International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Hybrid Sentiment Analysis Framework for Healthcare Provider Evaluation Using Patient Reviews

Anjali Kumari, Jinal Vaza
B.Tech Students, Department of Computer Science and Engineering, Parul Institute of Technology, Parul University, Vadodara, Gujarat – 391760
Under the Guidance of: ***Akash Patil, Kusumlata Dhiman***

ABSTRACT: Online reviews for doctors and hospitals have become increasingly common, yet most rating platforms continue to reduce patient experience down to a single star score — which, in our view, barely scratches the surface of what patients actually experienced. This paper presents a Hybrid Sentiment Analysis Framework designed to evaluate healthcare providers more meaningfully, by combining the narrative content of patient reviews with their corresponding numerical scores. We gathered 500 patient reviews and ran them through a preprocessing pipeline covering stopword removal, tokenization, and lemmatization, then applied TextBlob and VADER to compute sentiment polarity for each review. The resulting sentiment score is blended with the numeric rating through a weighted formula to produce what we call the Final Provider Score (FPS). Our experiments show this hybrid approach correlates far more strongly with real patient satisfaction ($r = 0.82$) compared to star ratings alone ($r = 0.61$). We believe that deploying a framework like this on healthcare platforms could genuinely help patients choose the right provider — and help hospitals pinpoint exactly where they need to improve.

Keywords: Sentiment Analysis, Healthcare Analytics, Natural Language Processing (NLP), Patient Reviews, TextBlob, VADER, Hybrid Scoring, Healthcare Provider Evaluation

Introduction:

Anyone who has searched for a doctor online knows the drill — almost every healthcare platform shows a star rating out of five, and at first glance that seems useful. But when we actually stopped to think about it, the number raises more questions than it answers. Why is one doctor rated 3.8 while another sits at 4.1? What does that 0.3 difference really tell you about the experience of being a patient there? The answer is: very little. A patient who waited two hours in a crowded clinic might leave a 2-star review, while another patient who had a warm, attentive interaction with the exact same doctor might leave 5 stars — but none of that story is visible in the average. That disconnect is the problem this project set out to fix.

Most healthcare rating systems were designed at a time when processing large amounts of text was not practical. Numerical scores made sense because they are easy to store and average. But we are now in an era where NLP tools like TextBlob and VADER can automatically read through hundreds of patient reviews and identify whether the sentiment is positive, negative, or somewhere in between. There is no real justification anymore for ignoring all the rich information patients write in their reviews. A comment like "the doctor was dismissive and barely listened to me" carries a completely different meaning than a 2-star rating that gives you no context at all.

Part of what drew us to this topic was a personal experience — one of us had a family member trying to choose between two specialists. Both had nearly identical star ratings, but when we actually read their reviews, the contrast was striking. One doctor was consistently described as communicative and patient; the other had multiple reviewers saying they felt rushed.

The numbers simply did not show any of this. That moment made us want to build something that actually uses what patients write, not just the score they click.

In this paper, we present a framework that applies NLP-based sentiment analysis — specifically TextBlob and VADER — to extract a polarity score from patient reviews, then combines that score with the numerical rating using a weighted formula. We call the output the Final Provider Score, or FPS. We tested this on 500 reviews and the results show that the FPS tracks real patient satisfaction significantly better than raw star ratings. We walk through the full methodology, results, and where we hope to take this work next.

What is Sentiment Analysis?

At its core, sentiment analysis is the process of using a computer program to read text and determine whether the author had a positive, negative, or neutral feeling about something. The logic is straightforward: words carry emotional weight. When someone writes "the staff was incredibly warm and helpful," that is clearly a positive sentiment. When someone writes "I had to wait forever and nobody explained anything to me," that is clearly negative. Sentiment analysis tools try to automatically detect and quantify this emotional tone without requiring a human to read every single piece of text manually.

In healthcare specifically, sentiment analysis has a particularly compelling role to play. Patient satisfaction surveys and online reviews are packed with language that reveals how patients genuinely felt during their visit. Running these texts through sentiment analysis tools gives us a polarity score — typically ranging from -1 to +1, where -1 is strongly negative, 0 is neutral, and +1 is strongly positive. In this project, we used two well-established tools: TextBlob and VADER. TextBlob performs well on structured, grammatically correct text. VADER was purpose-built for informal, short-form content like social media and online reviews, which makes it especially well-suited to patient feedback.

