The Superficiality Bias: Community Votes and Answer Utility in Portuguese Health Question Answering

Carlos Henrique Santos Barros
Federal University of Piauí (UFPI)
carlos.barros.cb@ufpi.edu.br

Gustavo Figueredo Rodrigues de Sousa
University of São Paulo (USP)
gustavofigueredo@usp.br

Rogério Figueredo de Sousa
Federal Institute of Piauí (IFPI)
rogerio.sousa@ifpi.edu.br

Abstract

Supervised models trained on community-labeled data have shown promise in Health Question Answering (HQA), but relying on "likes" as a proxy for clinical usefulness remains controversial. This work investigates the alignment between automated predictions and human perception in Portuguese HQA. Using a subset of the SaudeBR-QA corpus, we compare a Random Forest classifier against a controlled evaluation conducted by laypeople and healthcare professionals. Our results reveal a recurring divergence that we term **Superficiality Bias**: human evaluators frequently validate very brief answers, whereas the classifier often labels these cases as non-useful under its learned criteria. Rather than indicating that the model is inherently more clinically accurate, this pattern suggests a misalignment between community feedback and feature-driven utility judgments. We argue that crowd-based labels in medical domains should be treated cautiously and complemented with more rigorous annotation protocols.

1 Introduction

Access to online medical information has grown exponentially, fundamentally reshaping how individuals seek clarification on symptoms, diagnoses, and treatments (OECD, 2023). In this landscape, Health Question-Answering (HQA) platforms play a central role by enabling direct interaction between patients and professionals, relying heavily on "crowd wisdom" expressed through votes and likes to filter content and signal relevance (Verma et al., 2021). In Brazil, platforms such as *CatálogoMed* exemplify this ecosystem, generating vast amounts of User-Generated Content (UGC). However, the clinical reliability of these social signals remains a critical challenge. Previous studies warn that the popularity of an answer does not guarantee its medical accuracy (Zoratto et al., 2023), and relying solely on community feedback may favor accessible but incomplete information over technically accurate advice.

To address the lack of resources for Portuguese, the *SaudeBR-QA* corpus was recently released, providing the dataset for this domain (Barros et al., 2025). Initial benchmarks demonstrated that supervised models trained on these community labels can achieve high computational performance, with accuracy rates reaching nearly 79% (Barros et al., 2025). Yet, high metric performance against noisy labels does not necessarily imply high clinical utility. Little has been discussed about the gap between what these models learn from community votes and what constitutes a truly useful medical response.

This work shifts the focus from resource construction to critical validation. We investigate whether models trained on community likes learn signals of answer utility or simply reproduce patterns associated with popularity and brevity. By contrasting the decisions of a supervised classifier with a controlled human evaluation involving both laypeople and healthcare professionals, we identify a recurrent divergence that we term **Superficiality Bias**. Our findings suggest that community feedback may favor accessible and concise responses, whereas the classifier tends to penalize very short answers under its learned feature representation. Rather than treating either side as a definitive gold standard, we interpret this result as evidence that utility in HQA is operationalized differently by community feedback, model predictions, and controlled human judgment.

2 Related Work

The assessment of answer quality in Community Question Answering (CQA) has been extensively studied, often relying on metadata such as upvotes, user reputation, and answer length as proxies for utility. Wongchaisuwat et al. (2016) demonstrated that combining textual features with community

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1074–1078

signals significantly improves classification performance in health forums. Similarly, Tian et al. (2013) found that answer length and early arrival time are strong predictors of the "best answer" selection in general domains.

However, in the medical domain, the reliability of crowd wisdom is contested. Zoratto et al. (2023) warned that popularity metrics often reflect user engagement rather than clinical accuracy. Roy et al. (2023) further argued that hybrid approaches, integrating deep learning representations (e.g., BioBERT) with expert validation, are necessary to mitigate the noise inherent in community labels.

Regarding Portuguese resources, Ferreira et al. (2023) introduced *SemClinBr* for clinical semantic similarity, but datasets focused specifically on consumer health questions remain scarce. Recently, Barros et al. (2025) released *SaudeBR-QA*, the HQA corpus for Brazilian Portuguese. While their work focused on resource construction and baseline benchmarks, this study extends that effort by critically auditing the alignment between those automated baselines and controlled human judgment in a healthcare-oriented setting.

