California Hospital Performance Analysis

Kinaz Abdulsalam, Hajra Ibrahim, Daniel Tehrani

George Washington University

2242 Natural Language Processing for Healthcare

Muhammad Rahman

December 9, 2025

Abstract

Hospital performance ratings play an essential role in shaping patient choices, guiding policy decisions, and evaluating the overall quality of care delivered by hospitals. However, these ratings often rely heavily on structured variables such as complication rates and case volumes, which may overlook important contextual details found in written performance descriptions. This study investigates whether narrative hospital text can contribute meaningful insight into quality variation across California. Using TF-IDF, we transformed narrative description into quantitative features and examined their relationship to rating categories. Two logistic regression models were developed: one using structured clinical variables alone and another combining structured and text-derived variables. The combined model demonstrated improved recall for minority classes, suggesting that narrative information adds important nuance beyond traditional metrics. Overall, the findings support the idea that integrating structured and unstructured data can offer a more complete understanding of hospital quality across the states.

Background

Hospital performance ratings play an increasingly important role in healthcare by influencing patient choices, shaping policy decisions, and helping hospitals demonstrate accountability for the care they provide. These ratings are typically based on structured clinical variables such as complication rates, procedure counts, and risk-adjusted outcomes. While these metrics are essential, they do not always capture the underlying circumstances that contribute to hospital performance. Many hospitals provide written descriptions that explain factors like patient complexity, staffing limitations, and ongoing improvement efforts, which are all contextual information that cannot be fully represented through numbers alone. Understanding whether these narratives contain meaningful patterns may help provide a more complete picture of hospital quality and variation across California.

Recent research highlights the growing value of clinical text in healthcare analytics. Studies reviewing natural language processing applications to clinical narratives have shown that unstructured text often includes important clinical context that structured data may miss (Davidson, 2021). At the same time, work examining problem-oriented medical records demonstrates that converting narrative information inti structured formats can

improve the usability and interpretability of clinical data (Canales et al., 2021). Broader reviews of clinical text mining emphasize that transforming narrative descriptions into measurable features, such as TF-IDF, can support prediction and classification tasks in healthcare settings (Percha, 2021). Together, this body of research suggests that narrative hospital descriptions may hold valuable insight that can enhance quality assessment.

However, the usefulness of hospital ratings also depends on the fairness and completeness of the information used to generate them. Prior work has shown that public reporting systems can oversimplify hospital performance and fail to account for contextual differences across facilities. Additionally, geographic variation in hospital quality has been documented across the United States, indicating that regional factors may influence performance outcomes (Tsai et al., 2013). These findings motivate a deeper examination of both the language hospitals use in their written descriptions and geographic patterns present across California.

This study investigates whether narrative performance descriptions contain measurable signals that relate to hospital ratings and whether these signals can improve predictive modeling when combined with traditional variables. Using TF-IDF, we converted the written narratives into quantitative features and analyzed language differences across rating categories. We then developed two logistic regression models, one using only structured variables and one combining structured and text-derived features to evaluate the added value of narrative information. Collectively, these analyses aim to determine whether text data can enhance understanding of hospital quality and provide greater insight into variation across California hospitals. We also attempted to find out if any character fields in the data set existed that were actually very small, such as the descriptions for “Type of Report” per procedure, and if they contained a little bit more information concerning the quality of the respective hospitals.

Methodology

The dataset was from Data.gov, and is updated continuously by the California Department of Health Care Access and Information. It was imported into Google Colab, and cleaned by standardizing the encoding, stripping whitespace from the “Hospital Ratings” column, and converting all performance measure text to lowercase. Rows containing missing values for either the rating or the performance measure text were removed, and the index was reset to ensure alignment with later text-processing outputs. The original rating labels

(“Better,” “As Expected,” “Worse”) were recoded into standardized categories—“Above Expected,” “Meets Expectations,” and “Below Expected”—and then numerically encoded as 2, 1, and 0, respectively, for use in model training.

Text features were generated using a TF–IDF vectorizer with English stop-words removed and a maximum of 2000 features, transforming each hospital’s narrative performance measure description into a sparse matrix of weighted keyword frequencies. Structured features, including the number of adverse events, number of cases, and the risk-adjusted rate, were extracted from the dataset, had missing values imputed with zero, and were converted into a numerical sparse matrix. These structured features were horizontally stacked with the TF–IDF matrix to create a combined feature set capturing both numeric and text-based information.