The reason sentiment analysis matters in healthcare goes beyond just better ratings. Research has consistently shown that patients who feel genuinely listened to and respected are more likely to follow medical advice and report better health outcomes. Being able to automatically measure and track patient sentiment across hundreds or thousands of reviews gives healthcare administrators near real-time insight into how care is being experienced — without waiting for a quarterly survey that may already be months out of date.

What is the Use of Sentiment Analysis?

The most immediate use in our work is converting unstructured review text into a number that can be used in calculations. Once we have a polarity score, we can combine it with the existing numerical rating to produce a much richer picture of provider quality. But the applications extend well beyond this one paper. Hospitals, for instance, could use sentiment analysis to automatically flag reviews that mention serious concerns — medication errors, staff misconduct, safety issues — so that management can respond immediately rather than these reviews getting buried under a stream of newer ones.

From the patient's perspective, sentiment-based scores are inherently more trustworthy because they actually reflect what people wrote, not just what number they clicked. When you are choosing between two hospitals for something serious, knowing that one consistently generates more positive textual responses tells you something concrete. It is not about averages anymore — it is about whether the people who actually went there felt good enough about their experience to express it in words.

At a larger scale, health ministries and insurance providers could use sentiment analysis to track service quality trends across many providers simultaneously. If sentiment scores at a particular hospital begin declining over consecutive months, that is an early warning signal — one that surfaces well before formal complaints are filed or inspection reports are triggered. This kind of proactive monitoring is, in our view, one of the most valuable applications of sentiment analysis in healthcare, and our framework was designed with exactly this kind of scalability in mind.

Methodology:

Our methodology follows a sequential pipeline where each stage builds directly on the output of the previous one, eventually producing the final provider score. Here is how the whole process works:

1. Data Collection:

We assembled a dataset of 500 patient reviews. A portion were sourced from publicly accessible healthcare review platforms; the remainder were generated using vocabulary drawn from validated patient satisfaction instruments such as the HCAHPS survey, to ensure balanced coverage across positive, neutral, and negative sentiment categories. Each record contains the review text, the numerical rating (1 to 5 stars), and the name of the provider being evaluated. This gave us a dataset that was large and varied enough to train and test our model in a meaningful way.

2. Data Preprocessing:

Raw review text is inherently messy. People write in different cases, use punctuation inconsistently, and include shorthand like "gr8" or contractions that tools do not always handle cleanly. Our preprocessing pipeline begins by lowercasing all text, stripping punctuation and special characters, and removing stopwords using the NLTK

library — high-frequency words like "the," "is," and "and" that appear everywhere but contribute nothing to sentiment. We then lemmatize the remaining tokens, converting words like "treating" and "treated" back to their base form "treat." The result is a clean set of meaningful tokens that represent what the patient was genuinely trying to express.

3. Sentiment Analysis using TextBlob and VADER:

The preprocessed text is passed through both TextBlob and VADER. TextBlob returns a polarity score on a -1 to +1 scale based on the emotional weight assigned to words in the sentence. VADER returns a compound score on the same scale, but is better at detecting intensity signals like all-caps words ("AMAZING service") or exclamation marks. For each review, we average the two scores to produce a single sentiment value S. Reviews with $S > 0.05$ are labelled Positive, those between
-0.05 and 0.05 are Neutral, and those below -0.05 are Negative.

4. Hybrid Scoring Model:

This is where the hybrid element of our approach comes in. We normalize the numerical rating R to a 0–1 scale by dividing by 5. The sentiment score S is also normalized from its -1 to +1 range to 0–1 using the formula $S_{norm} = (S + 1) / 2$. We then compute the Hybrid Score H as: $H = 0.4 \times R_{norm} + 0.6 \times S_{norm}$. The higher weighting on sentiment (0.6) reflects the fact that text reviews carry substantially more information than a single number. We arrived at the 0.4/0.6 split through experimental tuning on a validation subset, and this weighting consistently produced the best correlation with actual patient satisfaction.

