Identifying Themes of GP Surgery Reviews via Text Analysis

Acey Vogelstein, Vaarun Muthappan

Center for Data Science

New York University

av3848@nyu.edu, vkm9614@nyu.edu

May 2025

1 Introduction

According to a customer survey, 94% of patients consult online reviews before choosing a healthcare practice.[1] As hospitals increase their digital presence and internet usage expands worldwide, patients are no longer limited to the nearest hospital; rather, they can evaluate providers across a broader landscape. Online reviews have largely replaced word-of-mouth recommendations. This shift toward online reviews presents an opportunity for hospitals: by analyzing online feedback, they can derive actionable insights into patient experiences and improve quality of care. With the rise of text-based analytics and modeling, distilling a large volume of text-based reviews is increasingly feasible. Such analysis is also ethical if personal identifiers are removed from the dataset, as is the case with the dataset used in our analysis. In this paper, we go beyond basic sentiment classification of text-based feedback and extract thematic insights within the sentiment classes. Hence, the research question we address in this paper is:

What are the key areas of praise and criticism of hospital care gathered from text-based patient feedback?

2 Description of the Text Data

To answer our research question, we analyze text data from a publicly available dataset on Hugging Face.[2] We selected this dataset as it contains thousands of text-based reviews as well as a numerical feedback column to serve as the ground truth for our sentiment classification. This dataset contains patients' feedback following their visits to practices in the "Brompton Health PCN North West London" network. The original dataset includes 12,000 rows and 10 columns, but we focus on two columns: 'free_text' (which includes text-based feedback from a given patient) and 'rating' (which includes an integer rating on a 1-5 scale from a given patient). Additionally, the original dataset includes a binary column called 'label', where 0 indicates a human-generated review and 1 indicates an AI-generated review. Our paper focuses on human-written reviews, so AI-generated reviews are removed. Furthermore, the original dataset includes approximately 100 duplicated rows, seemingly in error. After dropping duplicates as well as null entries in the 'free_text' and 'rating' columns, our final cleaned dataset consists of 6,058 rows.

3 Methods & Presentation of Results

BERT-Based Sentiment Classification

To examine the extent to which ratings align with the review sentiment, we apply a BERT-based model to classify the sentiment of each review. We use the model *cardiffnlp/twitter-roberta-base-sentiment* from Hugging Face to classify the reviews into 'positive', 'negative', and 'neutral' categories. Although BERT shows decent results in Figure 1, we decide not to rely on BERT for classification as the model produces many

obvious mismatches. For example, BERT classifies the review "*Nothing to be improved very happy with this type of appointment"* as negative, whereas the rating is 5 stars and the sentiment is clearly positive. Therefore, we instead use the numerical ratings to classify review sentiment, as the ratings directly given by the patients are more strongly indicative of their sentiment toward their hospital experience.

