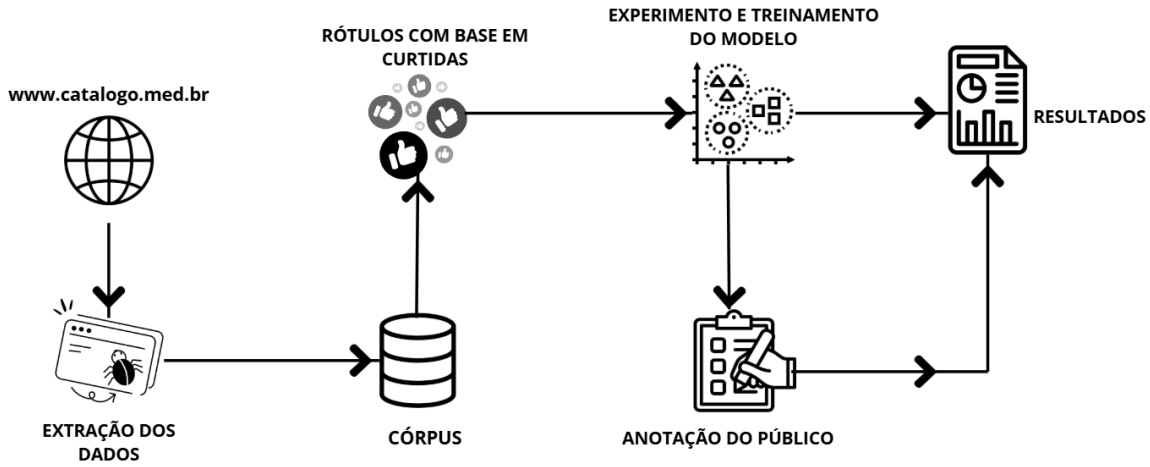

Figura 1 – Fluxo de trabalho para construção do modelo de aprendizado de máquina.

4.1. COLETA E ORGANIZAÇÃO DO CÓRPUS

Nesta seção, são detalhados a origem e organização dos dados para realização da tarefa.

4.1.1. Extração dos Dados

A primeira etapa consistiu na coleta dos dados por meio de técnicas de *web scraping*[5], utilizando a biblioteca *Selenium*[6]. O *web scraping* é um método automatizado de extração de informações disponíveis em páginas da Web, permitindo coletar grandes volumes de dados de forma estruturada. Já o *Selenium* é uma ferramenta que possibilita a automação de navegadores, simulando a navegação humana para percorrer o site, interagir com elementos dinâmicos e extrair o conteúdo desejado.

Um *script*[7] em linguagem Python (Python, 2025) foi desenvolvido para este trabalho e aplicado ao site *CatálogoMed*, a plataforma reúne uma grande quantidade de perguntas feitas por usuários sobre questões de saúde, com respostas fornecidas por médicos e avaliações atribuídas a cada resposta. Isso resultou em um *córpus* composto por mais de 15.000 perguntas, cada uma associada a até três respostas fornecidas por especialistas.

Foram extraídas as seguintes informações:

- **Pergunta:** O enunciado da dúvida postada pelo usuário.
- **Respostas:** Até três respostas fornecidas por especialistas para cada pergunta.
- **Especialidade:** A área médica à qual a pergunta se refere.
- **Avaliação da Resposta:** Sistema de estrelas, onde quanto maior a quantidade de estrelas, melhor é a avaliação da resposta.
- **Curtidas:** Número de curtidas atribuídas a cada resposta.

Cabe destacar que, por se tratar de uma plataforma voltada à orientação em saúde, as respostas disponibilizadas no CatálogoMed são elaboradas por profissionais cadastrados no sistema. Esse cadastro exige a vinculação do profissional a um número de registro em conselho de classe (como o Conselho Regional de Medicina), o que reforça a confiabilidade das informações fornecidas. Além disso, os perfis dos especialistas geralmente incluem nome completo, especialidade médica e local de atendimento, permitindo maior transparência e rastreabilidade das respostas publicadas.

4.1.2. Estruturação e Disponibilização do córpus

Os dados coletados foram organizados em um arquivo no formato CSV, em que cada linha representa uma instância composta por uma pergunta e suas respectivas respostas, juntamente com os metadados associados.

[5] <https://pypi.org/project/beautifulsoup4/>
[6] <https://pypi.org/project/selenium/>
[7] <https://github.com/CarlHenry670/Crawler-QA>

Essa estrutura permite a associação direta entre a dúvida do usuário, as respostas fornecidas, a especialidade médica envolvida e os indicadores de avaliação da comunidade. A padronização do córpus em um formato tabular facilita tanto o pré-processamento quanto a integração posterior com ferramentas de análise e modelagem de aprendizado de máquina, garantindo maior eficiência e consistência ao longo das etapas subsequentes.

O córpus, denominado SaudeBR-*QA*, foi publicado no GitHub[8], tornando-se acessível a qualquer pessoa que deseje utilizá-lo para pesquisas ou outras aplicações. A disponibilização pública do córpus facilita a replicação dos experimentos e promove a transparência, além de incentivar o uso da base de dados para outras análises e desenvolvimentos relacionados à área de saúde.

4.2. PRÉ-PROCESSAMENTO E MODELAGEM EXPERIMENTAL

Nesta seção, descrevemos o encadeamento de etapas que prepara os dados textuais e estabelece o protocolo de modelagem. Em linhas gerais, seguimos quatro frentes: filtragem e rotulação do córpus; pré-processamento linguístico; construção de representações e atributos; e protocolo experimental.

4.2.1. Conjunto de Dados e Rotulagem

Para este estudo, foram selecionados dois subconjuntos retirados do SaudeBR-*QA* que contêm pares de perguntas e respostas do domínio de saúde. O primeiro córpus, denominado `córpus_Duas_Respostas.csv`, consiste em 1.978 pares de perguntas e respostas. Cada pergunta é acompanhada de duas respostas, sendo que apenas uma delas foi curtida pela comunidade, ela foi rotulada como "útil"enquanto a outra "não útil/pouco útil". O segundo córpus, `Perguntas_Tres_Respostas.csv`, contém 1.032 pares de perguntas e respostas. Nesse caso, a resposta curtida, e única, foi automaticamente considerada "útil", enquanto o restante foi rotulada como "não útil/pouco útil".