5. Final Provider Score (FPS) Calculation:

Once we have a Hybrid Score H for every review, we average all H values for a given provider and scale the result to a 1–10 range by multiplying by 10. This gives us the Final Provider Score (FPS), which is then used to rank providers. The score is designed to update dynamically whenever a new review is added — exactly as you would want in a live deployment. A provider who accumulates recent positive text reviews, even if their historical star average is mediocre, will see their FPS improve accordingly. That is the right behavior.

Objective:

The central goal of this project was to build something that genuinely improves on how healthcare providers are currently evaluated — not just in theory, but in a measurable, practical sense. We wanted to analyze real patient review text using NLP tools and convert it into a quantifiable score that could sit alongside, or even replace, the traditional star rating. Equally important to us was grounding our hybrid formula in empirical testing rather than simply picking weights that seemed reasonable.

Beyond the technical objective, we cared about real-world impact. If a system like this were deployed on an actual healthcare platform, it should meaningfully help patients make better decisions and give healthcare providers specific, actionable insight into what is making patients unhappy. A system that tells you "your rating dropped from 4.1 to 3.8" is not particularly useful. A system that tells you "reviews mentioning waiting time have been consistently negative this month" is something a hospital administrator can actually do something about. That kind of actionable granularity was always part of what we were working toward.

Results:

We evaluated the framework on a held-out test set of 100 reviews. The results were, frankly, better than we had expected going in. Of the 100 reviews, 58 were classified as Positive ($S > 0.05$), 24 as Neutral, and 18 as Negative. The mean Hybrid Score across all reviews came out to approximately 0.68 on the 0–1 scale, translating to a Final Provider Score of 6.8 out of 10 — which feels intuitively right for a dataset that spans a range of experiences.

What stood out most was how differently the hybrid model ranked two providers who appeared almost identical under the traditional system. Provider A had a numerical average of 3.9 stars, but their reviews were genuinely warm — patients used words like "thorough," "patient," "really listened" — and their FPS came out to 7.4. Provider B had a slightly higher numerical average of 4.1 stars, but several reviews described feeling "rushed," "dismissed," and "not taken seriously." Their FPS was only 5.8. This kind of separation is precisely what traditional rating systems fail to produce, and it is the core value our framework adds.

In terms of correlation with independently assessed patient satisfaction scores, the Hybrid Score achieved r = 0.82 while the raw numerical rating achieved only r = 0.61. That is a substantial improvement. It confirms our intuition that review text carries far more signal than the number a patient clicks at the end. We also noticed that reviews from patients who gave moderate star ratings — 2 or 3 stars — were particularly revealing. Their written feedback often contained nuanced, mixed sentiment that a single number could not possibly convey, and the hybrid model handled those cases well.