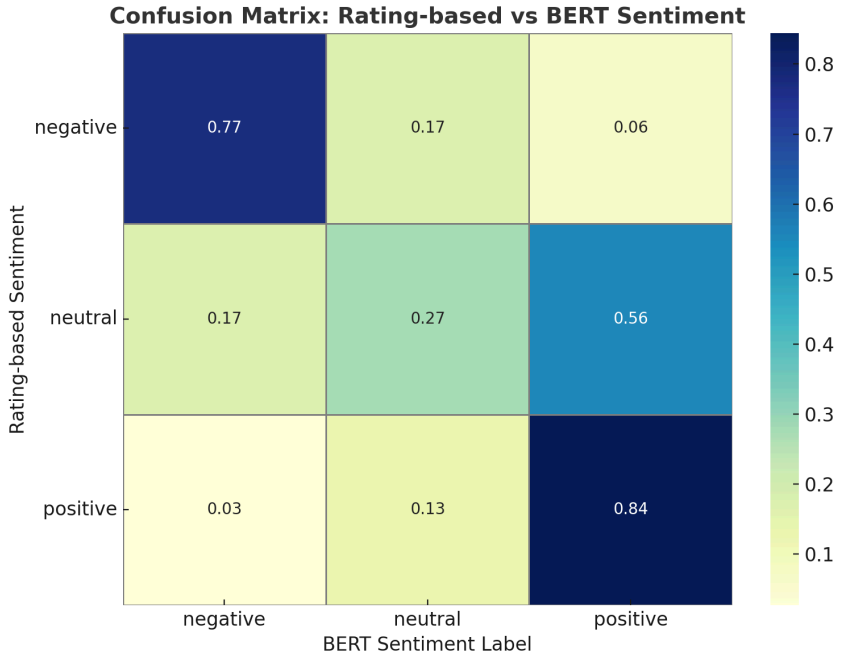

Figure 1: Heatmap of sentiment agreement between rating-based and BERT classifications.

Ratings-based Sentiment Classification

We introduce a new column called 'rating_based_sentiment'. Entries in this column are 'negative' where 'rating' = 1, 2, or 3; 'positive' where 'rating' = 5; and 'neutral' where 'rating' = 4. After subjective examination of sample reviews, we found that 4-star reviews are often ambiguous or weak in sentiment; therefore, we exclude all

reviews where ‘rating’ = 4 as we move forward in our analysis. The following example of a 4-star review illustrates this ambiguity:

- “*Online appointment booking is great however sometimes struggle to get through to the office And previously despite having said it was a face to face the wrong time was booked However the team is generally lovely*”.

As a result, the ‘rating_based_sentiment’ column effectively becomes binary. A vast majority (>70%) of the reviews come with a rating of 5 stars, as shown in Figure 2. This ratings-based sentiment classification (1-3 stars grouped as negative, 5 stars as positive) is applied to the methods below unless otherwise specified.

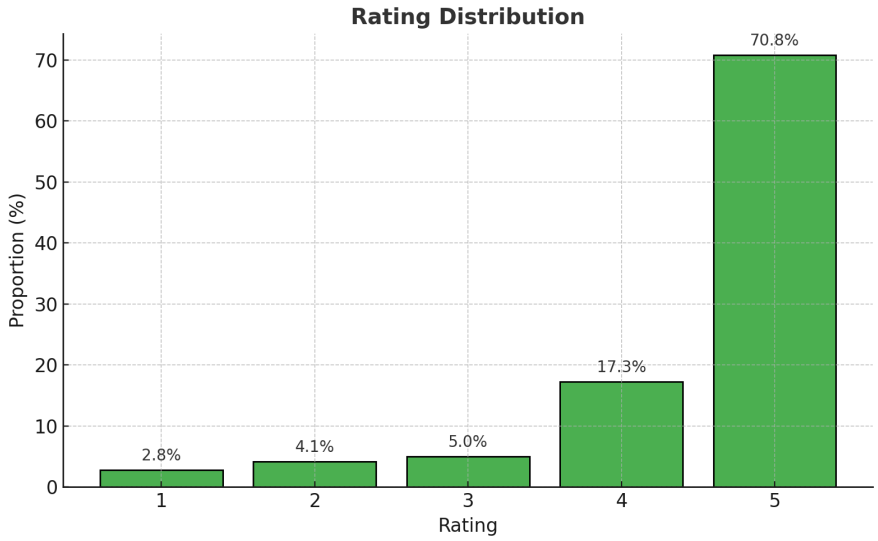

Figure 2: A bar chart showing the distribution of reviews by rating.

Logistic Regression

Using logistic regression regularized with elastic net, we perform binary sentiment classification on text reviews in order to identify the terms that serve as the top drivers of positive and negative sentiment. Ground truth sentiment labels come from

the ‘rating_based_sentiment’ column. We split the data into 80% training and 20% test. We also remove stop words and stratify the sampling, which is important due to the class imbalance. Additionally, we use CountVectorizer–if a word is relatively much more prevalent across reviews with a given sentiment label, then that word is likely indicative of the given sentiment. Next, we normalize (first by row, then by column) to ensure that varying review length does not bias the model, and that the features (words) are standardized for comparison of coefficients. Finally, we evaluate feature importance by identifying the magnitude of coefficients assigned to each word.