3 Methodology

This study adopts a mixed-method approach, combining supervised classification results with a controlled human evaluation campaign. The goal is not only to examine model performance, but also to analyze the degree of alignment between community-derived labels, model predictions, and human judgments in the health domain.

3.1 Dataset and Automated Classifier

We used the *SaudeBR-QA* benchmark as the data basis for this study. Rather than performing a new model comparison, we adopted the **Random Forest** classifier because it had previously shown the best balance between accuracy and F1-score for this dataset in the baseline benchmark (Barros et al., 2025). The classifier represents each question–answer pair through a hybrid feature space combining TF–IDF vectors, FastText-based sentence representations, and compact structural attributes, including answer length and question–answer similarity. In the present study, this previously validated classifier serves as an *algorithmic reference point* for comparison with controlled human judgments.

3.2 Sampling and Annotation Procedure

The human evaluation was conducted on a pool of 354 question–answer pairs drawn from the Random Forest test split. These pairs were balanced by class and organized into three thematic axes: **Symptoms, Diagnosis, and Consultations (SDC)**, **Treatment and Medication (TM)**, and **Physical Symptoms (SF)**. Each participant evaluated 118 items, randomly sampled from this pool, so that different portions of the test set could be judged across annotators.

Annotations were collected through a custom web application developed in Streamlit. Each item presented a question together with a candidate answer, and annotators were asked to provide a binary judgment indicating whether the answer was *Useful* or *Not Useful / Little Useful*. The interface did not reveal either the original community feedback or the model prediction. Before starting the task, participants selected their evaluator profile and received general instructions about the assessment procedure.

3.3 Participants and Consensus Formation

A total of 16 participants took part in the study: 10 lay users and 6 healthcare professionals. The latter group included students and professionals from Nursing, Nutrition, and Psychology. This composition provides a more domain-aware perspective than lay evaluation alone, but it does not constitute a topic-specialized physician benchmark for all assessed question types. Accordingly, the professional judgments reported in this study should be interpreted as a qualified healthcare-oriented perspective rather than as a definitive clinical gold standard.

For the model-versus-human comparison, human judgments were aggregated into a consensus label. A pair was assigned the label *Useful* when at least 50% of the evaluators who judged that item marked it as useful; otherwise, it was treated as *Not Useful*. Model accuracy against human judgment was then computed with respect to this aggregate label. In a complementary analysis, unanimity was examined separately for items with at least three valid human evaluations.

Inter-rater reliability was assessed using Fleiss' Kappa and Krippendorff's Alpha, allowing us to quantify the degree of agreement within each evaluator group and to contextualize the relationship between human judgment and classifier predictions.

4 Results and Discussion

Our experiments reveal a clear divergence between model predictions and the judgments collected in the controlled annotation study. Rather than assuming that any single signal fully captures the true notion of usefulness, we interpret the results as evidence that answer utility is operationalized differently across three layers: the original community-derived labels used in the benchmark, the predictions of the Random Forest classifier trained on those labels, and the controlled human judgments obtained in our evaluation protocol.

4.1 Agreement and Subjectivity

When compared against the aggregate human consensus, the classifier achieved an accuracy of 57.4%. This value should not be interpreted in isolation as either satisfactory or unsatisfactory, because the task itself proved to be substantially subjective. In other words, the disagreement between model predictions and human judgments must be read in the context of the limited agreement observed among human evaluators themselves.

As shown in Table 1, lay users achieved higher agreement, whereas healthcare professionals reached only moderate agreement ($\kappa = 0.475$; $\alpha = 0.484$). In addition, total unanimity was observed in fewer than half of the analyzed cases. These findings indicate that answer usefulness in health-related question answering is difficult to operationalize consistently, even when assessed by evaluators with healthcare backgrounds. Therefore, the observed 57.4% agreement between the classifier and the human consensus reflects not only the behavior of the model, but also the inherent ambiguity of the evaluation target.

Table 1: Inter-rater reliability by participant group.