Após a leitura e organização dos arquivos, os dados foram concatenados em um único *dataframe*, preservando a estrutura de múltiplas respostas para cada pergunta. Isso permite a análise de diferentes abordagens e pontos de vista para uma mesma dúvida, o que enriquece a tarefa de classificação. Cada instância no conjunto de dados corresponde a um par *(pergunta, resposta)*, com seus metadados associados.

4.2.2. Etapas de Limpeza e Normalização dos Textos

Os textos foram submetidos a um processo de pré-processamento. Para essa etapa, empregou-se o modelo linguístico `pt_core_news_sm` do *spaCy* (HONNIBAL *et al.*, 2020;

[8] <https://github.com/CarlHenry670/SaudeBR-QA>

Explosion, 2025), biblioteca de código aberto voltada a aplicações de produção em PLN. O *pipeline* do *spaCy* oferece recursos como tokenização, segmentação de sentenças, lematização, etiquetagem morfossintática (POS), reconhecimento de entidades nomeadas (NER) e análise de dependências. O modelo citado é especializado em português e foi utilizado na etapa de limpeza, que incluiu:

Normalização em minúsculas: Todos os textos foram convertidos para minúsculas para reduzir a variabilidade da escrita e garantir consistência na análise.

Remoção de stopwords: Palavras comuns que não contribuem significativamente para o significado do texto.

Remoção de tokens não alfabéticos: Tokens como números, pontuações e símbolos não alfabéticos foram descartados, pois não possuem valor semântico relevante para a análise.

Esse pré-processamento foi realizado tanto no campo de pergunta quanto no campo de resposta, assegurando que ambos os conjuntos de dados seguissem a mesma metodologia de limpeza e transformação. Como resultado, os campos *Pergunta_pp* (pré-processada) e *Resposta_pp* (pré-processada) foram gerados, contendo as versões limpas e normalizadas dos textos. Esse procedimento de limpeza e padronização tem como objetivo a redução do ruído lexical, o que melhora a qualidade dos dados e facilita as etapas seguintes de processamento.

4.2.3. Representação de Features com TF-IDF

Como primeira camada de representação textual, foi utilizada uma abordagem do TF-IDF (*Term Frequency-Inverse Document Frequency*)[9], que visa quantificar a relevância de uma palavra para um documento dentro de um córpus. A principal vantagem desta técnica é sua capacidade de destacar termos que são contextualmente importantes, indo além da simples contagem de frequência.

O cálculo foi realizado em duas etapas principais, utilizando o `scikit-learn`:

1. **Frequência do Termo (TF):** Primeiramente, através do `CountVectorizer`, construímos vocabulários distintos para o conjunto de perguntas e o de respostas. Para cada texto, foi gerado um vetor de contagens, onde cada posição representa a frequência de uma palavra do vocabulário naquele documento.

2. **Frequência Inversa do Documento (IDF):** Em seguida, com o `TfidfTransformer`, calculamos o peso IDF para cada palavra. Esse componente reduz o peso de termos muito comuns em todo o corpus (como "médico"ou "tratamento") e aumenta o peso de termos mais raros e específicos, que são potencialmente mais informativos para a tarefa de classificação.

[9] <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

O valor final do TF-IDF para cada palavra é o produto dessas duas métricas. As matrizes resultantes, uma para perguntas e outra para respostas. O resultado é um vetor de características de alta dimensionalidade que captura a importância lexical de cada termo no par pergunta-resposta.

4.2.4. Embeddings FastText

Para aprimorar a representação textual e capturar nuances semânticas foi incorporado uma representação vetorial densa através de *embeddings* de palavras. Foram utilizados vetores *FastText* de 300 dimensões (skip_s300) entre as embeddings disponíveis no NILC (Núcleo Interinstitucional de Linguística Computacional) [10] que se destaca por capturar relações semânticas entre palavras, especialmente quando se trata de palavras desconhecidas ou fora do vocabulário. Para cada texto, computamos a média ponderada por TF–IDF sobre as palavras presentes no vocabulário do modelo:

$$\mathbf{v}(t) = \frac{\sum_{w \in t} \mathsf{tfidf}(w,t)\, \mathbf{e}(w)}{\sum_{w \in t} \mathsf{tfidf}(w,t)},$$

em que $\mathbf{e}(w)$ é o vetor *FastText* de w. Esse procedimento foi aplicado separadamente à pergunta e à resposta, gerando dois vetores de 300 dimensões.

4.2.5. Engenharia de Atributos e Representação Final

Com o objetivo de fornecer ao modelo mais informações contextuais e aprimorar sua capacidade de análise, foram extraídos três atributos numéricos adicionais:

1. **Comprimento da resposta** (`ans_len`): número de palavras em `Resposta_pp`.
2. **Similaridade pergunta–resposta** (`similaridade_p_r`): cosseno entre os embeddings médios ponderados da pergunta e da resposta, $\cos(\theta) = \frac{\mathbf{v_p} \cdot \mathbf{v_r}}{\|\mathbf{v_p}\| \|\mathbf{v_r}\|}$.
3. **Comprimento relativo** (`ans_len_relativa`): razão entre o comprimento da resposta e o maior comprimento observado entre as respostas da mesma pergunta.

A adição desses atributos permite que o modelo capture nuances importantes nas respostas, como a sua profundidade, relevância e adequação. O comprimento da resposta e o comprimento relativo ajudam a entender o nível de detalhamento das respostas, enquanto a similaridade entre pergunta e resposta reforça a conexão semântica entre os textos.