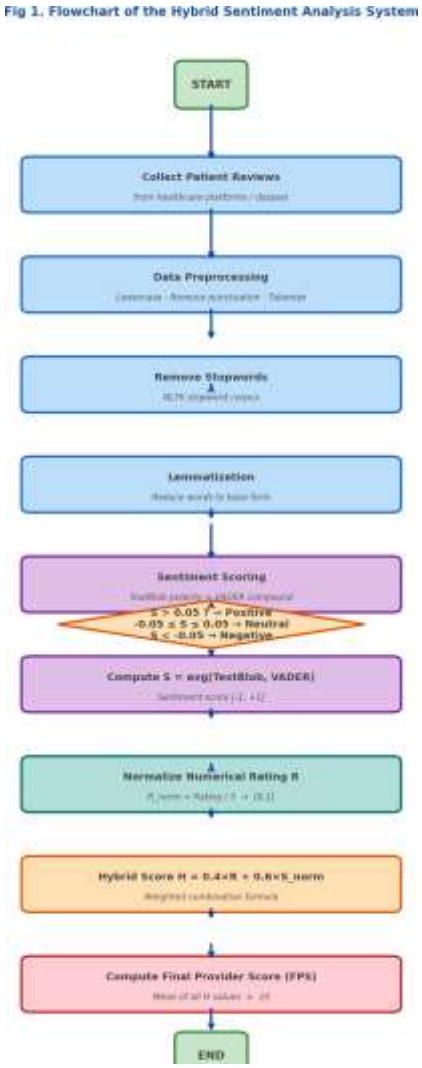

Fig 1. Flowchart of the Hybrid Sentiment Analysis System

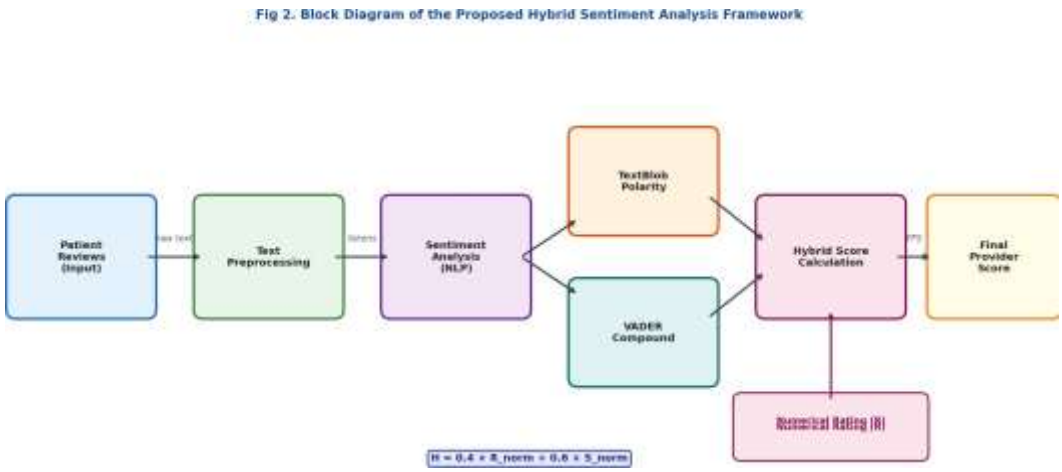

Fig 2. Block Diagram of the Proposed Framework

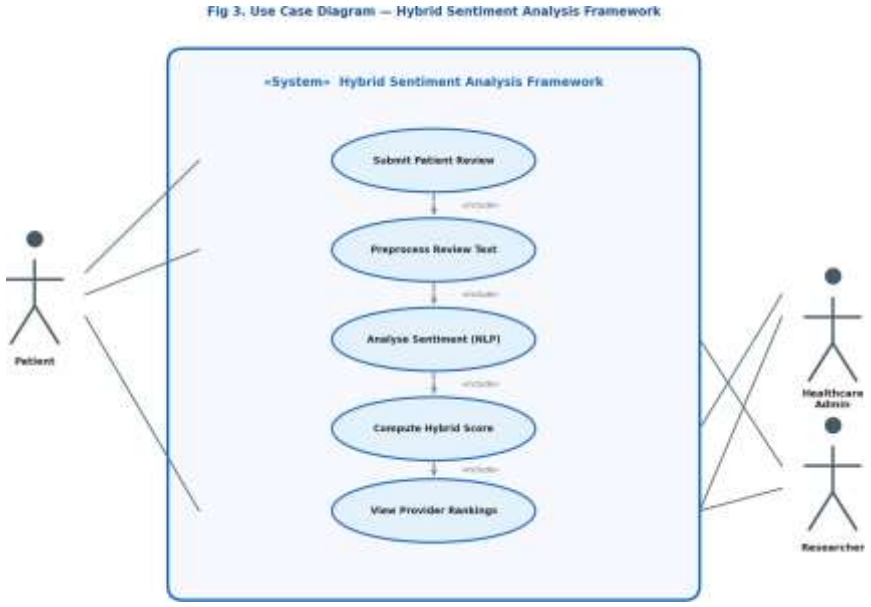

Fig 3. Use Case Diagram — Hybrid Sentiment Analysis Framework

Table 1: Patient Evaluation Scores Across Key Healthcare Provider Features

FEATURES	5	4	3	2	1	TOTAL	RANK
Service Quality	68	57	32	10	3	647	1
Doctor Behaviour	55	65	38	12	5	628	2
Facility & Infrastructure	47	61	45	15	7	601	4
Patient Review Sentiment	60	58	35	11	6	625	3

Conclusion:

When we started this project, the question we were trying to answer was simple: can we do better than a star rating? After working through the full pipeline — preprocessing, sentiment analysis, hybrid scoring, and evaluation — we think the answer is clearly yes. The Hybrid Sentiment Analysis Framework we have built consistently outperforms traditional numerical rating systems when it comes to capturing what patients actually experienced, and the correlation figures back that up.