Group	Fleiss' κ	Krippendorff's α
Common Users	0.638	0.644
Healthcare Professionals	0.475	0.484

This result is especially important because it weakens any simplistic interpretation of the task as a purely objective classification problem. If even healthcare professionals do not converge strongly on a single notion of usefulness, then differences between the model and human evaluators should be interpreted more cautiously as divergences between operational criteria, rather than as straightforward evidence of model failure or superiority.

4.2 The Superficiality Bias

The most salient qualitative pattern concerns the treatment of very brief answers. As illustrated in Table 2, human evaluators frequently judged short and direct answers as useful, even when those answers provided little elaboration. In contrast, the classifier often labeled these same cases as *Not Useful* under its learned feature representation.

Table 2: The Superficiality Bias: examples of disagreement between human judgments and model predictions.

Question (Snippet)	Answer	Human	Model
Can elderly people take ivermectin?	Yes.	✓	✗
I have cysts, does this prevent implants?	No.	✓	✗

Note: ✓= Useful; ✗= Not Useful.

We refer to this recurrent pattern as **Superficiality Bias**. In practical terms, the bias emerges when highly concise answers receive positive human judgments despite offering limited contextualization, justification, or actionable guidance. This does not mean that such answers are necessarily clinically unsafe, nor does it imply that the classifier is inherently more capable of identifying clinically useful responses. Instead, the pattern suggests that human evaluators, especially in an online-help context, may sometimes value brevity and direct reassurance, while the classifier tends to assign lower utility to answers with reduced informational density.

This point deserves careful interpretation. Because the Random Forest model was trained on community-derived labels and includes structural attributes such as answer length and question–answer similarity, its tendency to penalize very short answers may partly reflect a learned length-sensitive bias. For this reason, the observed pattern should not be presented as proof that the model correctly detects a lack of substance. A more defensible interpretation is that the experiment reveals a *misalignment* between community-oriented or human-perceived usefulness and the utility profile learned by the classifier from its feature space.

Seen from this perspective, the contribution of the present study is not to establish that one side is definitively correct, but to make this discrepancy explicit. In medical question answering, this is especially relevant because systems trained directly on community-derived signals may inherit and reproduce preferences for accessibility or reas-

surance that do not necessarily coincide with richer informational content.

4.3 Thematic Variations

We also observed that perceived utility varies across thematic categories, as summarized in Table 3. Questions in *Symptoms and Diagnosis* (SDC) obtained the highest overall utility rate, followed by *Treatment and Medication* (TM), while *Physical Symptoms* (SF) obtained the lowest percentage.

Table 3: Human perception of utility across thematic categories.

Code	Category	Utility (%)
SDC	Symptoms & Diagnosis	**79.9**
TM	Treatment & Medication	77.5
SF	Physical Symptoms	76.6

Although these differences are not large, they suggest that the notion of usefulness is not uniformly distributed across medical subdomains. One possible interpretation is that more general or explanatory questions, such as those in Symptoms and Diagnosis, allow evaluators to accept broader forms of guidance as useful, whereas questions involving more concrete bodily manifestations may create stronger expectations for specificity and detail. In this sense, usefulness may depend not only on answer content, but also on the informational demands associated with each topic.

At the same time, these thematic results should be interpreted with caution. In the present version of the analysis, the percentages are aggregated across all human evaluators. Therefore, they do not yet distinguish how lay users and healthcare professionals may have behaved differently within each thematic axis. This limits the granularity of the conclusions that can be drawn from the thematic comparison and should be treated as an important constraint of the current analysis.

4.4 Implications for Utility Modeling in HQA

Taken together, the results suggest that community-derived labels should not be treated as an unproblematic proxy for usefulness in medical HQA. On the one hand, the classifier reproduces patterns that are effective for predicting the original benchmark labels. On the other hand, the controlled human evaluation reveals that such labels do not fully align with later judgments made by either lay users or healthcare professionals in a blind setting.

This finding has two implications. First, it indicates that the utility signal available in real-world platforms is noisy and potentially shaped by social preferences that differ from richer informational criteria. Second, it suggests that benchmarking only against community-derived labels may overestimate the adequacy of a system for health-related settings, where the consequences of superficial but positively perceived answers may be non-trivial. For this reason, future HQA evaluation pipelines should rely on more explicit annotation protocols and should treat disagreement itself as part of the phenomenon to be modeled, rather than as mere annotation error.