Ao combinar essas características com as representações TF-IDF e *embeddings*, é possível obter um vetor de características mais completo, que proporciona uma análise mais precisa e contextualizada, aprimorando a capacidade do modelo de classificar corretamente a utilidade das respostas. O vetor final de características, representado

[10] <http://www.nilc.icmc.usp.br/embeddings>

por $\mathbf{x}$, que alimenta o classificador, é construído através da concatenação de cinco blocos de atributos distintos. A composição do vetor é formalmente dada por:

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

Posteriormente, os dados foram particionados em treino (80%) e teste (20%), com estratificação pelo rótulo. Todos os experimentos foram executados em ambiente Python no Google Colab[11].

4.3. CONSTRUÇÃO E ANÁLISE DOS CLASSIFICADORES

A etapa final do estudo consistiu no desenvolvimento e na avaliação de modelos de aprendizado de máquina para a classificação das respostas.

4.3.1. Otimização dos Modelos Supervisionados

Foram avaliados sete algoritmos de classificação: ***LinearSVC***, ***Random Forest***, ***Gradient Boosting***, ***MLP***, ***ComplementNB***, ***BernoulliNB*** e ***Logistic Regression***. Foi construído um pipeline que integrou as representações textuais (TF–IDF e *embeddings*), as representações numéricas (atributos) e o classificador.

Para encontrar a combinação ótima de hiperparâmetros, foi empregada a técnica de Busca em Grade com Validação Cruzada Estratificada (*GridSearchCV* com *StratifiedKFold* de 5 partições). A métrica de otimização utilizada foi o F1-Score, que foi escolhida por seu equilíbrio entre precisão e cobertura.

4.3.2. Métrica de Avaliação

O desempenho final dos modelos otimizados foi aferido no conjunto de teste. As seguintes métricas de avaliação foram calculadas:

- Acurácia;
- Precisão, Cobertura e F1-Score
- Matriz de Confusão

Este delineamento metodológico permite avaliar, de forma objetiva, se as características textuais são preditores eficazes na distinção entre respostas úteis e não úteis em serviços de HQA. Entre os modelos testados, aquele que apresentou o melhor desempenho foi o **Random Forest**, alcançando resultados superiores em equilíbrio entre acurácia e F1-Score quando comparados às demais abordagens.

[11] <https://colab.research.google.com/drive/1pjJbNicPIPCP5ELmjdK457cU5CekbioG?usp=sharing>

4.4. AVALIAÇÃO DAS RESPOSTAS DO MODELO

Realizou-se uma avaliação manual das respostas para comparar a *utilidade* estimada pelo modelo à *utilidade* percebida pelas pessoas, captando a percepção humana e eventuais divergências. Os julgamentos dos avaliadores permitem delimitar, com critérios explícitos, o limiar entre respostas úteis e não úteis e revelar padrões de acerto e de falha. Esta seção apresenta os resultados dessa avaliação sob a perspectiva do público.

4.4.1. Elaboração e Coleta do Formulário

Para a coleta dos julgamentos, foi desenvolvida uma aplicação web interativa em *Streamlit* [12]. O *Streamlit* é uma biblioteca em Python voltada à construção rápida de interfaces e aplicações de ciência de dados diretamente a partir de scripts, oferecendo componentes reativos. Em cada item, o participante deveria avaliar a resposta associada como útil ou não útil/pouco útil, registrando sua decisão por meio de um rótulo binário diretamente na interface da aplicação, conforme a Figura 2. Para maior fidelidade ao uso real, a coleta considerou dois perfis de avaliadores (público geral e profissionais de saúde), contrastando utilidade percebida e pertinência clínica. A tela inicial permitia a seleção do perfil e oferecia instruções sobre a tarefa de avaliação, conforme ilustrado na Figura 3.

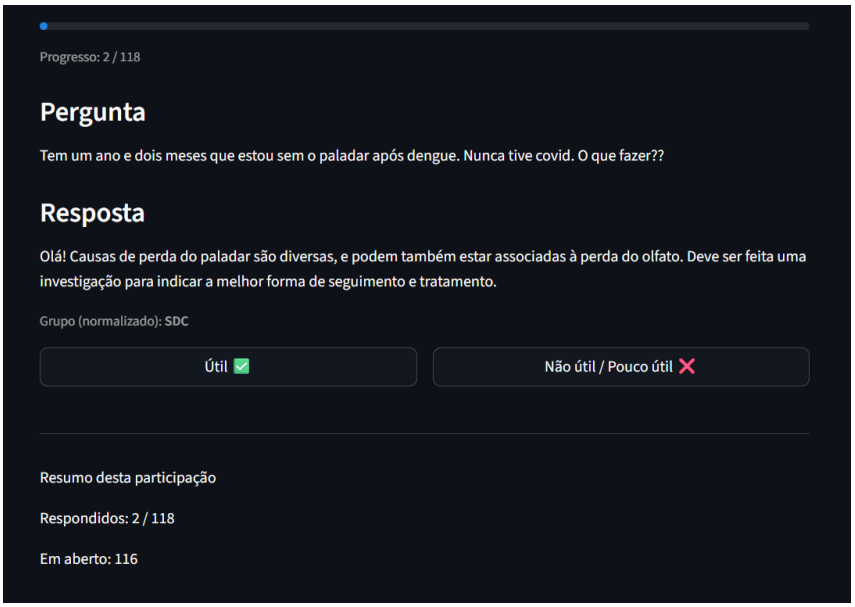

Figura 2 – Interface de julgamento: exibição do par pergunta e resposta e registro de utilidade.