The data was split into training and testing sets using an 80/20 ratio with stratification to maintain the distribution of the three rating categories. A multinomial logistic regression model was then trained using the LBFGS solver with class-balanced weighting to address rating imbalance and a high maximum iteration limit to support convergence with the high-dimensional sparse input. Model performance was evaluated on the test set using accuracy and the full classification report, and a structured-only logistic regression model was trained separately to compare its performance with the combined model.

For the purpose of determining whether the assumption made is correct or incorrect, a Naive Bayes classifier was developed to predict the hospital rating category using the ‘Type of Report’ text as input., based on the sample data available. The input text was transformed into a Bag-of-Words feature vector through the implementation of the CountVectorizer tool in order to determine the type of the ratings in the hospitals based on the input texts.

Results

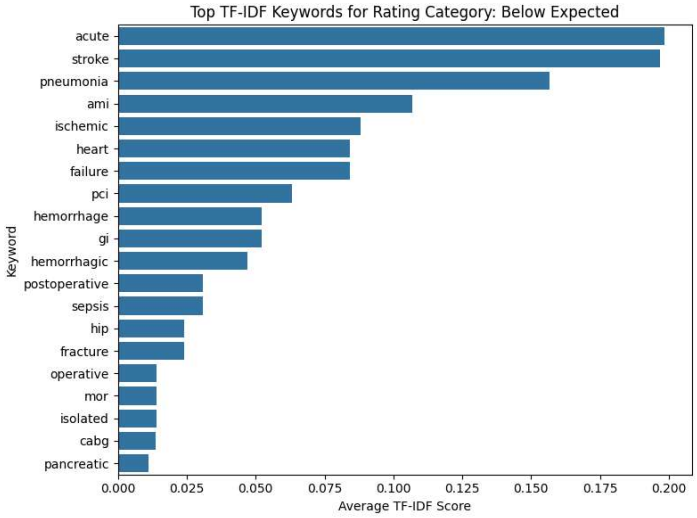

Figure 1.Barplot depicting the 20 highest TF-IDF scores in the “Below Expected” category.

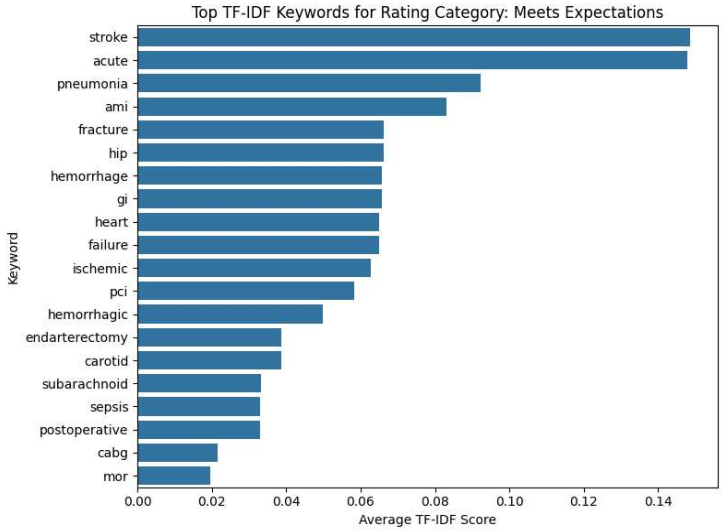

Figure 2. Barplot depicting the 20 highest TF-IDF scores in the “Meets Expectations” category.