Figures 4 and 5 capture the performance of our elastic net-regularized logistic regression model. The results indicate many false positives and false negatives, and that the recall for the negative class is much lower than that of the positive class. Even with the model’s imperfect performance, many coefficients of the terms remain informative for identifying words that are most associated with increases or decreases in rating, as shown in Figure 3. The top positive drivers reflect gratitude, competence, professionalism, and efficiency of care. Conversely, the top negative drivers reflect rudeness and other critical though less interpretable themes.

```
↑ Top Positive Drivers:
 ['helpful' 'professional' 'quickly' 'efficient' 'friendly' 'explained'
 'excellent' 'best' 'nurses' 'thank' 'thorough' 'quick' 'attentive'
 'promptly' 'satisfied']

↓ Top Negative Drivers:
 ['rude' 'disappointed' 'appointment' 'told' 'don' 'poor' 'health' 'blood'
 'request' 'difficult' 'doctor' 'called' 'definitely' 'worst' 'patient']
```

Figure 3: Words that most strongly drive positive and negative sentiment according to the coefficients from the logistic regression model.

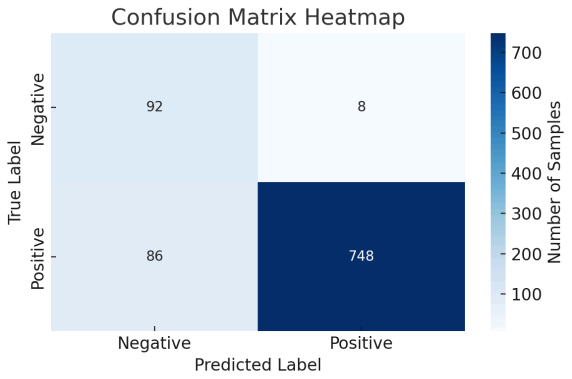

Figure 4: Confusion matrix of predicted class from logistic regression.

Accuracy: 0,862

Classification Report

	precision	recall	f1-score	support
Negative	0.52	0.66	0.58	140
Positive	0.94	0.90	0.92	834
accuracy			0.86	974
macro avg	0.73	0.78	0.75	974
weighted avg	0.88	0.86	0.87	974

Figure 5: Classification report from logistic regression.

Reviews Across Practices

The dataset consists of reviews across 12 GP practices in the Brompton Health network; however, the dataset does not label the specific practice tied to each review. In order to find any insights that might be associated with a health practice, we use a substring match based on Levenshtein distance to identify the reviews that mention the names of the practices. First, we clean the text by lowercasing and removing punctuation. We then import the *fuzz* module from the *rapidfuzz* library to calculate the similarity, and we set the similarity threshold to 80 (the range is 0 to 100, where 0

imposes no similarity requirement and 100 requires an exact match). This threshold offers flexibility in case a patient misspells the practice name or uses a variation of the proper spelling.

We decide to exclude “The Good Practice” from our fuzzy matching, as this practice’s name is difficult to distinguish from practices merely described as good. In Figure 6, we can see the ratings distribution among the other 11 practices. There are only 130 total reviews that fuzzy-match with any practice name, and there exists bias in terms of the frequency with which practice names happen to be mentioned in reviews. While it is interesting that reviews that include references to Stanhope and Chelsea have relatively many more 5-star ratings, we are constrained from reaching statistically significant conclusions on the sentiment around specific practices due to bias and the small sample size.

Number of Reviews by Practice and Rating (1–5):

matched_practice	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Stanhope Mews West Surgery	0	0	1	3	24
Health Partners At Violet Melchett	1	0	0	1	0
Emperor’s Gate Centre for Health	0	0	0	0	0
The Chelsea Practice	0	1	1	5	34
Abingdon Medical Centre	0	0	0	2	13
Earls Court Surgery	0	0	0	1	10
Earls Court Medical Centre	2	2	0	2	4
Knightsbridge Medical Centre	0	0	0	0	0
Kensington Park Medical Centre	0	0	0	0	3
Royal Hospital Chelsea	0	0	0	0	2
Scarsdale Medical Centre	0	2	0	0	11

Figure 6: Mentions of each practice across ratings.