[12] <https://pypi.org/project/streamlit/>

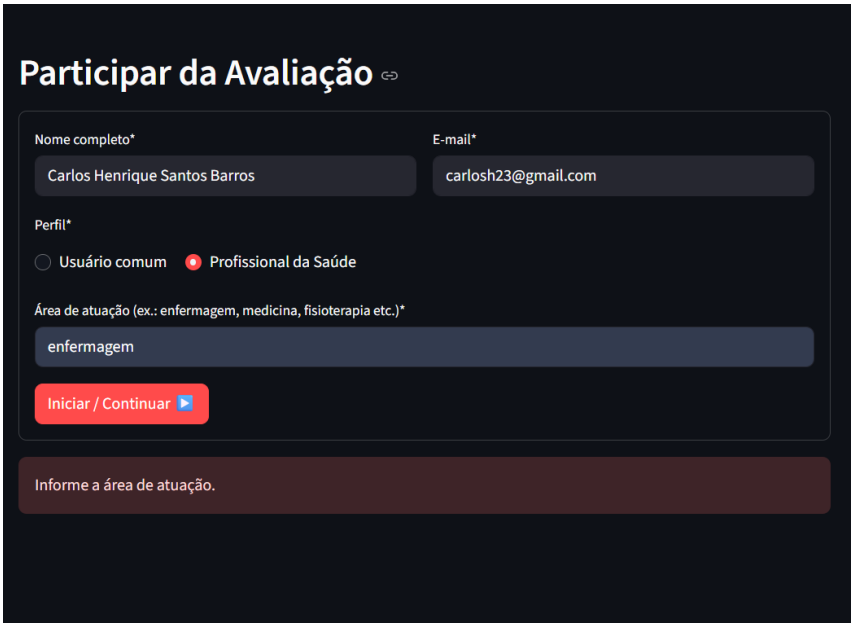

Figura 3 – Tela de seleção do perfil do participante na aplicação em Streamlit.

O *link* de participação foi compartilhado com esses públicos para recrutamento voluntário. Ao todo, foram coletadas 16 respostas, sendo 6 de participantes da área da saúde e 10 do público geral. Entre os participantes da saúde, incluem-se estudantes de nível superior e profissionais das áreas de Enfermagem, Nutrição e Psicologia.

Cada participante avaliou 118 pares de pergunta e resposta, amostrados de forma aleatória a partir do conjunto total de 354 pares balanceados por classe. Esse quantitativo de pares corresponde a aproximadamente um terço do total e foi definido para ampliar a variabilidade observada entre os diferentes tipos de pares. Os itens apresentados no formulário foram organizados em três eixos temáticos, de modo a verificar a análise dos julgamentos aos tipos de pergunta:

- **Sintomas, diagnósticos e consultas (SDC)**: questões em que os usuários descrevem sinais clínicos e buscam hipóteses diagnósticas ou recomendações sobre quando procurar atendimento.
- **Tratamento e medicação (TM)**: dúvidas relacionadas ao uso de medicamentos, efeitos colaterais, interações e condutas terapêuticas seguras.
- **Sintomas físicos (SF)**: perguntas centradas em manifestações corporais objetivas, como inchaços, manchas ou alterações visíveis.

Os pares de pergunta e resposta utilizados na aplicação foram retirados do conjunto de teste empregado na validação do classificador *Random Forest*. Esse modelo foi selecionado por ter apresentado a melhor performance entre os classificadores avaliados no estudo, de modo que a etapa com usuários funcionou como uma validação

externa da utilidade percebida das respostas no mesmo universo de itens que serviu para aferir o desempenho do modelo.

4.4.2. Cálculos de Concordância

Para quantificar a consistência dos julgamentos entre anotadores, foram selecionados três avaliadores de cada grupo (usuários comuns e área da saúde), escolhidos aqueles que apresentaram a maior quantidade de avaliações em comum, o que permite a comparação direta entre julgamentos. A concordância foi mensurada por meio de duas métricas para dados categóricos: *Fleiss'* κ e *Krippendorff's* α.

Fleiss' κ mede a concordância corrigida pelo acaso para *m* anotadores em categorias nominais, sendo definida por $\kappa = (P_o - P_e)/(1 - P_e)$, em que P_o é a proporção de concordância observada e P_e a esperada ao acaso. Para interpretar os valores de κ, foram adotados os limiares: $> 0{,}80$ (quase perfeito), $> 0{,}60$ (substancial), $> 0{,}40$ (moderado), $> 0{,}20$ (razoável), > 0 (leve) e ≤ 0 (sem acordo) (LANDIS; KOCH, 1977).

Krippendorff's α avalia a confiabilidade entre avaliadores e, no caso nominal, é dado por $\alpha = 1 - \frac{D_o}{D_e}$ (desacordo observado versus esperado). Seguindo a orientação de Marzi, Balzano & Marchiori (2024), a interpretação recomendada é: $\alpha = 1$ (concordância perfeita); $\alpha \geq 0{,}80$ (nível aceitável para conclusões trianguladas); $0{,}67 \leq \alpha < 0{,}80$ (uso tentativo/provisório); $\alpha < 0{,}67$ (baixa confiabilidade); $\alpha = 0$ (sem acordo) e $\alpha < 0$ (discordância sistemática).

5 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da aplicação da metodologia descrita anteriormente. O capítulo está organizado em duas partes principais: a primeira detalha a análise comparativa da modelagem supervisionada, e a segunda avalia o alinhamento entre o modelo e os julgamentos dos anotadores.

5.1. AVALIAÇÃO DA MODELAGEM SUPERVISIONADA

Esta seção, identifica o modelo com melhor *trade-off* de desempenho, diagnostica onde ocorrem confusões entre classes e isola quais atributos e etapas de pré-processamento mais impactam o resultado, orientando a escolha e a justificativa do modelo final.

5.1.1. Desempenho Comparativo dos Modelos

A etapa inicial da análise de resultados concentrou-se em comparar o desempenho dos modelos supervisionados na tarefa de classificar a **utilidade** de respostas em saúde. A Tabela 1 apresenta, lado a lado, a acurácia (Acc), a precisão (Prec), a cobertura

(Rec) e o F1 (macro) dos sete classificadores no conjunto de teste, permitindo avaliar o equilíbrio entre acerto global e capacidade de identificar respostas úteis.

Tabela 1 – Desempenho no conjunto de teste.