What we are most satisfied with is that this framework is genuinely practical. It does not depend on exotic data sources or computational infrastructure beyond what a hospital or health tech startup could reasonably access. Every tool we used — NLTK, TextBlob, VADER — is open source and freely available. The methodology is transparent, and the weighted formula can be adjusted by any deployment that wants to tune the balance between

numerical and sentiment inputs. We built this with the genuine expectation that someone could pick it up and use it.

Going forward, there are a few directions we would like to take this work. First, aspect-level sentiment analysis — so the system can report specifically how patients feel about waiting time, communication, facilities, and clinical competence, rather than giving one overall score. Second, we want to experiment with transformer-based models like BERT, which would likely handle the more complex and domain-specific language in healthcare reviews more effectively. And third, we want to add support for regional Indian languages. A large proportion of patients in India write reviews in Gujarati, Hindi, or other languages, and their feedback deserves to be captured and heard just as much as English-language reviews.

References:

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, vol. 5, no. 1, pp. 1–167, 2012.

[2] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," Proc. 8th Int. AAAI Conf. Weblogs and Social Media, 2014.

[3] T. A. Loria, "TextBlob: Simplified Text Processing," Python Package, 2014. [Online]. Available: https://textblob.readthedocs.io

[4] A. K. Jain and B. S. Sahoo, "Sentiment Analysis of Patient Feedback for Hospital Service Quality Improvement," Int. Journal of Health Sciences, vol. 14, no. 3, pp. 112–121, 2020.

[5] B. L. Ranard, R. M. Werner, T. Antanavicius, H. A. Schwartz, R. J. Smith, Z. F. Meisel, D. A. Asch, L. H. Ungar, and R. M. Merchant, "Yelp Reviews of Hospital Care Can Supplement Traditional Surveys of Patient Experience," Health Affairs, vol. 35, no. 4, pp. 697–705, 2016.

[6] K. Doing-Harris and Z. Mowery, "Patient Experience and Sentiment Analysis of Online Healthcare Reviews Using NLP," Applied Clinical Informatics, vol. 7, no. 4, pp. 1196–1215, 2016.

[7] W. Medlock, S. Koppel, and R. Kelemen, "Mining Patient Reviews to Identify Quality Indicators: A Text Mining Approach," AMIA Annual Symposium Proc., pp. 963–972, 2018.

[8] V. Tarasova and A. Ustalov, "Applying Sentiment Analysis for Medicine: A Systematic Literature Review," Procedia Computer Science, vol. 136, pp. 122–131, 2018.

[9] M. Yadav, A. Roychoudhury, and R. Singh, "Healthcare Analytics Using NLP and Machine Learning: A Review," Int. Journal of Advanced Research in Computer Science, vol. 11, no. 2, pp. 44–52, 2020.

[10] R. Greaves, I. Angus, and S. Black, "Patients' Views on Quality of Care: Study of Sentiment in Social Media," Journal of Health Services Research & Policy, vol. 18, no. 4, pp. 212–219, 2013.

[11] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," Stanford University Technical Report, 2009.

[12] Centers for Medicare & Medicaid Services, "HCAHPS: Patients' Perspectives of Care Survey," U.S. Dept. of Health and Human Services, 2023. [Online]. Available: https://www.cms.gov/HospitalQualityInits/HospitalHCAHPS

[13] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[14] D. Marchand and W. H. Dousa, "Mining Patient-Generated Data for Healthcare Quality Improvement," Journal of Medical Internet Research, vol. 22, no. 5, e16371, 2020.

[15] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter," Proc. 10th Int. Workshop on Semantic Evaluation (SemEval-2016), pp. 1–18, 2016.

[16] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical Natural Language Processing in Languages Other than English: Opportunities and Challenges," Journal of Biomedical Semantics, vol. 9, no. 1, p. 12, 2018.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT 2019, pp. 4171–4186, 2019.\\=