LDA

We run an Latent Dirichlet Allocation (LDA) model on each binary class to identify themes across the reviews. Rather than running LDA across the entire dataset, we split

the LDA analysis by sentiment class in order to reveal sentiment-specific insights. We also tune the optimal number of topics (k) via coherence scoring.

Based on coherence scoring, we select 3 topics (k=3) for both LDA analyses, with optimal coherence scores of 0.466 for negative reviews and 0.544 for positive reviews. After receiving the terms in each LDA topic as shown in Figure 7, we prompt ChatGPT for topic descriptions from each LDA topic to obtain more objective insights into each topic. More specifically, we uploaded Figure 7 to ChatGPT and wrote, “Please interpret these topics.” For negative reviews, Topic 1 focuses on frustration with access and communication; Topic 2 emphasizes medical testing issues, such as long wait time; and Topic 3 points to the lack of face-to-face appointments with doctors. For positive reviews, Topic 1 accounts for attentiveness of the staff; Topic 2 accounts for efficient care delivery; and Topic 3 accounts for general satisfaction about the care.

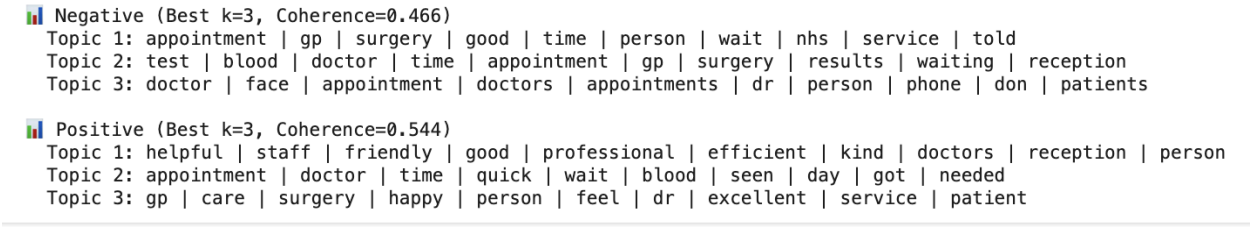

Figure 7: Topics from the negative and positive class.

Decision Trees

To identify words that are strongly indicative of a specific sentiment, we run a decision tree classifier trained to predict each of the five rating scores from the words in each review. We split 60% of the data into the training set, 20% into the validation set, and 20% into the test set. We then lemmatize the reviews in order to group all the variations of a word together before running CountVectorizer on the reviews (for the

same reason covered in the Logistic Regression section). Different max_depth and min_samples parameters–[5, 10, 20] and [2, 5, 10], respectively–were tuned on the validation set before using the optimally tuned parameters to find the test set loss. We use cross-entropy loss, as it is an appropriate metric for categorical predictions.

Of all the hyperparameters tested, the model with a max depth of 5 and min_samples_split of 2 produce the lowest loss on the validation set. The cross entropy loss on the test set is 0.9609 (better than random for the 5 possible rating values). In Figure 8, we see the most important words, such as: *tell*, *say*, *rude*, *send*, *listen*, *end*, *practice*, *and save.* The words *tell*, *say*, and *rude* all pertain to communication, highlighting the importance of effective communication to the patient experience. Additionally, after checking reviews with the substring “save”, it appears that these patients prioritize “saving” time, money, and, of course, their lives. Since *end* and *practice* appear in a broader range of contexts, their importance is less interpretable.