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
MLP	66.11	66.00	66.00	66.00
ComplementNB	55.15	55.00	55.00	55.00
BernoulliNB	53.16	53.00	53.00	53.00
Logistic Regression	62.13	62.00	62.00	62.00

Nota: valores estão aproximados. Melhor F1 (macro) em negrito.

O *Random Forest* apresentou o melhor equilíbrio entre *precision*, *recall* e F1 (macro), com acurácia de 79,07% e F1(macro) de ≈79%. *Gradient Boosting* veio em seguida (≈77%), e o *LinearSVC* obteve ≈73%. Modelos lineares simples (*Logistic Regression*) e Naive Bayes tiveram desempenho inferior (≈62% e ≈55–53%), sugerindo que as relações entre as features e o rótulo exigem modelagem não linear e/ou agregação de múltiplas árvores.

5.1.2. Matriz de Confusão do Modelo Final

A seguir, a matriz de confusão do modelo final. No conjunto aproximadamente balanceado (classe 0: 303; classe 1: 299). A Figura 4 exibe a matriz de confusão do classificador escolhido (Random Forest).

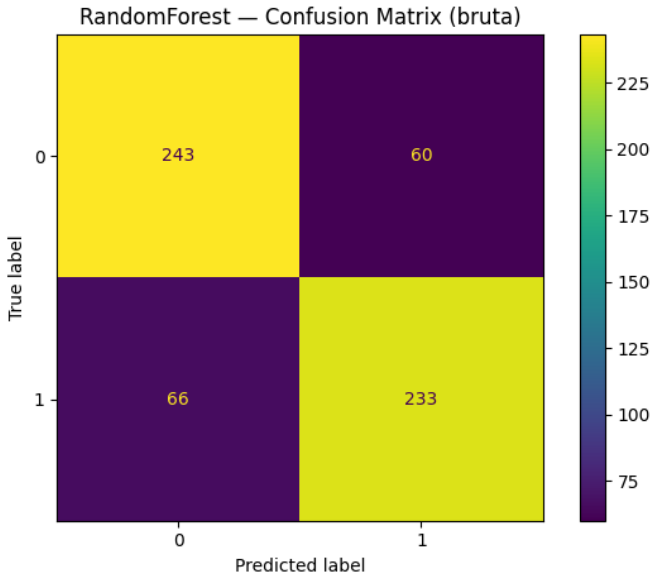

Figura 4 – Matriz de confusão do modelo final no teste.

Observa-se a seguinte distribuição na matriz de confusão: *verdadeiros negativos* (243; respostas não úteis corretamente rejeitadas), *falsos positivos* (60; respostas não úteis classificadas como úteis), *falsos negativos* (66; respostas úteis não reconhecidas) e *verdadeiros positivos* (233; respostas úteis corretamente identificadas), indicando leve predominância de perdas de respostas úteis.

5.1.3. Estudo de Ablação

O estudo de ablação é uma abordagem em experimentos de aprendizado de máquina, onde se avalia a contribuição de diferentes componentes ou características de um modelo. Isso é feito removendo ou modificando partes do modelo para observar como a performance é afetada (SHEIKHOLESLAMI, 2019). Neste estudo, a avaliação a partir do TF-IDF ao devido à sua eficácia nos testes iniciais.

Tabela 2 – Ablação por blocos de *features* (Random Forest).

Conjunto de features	**#feat**	**F1 (macro) hold-out**	$\triangle$F1 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: $\triangle$F1 calculado em relação a TF–IDF apenas. Valores em %.

O melhor arranjo de *features* foi o conjunto completo, atingindo $F1_{macro} \approx 79{,}23\%$ no *hold-out* e superando TF–IDF isolado em $+8{,}16$ p.p. Sequências com *Embeddings* (sozinhos ou combinados) e até mesmo os Extras isolados já superam TF–IDF puro, indicando que informações semânticas densas e atributos manuais compactos (apenas 3 variáveis) são particularmente informativos. A proximidade entre as médias de CV (70–72%) e o desempenho no *hold-out* sugere boa generalização e baixo *overfitting*.

5.2. VALIDAÇÃO COM USUÁRIOS E ALINHAMENTO MODELO–PESSOAS

A segunda etapa dos resultados foca nas avaliações humanas sobre as respostas do modelo e as conclusões decorrentes dos achados.

5.2.1. Visão geral: (modelo × consenso humano)

Figura 5 – início do painel de resultados do formulário.

Tabela 3 – Visão geral dos julgamentos e desempenho do modelo.

Métrica	**Valor**
Pares no córpus	354
Pares já avaliados	352
Cobertura	99,4%
Acurácia (modelo × consenso humano)	57,4%

A Figura 5 indica que o conjunto possui 354 pares de pergunta–resposta, dos quais 352 foram avaliados, resultando em cobertura de 99,4%. A acurácia do modelo em relação ao consenso humano aparece como 57,4%. Foi considerado como útil os pares em que 50% ou mais dos avaliadores classificaram como úteis, pois, mesmo não sendo a maioria absoluta, esse valor indica que pelo menos metade dos avaliadores concordou na utilidade da resposta. Portanto, A acurácia de 57,4% representa o alinhamento do modelo com o consenso dos avaliadores humanos. Ou seja, em 57,4% das vezes, a classificação do modelo coincidiu com a dos humanos. Esse número resume, de forma agregada, o alinhamento do classificador com o rótulo de consenso dos avaliadores humanos, servindo como uma medida do sistema às decisões humanas.

5.2.2. Percentual de utilidade por grupo de avaliadores

O painel de utilidade média por grupo mostra três visões: Usuários comuns, Profissionais da Saúde e o agregado Geral. Além dos percentuais, são explicitados os números de pares considerados em cada cálculo.

Tabela 4 – Utilidade média por grupo (com pares considerados).

Grupo	**Pares considerados**	**Utilidade média**
Usuários comuns	350	77,7%
Profissionais da Saúde	315	81,9%
Geral	352	78,7%

Profissionais da Saúde classificam mais respostas como úteis ($+4{,}2$ pp acima dos usuários comuns). O agregado "Geral" fica entre os dois grupos, refletindo a composição e o peso relativo de cada estrato no conjunto avaliado.