4.5 Threats to Validity

This study has limitations. First, the analysis relies on a single previously validated classifier, namely Random Forest, which restricts claims about the generality of the observed pattern across model families. Without comparisons against Transformer-based or other modern architectures, it remains unclear to what extent the observed divergence is a broader property of the task or a behavior associated with this specific model configuration.

Second, because the classifier includes structural features such as answer length, its tendency to penalize short answers may partly reflect a learned length-sensitive artifact. Accordingly, the present results should not be interpreted as demonstrating that the model is inherently more capable of identifying clinically useful answers than human evaluators.

Third, the healthcare professional group was heterogeneous, including participants from Nursing, Nutrition, and Psychology. Although this provides a more domain-aware perspective than lay evaluation alone, these backgrounds do not necessarily align equally with all thematic axes, especially questions involving medication use, diagnosis, and treatment decisions. Therefore, the professional judgments reported here should be interpreted as a qualified healthcare perspective rather than as a definitive specialist gold standard.

Finally, the moderate agreement observed among healthcare professionals indicates that no definitive gold standard for answer utility was established in this study. For this reason, the contribution of the paper is best understood as an analysis of *misalignment* between evaluation signals—community-derived labels, controlled human judgments, and feature-based model predictions—

rather than as the establishment of a single correct notion of usefulness.

5 Conclusion

This study examined the reliability of community-derived labels as a proxy for answer utility in Portuguese Health Question Answering. By comparing a previously validated Random Forest classifier with controlled human judgments from lay users and healthcare professionals, we identified a consistent divergence between different signals of usefulness. In particular, very brief answers were often positively evaluated by human participants, while the classifier frequently labeled these same cases as non-useful under its learned feature representation.

Rather than demonstrating that the model is inherently better at identifying clinically useful answers, these findings indicate that utility is operationalized differently across community-derived labels, controlled human judgments, and feature-based model predictions. This divergence is especially relevant in health-related settings, where concise and reassuring responses may be perceived as helpful even when they provide limited contextualization or actionable guidance. In this sense, the main contribution of this work is not to establish a single correct notion of usefulness, but to make explicit the misalignment among the evaluation signals commonly used in HQA research.

The results also reinforce that usefulness in medical question answering is not a fully objective construct. The moderate agreement observed among healthcare professionals suggests that even qualified evaluators do not converge on a single stable interpretation of utility across all cases. For this reason, community feedback should be treated cautiously when used as direct ground truth for model development and evaluation in medical NLP pipelines.

Future work should move toward more rigorous annotation protocols, including clearer task guidelines, topic-specialized evaluators, multi-annotator adjudication, and comparisons with additional model families. More broadly, disagreement itself should be treated as part of the phenomenon to be modeled, rather than as mere annotation noise. Under this view, improving Portuguese HQA systems depends not only on stronger classifiers, but also on more reliable and clinically informed evaluation frameworks.

References

Carlos Henrique S. Barros, Gustavo F. R. de Sousa, and Rogério F. de Sousa. 2025. Saudebr-qa: Um córpus de perguntas e respostas para o domínio da saúde em português brasileiro. In *Proceedings of the Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI)*. Brazilian Computer Society (SBC). To appear.

Debora Ferreira, Fernando Alva-Manchego, Saturnino Luz, and 1 others. 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.

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

Roy, Pradeep, Kumar, Saumya, Sunil, Singh, Jyoti, and Prakash. 2023. Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, 8(1):95–117.

Qiongjie Tian, Peng Zhang, and Baoxin Li. 2013. Towards predicting the best answers in community-based question-answering services. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 725–728. AAAI Press.

Verma, Mayank, Thadani Khyati, and Shubham Mishra. 2021. Powering covid-19 community q&a with curated side information. *arXiv preprint arXiv:2101.11556v1*.

Papis Wongchaisuwat, Diego Klabjan, and Siddhartha Reddy Jonnalagadda. 2016. A semi-supervised learning approach to enhance health care community–based question answering: A case study in alcoholism. *JMIR medical informatics*, 4(3):e5490.

Zoratto, Victoria, Godoy, Daniela, and Gabriela N. Aranda. 2023. A study on influential features for predicting best answers in community question-answering forums. *Information*, 14(9):496.