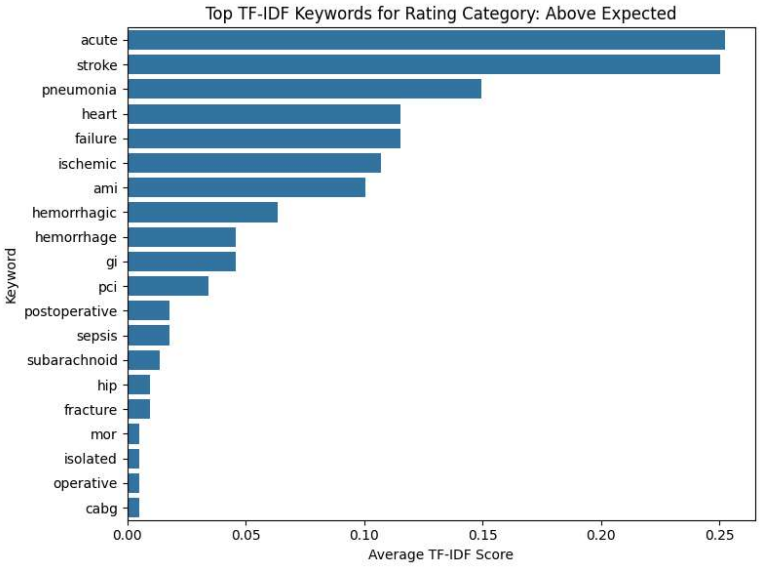

Figure 3. Barplot depicting the 20 highest TF-IDF scores in the “Above Expected” category.

Full-Model Accuracy = 70.5%	Structured Variables Only Model Accuracy = 91.6%
Below Expected Recall = 0.89	Macro-F1 Score = 0.52
Meets Expectations Recall = 0.68	Weighted F1 Score = 0.77
Above Expected Recall 0.95	

Figure 4. Table of Logistic Regression Results

Naive Bayes Model Accuracy: 0.9127757824525398

Classification Report:

Class	Precision	Recall	F1-Score	Support
0.0	0.00	0.00	0.00	165
1.0	0.91	1.00	0.95	3558

2.0	0.00	0.00	0.00	175
Accuracy			0.91	3898
Macro Avg	0.30	0.33	0.32	3898
Weighted Avg	0.83	0.91	0.87	3898

Figure 5: Classification Report Table and Naive Bayes Model Accuracy Score

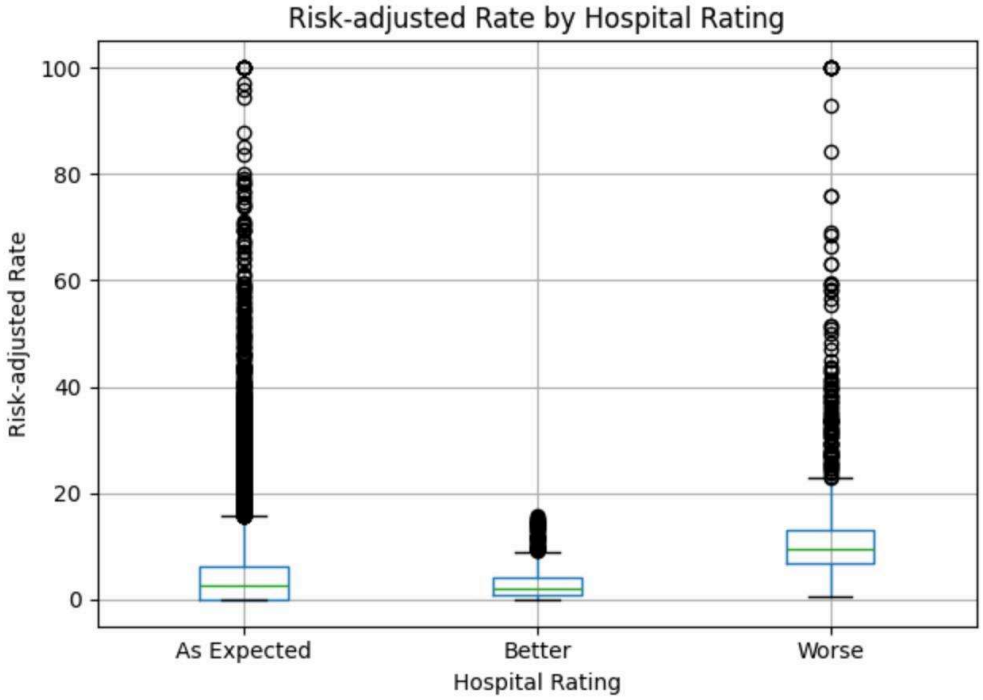

Figure 6. Boxplot of Risk-Adjusted Rate by Hospital Rating.