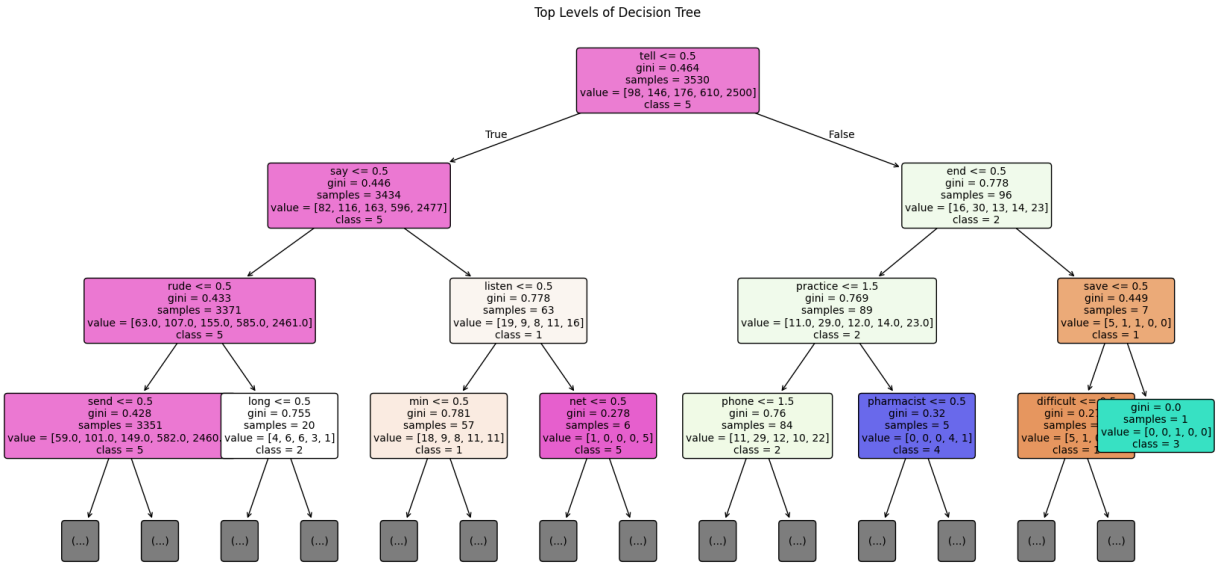

Figure 8: Decision tree diagram representing the best-performing model.

Analyzing Sentiment Towards Employees

Since the reviews contain specific mentions of doctors and staff (where the names are replaced with the word “PERSON” to protect privacy), we analyze the sentiment toward the staff by splitting these reviews into sentiment classifications. Since there are only 720 of these reviews, we read through them to see what patients are criticizing or praising in this context. Table 1 shows the count of reviews containing ‘PERSON’ by rating.

Rating	Count of Reviews with 'PERSON'	Total Reviews	% of Reviews with 'PERSON'
1	18	166	10.84%
2	22	1175	1.87%
3	19	2267	0.84%
4	89	3917	2.27%
5	572	4273	13.39%

Table 1: Mentions of review across ratings.

The analysis indicates that most mentions of specific doctors are strongly positive. Example reviews include:

- “*the doctor PERSON was great*”
- “*Every time when i go to the to ask for an appointment At the reception I have very friendly staff my GP is very nice and understanding and easy to talk with about my health issues I m very happy and grateful 🙏 patient Many thanks PERSON*”

On the other end, a relatively high proportion of 1-star reviews reference a doctor or staff member as well. This trend implies that doctors play a key role in shaping both highly positive and highly negative experiences. For example, a 1-star review says the following:

- "*I left the appointment feeling frustrated and still without clear answers I would not recommend Dr PERSON if you are looking for a doctor who genuinely cares about their patients*".

Frequency of Specific Nouns in Each Binary Class and Their Most Frequent Co-Occurring words

We identify the most common nouns in the text to attempt to capture a comprehensive list of topics that patients focused on. We also examine the words that most frequently co-occur with nouns and the rating distribution associated with each noun to assess the binary sentiment around each noun. Co-occurring words are defined as words appearing in the same review as the given noun.