5.2.2.1. Análise Crítica da Percepção de Utilidade das Avaliações

A análise da percepção de utilidade atribuída pelos avaliadores mostra que muitas respostas foram marcadas como úteis mesmo quando continham apenas informações mínimas. Em alguns casos, bastava a indicação de uma especialidade médica ou mesmo uma resposta extremamente curta para que a comunidade atribuísse o rótulo de utilidade.

Tabela 5 – Exemplos de respostas consideradas úteis pela comunidade

Pergunta	**Resposta**	**Label Comunidade**
Tenho cistos mamários, isso impede de fazer implante de silicone?	Em pricípio, não.	1 (Útil)
Pessoas idosas podem tomar ivermectina?	Sim.	1 (Útil)

Os exemplos evidenciam que, para os avaliadores, qualquer orientação prática, ainda que superficial, já é suficiente para ser considerada útil. Esse comportamento explica o alto índice de marcações positivas mesmo em respostas com baixo nível de detalhamento informacional.

5.2.2.2. Percepção do modelo

Em controvérsia com o comportamento observado nos rótulos da comunidade, o modelo supervisionado aprendeu — a partir dos padrões do conjunto — a que a mera assertiva curta, desacompanhada de evidências mínimas ou orientação prática, tende a ser classificada pelo modelo como não útil/pouco útil.

Os sinais mais associados à utilidade, segundo o modelo, incluem:

- **Contextualização do caso**: referência a sintomas, duração, fatores de risco ou condições que enquadram a resposta à pergunta específica;
- **Justificativa/explicação**: presença de conectores explicativos ("porque", "pois", "uma vez que") e raciocínio clínico elementar que sustente a conclusão;
- **Orientação acionável**: indicação de próximos passos (exames a considerar, quando procurar atendimento, condutas iniciais) em linguagem objetiva;
- **Consideração de alternativas e alertas**: menção a diagnósticos diferenciais, limites de segurança e sinais de alerta;

- **Alinhamento semântico com a pergunta**: alta sobreposição semântica/lexical entre pergunta e resposta, evitando genericidade;

Em síntese, enquanto a comunidade frequentemente atribuiu utilidade a respostas breves, o modelo privilegiou conteúdos que adicionam justificativa, contexto e instruções claras, por entender que tais elementos aumentam a utilidade prática para o usuário final.

5.2.3. Unanimidade dos julgamentos humanos

Para esta análise, foram considerados apenas os pares de pergunta–resposta que receberam pelo menos três avaliações humanas, de modo a garantir o impacto de julgamentos isolados.

O painel de Unanimidade Geral indica que, dentre 333 pares considerados, houve 163 pares unânimes, o que equivale a 48,9% de unanimidade.

Tabela 6 – Unanimidade (visão geral).

Métrica	**Pares considerados**	**Pares unânimes**	**Percentual**
Unanimidade (Geral)	333	163	48,9%

Aproximadamente metade dos itens recebeu a mesma decisão de todos os avaliadores (“útil” ou “não útil/ pouco útil”). A diferença entre *pares já avaliados* (352) e *pares considerados* para unanimidade (333) decorre justamente do critério de elegibilidade (mínimo de três votos válidos) e da exclusão de casos com empate/tie. A taxa de unanimidade moderada sugere variabilidade real na tarefa de julgamento de utilidade, o que ajuda a contextualizar a acurácia do modelo.

5.2.4. Distribuição por eixo temático

A distribuição normalizada por eixo temático aponta três grupos: Sintomas físicos, Tratamento/Medicação e Sintomas/Diagnósticos/Consultas. Além do volume por eixo, foi reportado o percentual médio de itens marcados como “útil”.

Tabela 7 – Distribuição por eixo temático (normalizado).

Eixo	**Itens**	**Avaliados**	**% “útil” (média)**
SF	52	52	76,6%
TM	105	105	77,5%
SDC	197	195	79,9%

O eixo **SDC** concentra mais itens e apresenta a maior taxa média de utilidade (79,9%), seguido de **TM** (77,5%) e **SF** (76,6%). As diferenças são mínimas, sugerindo

que perguntas ligadas a diferentes tipos tendem a ter a mesma percepção pelos avaliadores.

5.2.5. Confiabilidade entre avaliadores

Os coeficientes *Fleiss'* κ e *Krippendorff's* α foram calculados com a ferramenta *ReCal*[13]. A Tabela 8 resume os resultados por grupo de participantes.

Tabela 8 – Confiabilidade entre avaliadores por grupo.

Grupo	**Fleiss'** κ	**Krippendorff's** α
Usuários comuns	0,638	0,644
Saúde	0,475	0,484

Ao grupo de Usuários comuns, observou-se $\kappa = 0{,}638$ e $\alpha = 0{,}644$. Pela interpretação de Landis & Koch (1977), o valor de κ caracteriza concordância substancial ($0{,}61$–$0{,}80$). Já o α permanece abaixo do limiar de uso tentativo ($0{,}67 \leq \alpha < 0{,}80$) recomendado por Marzi, Balzano & Marchiori (2024), enquadrando-se como baixa confiabilidade ($\alpha < 0{,}67$), ainda que próximo ao limiar. Em relação ao grupo Saúde, foram obtidos $\kappa = 0{,}475$ e $\alpha = 0{,}484$. O κ indica concordância moderada ($0{,}41$–$0{,}60$) segundo Landis & Koch (1977). O α permanece em baixa confiabilidade por estar abaixo de $0{,}67$ (MARZI; BALZANO; MARCHIORI, 2024). Em ambas as métricas, o grupo de Usuários comuns apresentou valores superiores ao grupo da Saúde e o ordenamento entre grupos foi consistente entre κ e α, o que reforça o achado de maior consistência relativa no primeiro grupo.