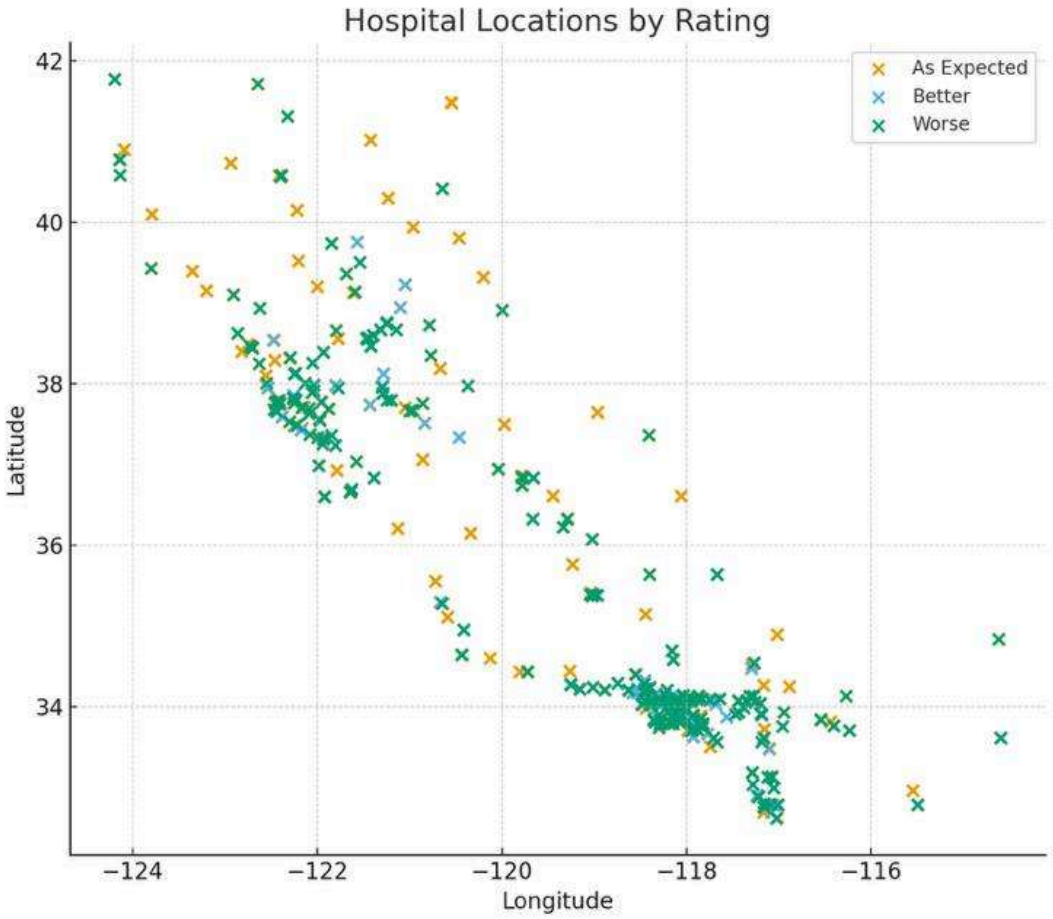

Figure 7. Map of hospital locations by Rating.

Discussion

The TF-IDF results show clear differences in the language used in clinical documentation across hospital performance categories. “Below Expected” hospitals had the highest weighting for acute, high-severity terms such as *acute*, *stroke*, and *pneumonia*, suggesting either a greater burden of complex cases or documentation patterns reflecting complications. “Meets Expectations” hospitals showed a more balanced mix of acute, postoperative, and cardiovascular terms, while “Above Expected” hospitals still referenced acute conditions but with comparatively lower weighting for complication-related terms. These patterns indicate that free-text clinical notes capture meaningful distinctions in case mix and documentation style across rating groups.

However, the logistic regression results show that structured variables remained far superior for prediction. The structured-only model reached 91.6% accuracy, far outperforming the full model with text

(70.5% accuracy). This suggests that TF-IDF features added noise rather than enhancing predictive signal. Even so, the full model achieved high recall for “Below Expected” (0.89) and “Above Expected” (0.95) hospitals, meaning text may still help identify the extremes of performance. The lower macro-F1 score (0.52) reflects class imbalance and difficulty classifying middle-category hospitals.Together, these findings show that clinical text contains meaningful qualitative differences across performance levels but, in its current form, does not improve model accuracy. More advanced NLP methods may be needed to capture deeper linguistic patterns and add predictive value beyond structured data.