Table 2 shows the most common nouns within each binary sentiment class. Interestingly, there is no clear separation of themes between the classes–the most common nouns are similar in meaning, though in a different order. In general, the prevailing content relates to the service and people, with top words including *doctor*, *staff*, *service*, *person*, *receptionist,* and *nurse.*

Nouns from a Negatively Classified Review	Nouns from a Positively Classified Review
surgery: 15138	doctor: 1081
staff: 13753	appointment: 865
time: 10004	staff: 857
receptionist: 9938	person: 692
appointment: 9784	time: 648
nurse: 9627	surgery: 490
doctor: 8730	service: 489
experience: 8519	reception: 399
visit: 5184	care: 367
concern: 4580	nurse: 311

Table 2: 10 most frequent nouns from each sentiment class.

4 Discussion of the Results and Takeaways:

This paper moves past simple sentiment classification of text-based feedback to uncover thematic patterns within each sentiment category. The analysis of patient reviews from the “Brompton Health PCN North West London” healthcare network reveals several key insights. Notably, 4-star reviews emerge as an ambiguous category. They often contain mixed sentiments, so they weakly indicate overall patient satisfaction. In contrast, 3-star ratings or below are strongly indicative of negative feedback, whereas 5-star reviews tend to express praise. Secondly, sentiment classification via BERT appears unreliable in this context, apparently due to challenges in capturing subtle emotional tones and hospital-specific language. Additionally, due to a limited sample of reviews that directly mention practice names, our attempt to analyze reviews by practice draws inconclusive results.

Furthermore, thematic analysis of negative reviews points to dissatisfaction with outcomes, poor communication, and logistical challenges. In particular, logistic regression within the negative sentiment class exposes rudeness, and LDA reveals critical themes of long wait times and difficulty accessing appointments. Decision tree analysis reveals that service-related qualities such as *listen* and *rude* are the most important identifiers of whether a review is poorly rated or highly rated. On the other hand, the same methods reveal that positive reviews are driven by attentiveness, care, professionalism, and successful patient outcomes. In addition, noun usage patterns reinforce that service quality shapes patient feedback, as the most frequent nouns refer to service-related qualities. Reviews rated at the extremes–i.e., 1 or 5 stars–are relatively much more likely to mention doctors and staff by name. Nevertheless, the

substantially higher volume of 5-star reviews than 1-star reviews suggests an overall high level of satisfaction with employees.

5 Limitations & Future Work

Certain limitations in the dataset have constrained our analysis. First, the dataset is imbalanced–just 2.8% of reviews receive 1 star and 4.1% of reviews receive 2 stars. With approximately 6,000 reviews overall, strong negative sentiment is underrepresented. Additionally, the dataset does not include patient characteristics such as age, gender, or visit type, which prevents any subgroup analysis. We also lack labeling of the 12 different health practices in the network. Furthermore, due to ethical considerations, all doctor names are anonymized as “PERSON” in the reviews, so we are unable to identify sentiment themes around specific doctors.

We also acknowledge potential self-selection bias, as it is unclear whether every patient was required to fill out the feedback form. In addition, the dataset exhibits post-visit feedback bias: the reviews naturally exclude patients who died or were unfit to provide feedback following a hospital visit. The outcome of the patient’s treatment, which is not directly available in the dataset, remains a potential confounder. Many negative reviews note unresolved health issues or failed follow-ups. However, without explicit outcome data, we can only guess–without significant statistical inference–that treatment success may heavily influence sentiment.

With a dataset featuring more descriptive columns, future work could evaluate sentiment themes within subcategories such as demographic groups, health practices, visit types, and staff members. Additionally, a higher volume of strongly critical reviews

would enable deeper insights into negative sentiment themes. Future studies also could explore the extent to which text-based feedback is associated with clinical outcomes or objective measures of care quality. Lastly, other forms of feedback such as verbal input and patient outcomes should be leveraged as well to improve the hospital experience.

6 References

1. https://www.softwareadvice.com/resources/how-patients-use-online-reviews/#back
2. https://huggingface.co/datasets/janduplessis886/gp_surgery_reviews_fake_and_real
3. AI disclosure: We used ChatGPT to help troubleshoot our code, help improve our writing in the report, and help generate our figures.