6 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Os resultados apresentados neste trabalho, detalhados nos capítulos anteriores, fornecem uma visão profunda sobre a aplicação de técnicas de aprendizado de máquina para classificar a utilidade das respostas em plataformas de perguntas e respostas na área da saúde.

6.1. CONTRIBUIÇÕES DO ESTUDO

Este estudo contribui significativamente para o campo da saúde digital ao propor uma abordagem automática e eficiente para classificar a utilidade de respostas, complementando sistemas de avaliação comunitária existentes, como os votos e curtidas. Além disso, a construção do córpus SaudeBR-QA, a implementação do crawler para coleta de dados e a análise detalhada de modelos de aprendizado de máquina permitem que futuros estudos repitam e expandam os experimentos realizados, promovendo maior reprodutibilidade e transparência na pesquisa.

[13] <https://dfreelon.org/>

6.2. RESULTADOS E IMPLICAÇÕES PRÁTICAS

O modelo Random Forest, com o uso combinado de TF-IDF, *embeddings* e atributos adicionais, demonstrou a maior acurácia na classificação da utilidade das respostas. Esse desempenho foi validado através da análise comparativa com julgamentos de profissionais da saúde e do público geral, reforçando a validade da abordagem proposta. A precisão dos modelos evidencia que as características textuais, como o conteúdo e o formato das respostas, desempenham um papel fundamental na identificação de respostas úteis, oferecendo um método leve, eficiente e com alta capacidade de generalização.

Os resultados também destacam a importância de integrar múltiplos critérios na avaliação da utilidade das respostas. A dependência exclusiva de votos comunitários, como mostrado neste estudo, pode ser insuficiente em contextos onde a precisão clínica é crucial. Nesse sentido, os modelos de aprendizado de máquina podem servir como um primeiro filtro ágil, priorizando respostas úteis de forma mais assertiva.

6.3. LIMITAÇÕES E TRABALHOS FUTUROS

Embora os resultados deste estudo sejam promissores, há algumas limitações que devem ser consideradas. A análise foi conduzida em um único conjunto de dados específico para o domínio da saúde, o que exige validação em outros córpus para testar a generalização dos achados. Além disso, a dependência de características como TF-IDF e *embeddings* pode não capturar nuances mais complexas de respostas em contextos muito específicos ou com vocabulário técnico altamente especializado.

Como sugestões para trabalhos futuros, propõe-se:

- **Avaliação com novos córpus:** Validar os resultados com outros conjuntos de dados do domínio da saúde ou de outros domínios para verificar a performance do modelo.
- **Aperfeiçoamento do modelo:** Explorar a fusão de mais fontes de dados, como interações entre usuários e respostas, ou características de textos médicos específicos, para melhorar ainda mais a precisão da classificação.
- **Variação de *embeddings:*** Testar diferentes famílias e dimensões de embeddings em português (por exemplo, *Word2Vec* e *Glove* em 50, 100, 300 e 600 dimensões), bem como combinações com outras representações, avaliando o impacto nas métricas e no custo computacional.
- **Teste com redes neurais profundas:** Avaliar o desempenho de redes neurais profundas, como o *BERTimbau*, uma versão do *BERT* treinada especificamente

para o português, para a classificação de respostas. Essas redes e poderiam fornecer uma abordagem mais poderosa e precisa para a classificação de utilidade.

- **Ampliação da amostra de avaliadores:** Recrutar mais participantes em ambos os perfis (público geral e saúde) para aumentar o poder estatístico, reduzir a variância das estimativas e possibilitar conclusões mais robustas.

Em suma, a utilização de modelos de aprendizado de máquina para a classificação de utilidade de respostas em plataformas de saúde representa uma alternativa eficiente e complementária aos métodos tradicionais baseados em votos e curtidas. A abordagem proposta neste estudo, ao considerar características textuais e interativas, oferece uma solução com alta aplicabilidade para melhorar a qualidade das informações acessíveis ao público. Com a contínua evolução das técnicas de PLN e aprendizado de máquina, a integração dessas abordagens pode transformar a forma como as respostas médicas são classificadas, garantindo maior facilidade na tomada de decisões.

REFERÊNCIAS

AN, X.; LI, Y.; ZHOU, J. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities, and challenges. **Sensors**, MDPI, v. 23, n. 8, p. 4032, 2023. Citado na página 6.

CAO, C.; YANG, H. Explore how factors that contribute to online review helpfulness in automotive marketing. In: **Proceedings of the International Conference on Information Systems (ICIS)**. [S.l.: s.n.], 2022. Citado na página 8.

DIAZ, O.; NG, V. Modeling and prediction of online product review helpfulness: A survey. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)**. [S.l.: s.n.], 2018. p. 698–708. Citado na página 8.

Explosion. **pt_core_news_sm: Portuguese pipeline for spaCy**. 2025. Modelo pequeno para Português; Acesso em: 24 ago. 2025. Citado 2 vezes nas páginas 11 e 12.

GHOSE, A.; IPEIROTIS, P. G. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In: **Proceedings of the Ninth International Conference on Electronic Commerce**. [S.l.: s.n.], 2007. p. 303–310. Citado na página 7.

HONNIBAL, M. *et al.* **spaCy: Industrial-strength Natural Language Processing in Python**. 2020. Acesso em: 24 ago. 2025. Citado 2 vezes nas páginas 11 e 12.

HU, Z. *et al.* A deep learning approach for predicting the quality of online health expert question-answering services. **Journal of Biomedical Informatics**, v. 71, p. 241–253, 2017. Citado na página 8.

KIM, S. *et al.* Automatically assessing review helpfulness. In: **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2006. p. 423–430. Citado na página 7.

LAKATOS, E. M.; MARCONI, M. d. A. **Fundamentos de metodologia científica**. 8. ed. São Paulo: Atlas / Grupo GEN, 2017. ISBN 9788597010763. Citado na página 9.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977. Citado 2 vezes nas páginas 17 e 23.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan & Claypool, 2012. Citado na página 8.