The fact that the results confirm the significance of even short free-text entries indicates that the presence of free-text fields has a positive effect, and thus impacts the belief in the positive effect of the combination of free-text fields and structural variables in the understanding of the performance of the hospitals. The Naive Bayes model reached an accuracy of 0.9128 based solely upon the short procedure labels. This is consistent with what we expect from a Naive Bayes Model used on short text data. Though less than our model based upon the structured data, it obviously indicated a pattern linking the type of procedure to the levels of performance. This indicates that even minimal textual data conveys important information about hospital quality. In the context of the study question, the results support the premise that textual features can provide insight into the variation of hospital quality and should be explored prospectively along with other structured variables.

The visualizations helped make the patterns in the data a lot easier to understand. In figure 6, the boxplot comparing risk-adjusted rates across the three hospital rating categories, the differences were very clear. Hospitals that were rated “Below Expected” consistently showed higher risk-adjusted rates, while “Above Expected” hospitals had noticeably lower ones. This lineup with the structured-only logistic regression showed that the numerical clinical variables already carry most of the predictive power. Basically, if a hospital has higher complication-adjusted rates, it almost always falls into the lower performance categories. Seeing that same pattern visually makes the relationship between these structured variables and the ratings more obvious.

Furthermore, figure 7 is a geographic plot that adds more context to the findings. Even though the model itself didn't directly use geographic information, plotting the hospital locations by rating helped show that performance isn't evenly spread across California. Some regions had more lower-rated hospitals clustered together, while others had mostly average or above-average ratings. This supports the idea that regional differences can play a role in hospital performance, whether that's due to population characteristics, resources, or other local factors. Overall, these results show that both the text features and structured variables offer different kinds of insight into hospital performance, even if they don't contribute equally to prediction accuracy.

Conclusion

These findings from this analysis demonstrate that narrative hospital descriptions contain meaningful information that structured variables alone cannot capture. The TF-IDF results revealed distinct language patterns across rating categories, suggesting that hospitals performing below expectations often emphasize risk-heavy complications or complication-related terminology, while higher-performing hospitals use more positive or improve-oriented language. Although the structured-only logistic regression model achieved strong accuracy, its performance was influenced by class imbalance, limiting its usefulness in identifying minority rating groups. In contrast, the combined model incorporating TF-IDF features showed improvement recall for underrepresented classes, indicating that narrative information can strengthen predictive modeling and provide additional insight into performance differences. These results align with prior research that highlights the added value of clinical text in healthcare analytics (Davidson et al., 2021; Percha., 2021). The Naive Bayes test demonstrates the relative simplicity of NLP procedures for implementation and the potential for much more to be learned from the pattern in the data than if the data is put into a structured form to be understood by the computer algorithm. Naive Bayes shows that simple text features can still reveal small patterns the regular data doesn't capture. Overall, this study supports the idea that text data can enrich hospital quality assessment, improve minority class detection in predictive models, and provide deeper contextual insight into dynamics shaping healthcare performance across the state. Future research may expand on these findings by incorporating sentiment analysis, topic modeling, or longitudinal text data to uncover more detailed semantic patterns in hospital narratives.

References

Canales, L., Menke, S., Marchesseau, S., D'Agostino, A., del Rio-Bermudez, C., Taberna, M., & Tello, J. (2021). Assessing the Performance of Clinical Natural Language Processing Systems: Development of an Evaluation Methodology. *JMIR Medical Informatics*, *9*(7), e20492. https://doi.org/10.2196/20492

Tobin, R., Whalley, H., Wu, H., Alex, B., & Whiteley, W. (2021). The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Medical Imaging*, *21*(1). https://doi.org/10.1186/s12880-021-00671-8

Percha, B. (2021). Modern Clinical Text Mining: A Guide and Review. *Annual Review of Biomedical Data Science*, *4*(1). https://doi.org/10.1146/annurev-biodatasci-030421-030931

Tsai, T. C., Joynt, K. E., Orav, E. J., Gawande, A. A., & Jha, A. K. (2013). Variation in Surgical-Readmission Rates and Quality of Hospital Care. *New England Journal of Medicine*, *369*(12), 1134–1142. https://doi.org/10.1056/nejmsa1303118

McCallum, A., and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 41–48. AAAI Press.