LIU, J. *et al.* Low-quality product review detection in opinion summarization. In: **Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2007. p. 334–342. Citado na página 7.

MARZI, G.; BALZANO, M.; MARCHIORI, D. K-alpha calculator—krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. **MethodsX**, v. 12, p. 102545, 2024. Citado 2 vezes nas páginas 17 e 23.

OECD. **Health at a Glance 2023: OECD Indicators**: Digital health. Paris: OECD Publishing, 2023. Chapter: Digital health. Accessed: 2025-08-30. Citado na página 4.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, v. 2, n. 1–2, p. 1–135, 2008. Citado na página 7.

Python. **Python Programming Language – Official Website**. 2025. Acesso em: 26 ago. 2025. Disponível em: <https://www.python.org/>. Citado na página 10.

QIU, Y. *et al.* Predicting the quality of answers with less bias in online health question answering communities. **Information Processing & Management**, v. 59, n. 6, p. 103112, 2022. Citado na página 8.

ROY *et al.* Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. **CAAI Transactions on Intelligence Technology**, v. 8, n. 1, p. 95–117, 2023. Citado na página 4.

SANTIN, L. G. *et al.* O processamento de língua natural permite a classificação correta de laudos radiológicos em doenças benignas da vesícula biliar. **Radiologia Brasileira**, v. 57, p. e20230096, 2024. Citado na página 6.

SHAH, A. M.; MUHAMMAD, W.; LEE, K. Examining the determinants of patient perception of physician review helpfulness across different disease severities: A machine learning approach. **Computational Intelligence and Neuroscience**, Hindawi, 2022. Citado na página 8.

SHEIKHOLESLAMI, S. **Ablation Programming for Machine Learning**. Dissertação (Mestrado) — KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science (EECS), Stockholm, Sweden, 2019. Citado na página 19.

TIAN, Q.; ZHANG, P.; LI, B. Towards predicting the best answers in community-based question-answering services. In: **Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media**. [S.l.]: AAAI Press, 2013. p. 725–728. Citado na página 6.

VERMA *et al.* Powering covid-19 community q&a with curated side information. **arXiv preprint arXiv:2101.11556v1**, 2021. Acesso em: 01 abr. 2025. Citado na página 4.

WONGCHAISUWAT, P.; KLABJAN, D.; JONNALAGADDA, S. R. A semi-supervised learning approach to enhance health care community-based question answering: A case study in alcoholism. **arXiv preprint arXiv:1607.00706**, 2016. Citado na página 5.

XIA, L. The impacts of geographic and social influences on review helpfulness perceptions: A social contagion perspective. **Tourism Management**, v. 95, p. 104687, 2023. Citado na página 8.

YADAV, H.; YADAV, P. *et al.* Ai in healthcare: A survey on medical question answering system. **Studies in Engineering and Technology (SEEJPH)**, 2024. Citado 2 vezes nas páginas 5 e 8.

ZHANG, Z.; VARADARAJAN, B. Utility scoring of product reviews. In: **Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)**. New York, NY, USA: ACM, 2006. p. 51–57. Citado na página 7.

ZHENG, J. *et al.* Helpfulness prediction of online drug reviews. In: IEEE. **2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)**. [S.l.], 2021. p. 528–537. Citado na página 8.

ZORATTO *et al.* A study on influential features for predicting best answers in community question-answering forums. **Information**, v. 14, n. 9, p. 496, 2023. Citado na página 4.

AGRADECIMENTOS

Este trabalho — e a própria trajetória do ensino fundamental ao ensino superior — é fruto de um esforço coletivo, resultado de presença, cuidado e confiança de muitas pessoas ao meu redor. Nada aqui foi e será construído de forma isolada. A família e aos amigos ofereceram abrigo nos dias difíceis e comemoraram as pequenas e grandes vitórias junto comigo;

Agradeço à minha madrinha e mãe, Vana, minha maior inspiração e professora, por me ensinar todos os dias o significado do que me faz ser homem. Também, aos meus irmãos, Ana Vitória e Samuel, vocês são a base da minha história e a certeza de um porto seguro. Tudo que eu construir também é e será de vocês. Obrigado por tanto!

Agradeço aos meus pais, tios e tias, cuja presença em minha vida sempre foi paternal. Sou imensamente grato por todos os sacrifícios e preocupações que me trouxeram até este momento. O cuidado de vocês sempre esteve presente em tudo, desde um 'boa noite' a um doce de leite.

Agradeço ao meu orientador, amigo e melhor professor que eu poderia ter, Rogério, por cada correção e ensinamento repassado. Você foi essencial a conclusão deste trabalho e ao longo de todo o meu curso. Sou profundamente grato pela confiança, disponibilidade e apoio demonstrados durante todo o desenvolvimento desta pesquisa e em outros trabalhos relacionados.

Agradeço ao LIARA e aos amigos de sala, meu agradecimento gigante por cada dia que passei com vocês. Sou grato por cada trabalho, pesquisa, código, perrengue, café e partida de vôlei. Agradeço até pelo 'bullying' carinhoso que, no fim, nunca adiantou nada — e vocês sabem bem disso. Vocês fizeram parte dos meus melhores dias nestes últimos anos e levo cada um como parte da minha história universitária, que ficará marcada para sempre. Obrigada por me fazerem crescer. Vocês são os melhores e vou sentir muita saudade. Camilla, Avelar, Maykon, Jhulia, Gustavo, Anthony, Victor e Livya.

Aos meus amigos, companheiros insubstituíveis nesta jornada, minha eterna gratidão. Agradeço à Layla e à Amanda pela escuta, pela companhia e pelas risadas. À minha querida 'sister', Paulo Cesar, que anula qualquer distância vinda de São Paulo e permanece como um dos maiores presentes da minha vida. Sou grato ao meu primo, amigo e orientador nas horas vagas, Denílson, por toda a ajuda e pelos conselhos. Agradeço também ao meu amigo David, por dividir o mesmo cérebro que eu, e aos melhores grupos existentes, Pluck e Fofoquinha.