Hospital Reviews Topic Modelling and Sentiment Analysis

Analysis Report

Jishen Harilal

6/27/25

Natural Language Processing / Text Analytics

Table of Contents

Introduction

This report explores the use of text analysis techniques to help better understand public feedback by analysing online reviews. The goal is to extract meaningful insights from unstructured feedback that can guide service improvements and customer engagement strategies.

For this project, the **Hospital Reviews Dataset** from Kaggle (https://www.kaggle.com/datasets/junaid6731/hospital-reviews-dataset) was selected . It contains 996 sentiment-labelled reviews sourced from **Google Maps**, reflecting patient experiences at hospitals in **Bengaluru, India**. The format of these reviews - combine written feedback with a 1–5 rating, making the dataset highly relevant and suitable for this task.

The analysis focuses on two main objectives:

1. **Topic modelling**, to uncover recurring themes in the feedback; and
2. **Sentiment analysis**, to classify reviews as positive or negative and evaluate sentiment trends.

These combined approaches offer both thematic understanding and sentiment classification, enabling a deeper and more actionable view of customer concerns. This report details the process of dataset preparation, model development, evaluation, and key findings, with a focus on how these techniques can be applied to real-world feedback streams.

Dataset Overview

The dataset (hospital.csv) contains 996 records and 3 features:

Variables

- Feedback – The unprocessed text of the review (detailed feedback about hospital services.
- Sentiment Label – Binary value indicating positive or negative sentiment associated with the review.
- Ratings – A numerical score/star rating (1-5) representing the user's overall satisfaction.

Data Overview and Quality Considerations

The dataset is well-structured and contains clearly labelled columns relevant to the task, including written feedback, sentiment labels, and rating scores. However, a few quality issues were identified during initial inspection.

An unnamed column with no useful data was present and will be dropped during the cleaning phase. Duplicate entries were found and removed before data preprocessing continued, no missing data was found what remained.

The text feedback required standard natural language preprocessing, including converting text to lowercase, removing punctuation, tokenisation, stop word filtering, and lemmatisation. These steps were essential to prepare the data for topic modelling and sentiment classification.

One important consideration is the presence of a sentiment imbalance, with significantly more positive reviews than negative ones. This is addressed later in the report with findings to go with.

Exploratory Data Analysis (EDA)

Statistical Analysis

Summary Statistics:

	Sentiment Label	Ratings
count	996.000000	996.000000
mean	0.730924	3.567269
std	0.443703	1.408693
min	0.000000	1.000000
25%	0.000000	2.000000
50%	1.000000	4.000000
75%	1.000000	5.000000
max	1.000000	5.000000

Figure 1: Statistical summary

The summary in Figure 1, shows that the dataset contains 996 records across three columns: Feedback (text), Sentiment Label (binary: 0 = negative, 1 = positive), and Ratings (integer values from 1 to 5). All entries are complete with no missing values.

- The Sentiment Label is positively skewed, with a mean of 0.73, indicating that approximately 73% of reviews are positive. This confirms a noticeable class imbalance, which has implications for model training and evaluation.
- The Ratings column also reflects a positive trend, with a mean of 3.57 and a median of 4, suggesting that most reviewers rated their experiences above average.
- Ratings range from 1 to 5, with 75% of reviews scoring 2 or higher, and 25% scoring the maximum 5, reinforcing the general positivity of the dataset.

Data Preparation

	Feedback	Sentiment Label	Ratings	Cleaned_Feedback
0	Good and clean hospital. There is great team o...	1	5	good clean hospital great team doctor good ot ...
1	Had a really bad experience during discharge. ...	1	5	really bad experience discharge need sensitive...
2	I have visited to take my second dose and Proc...	1	4	visited take second dose process really smooth...
3	That person was slightly clueless and offered...	1	3	person slightly clueless offered one package g...
4	There is great team of doctors and good OT fac...	0	1	great team doctor good ot facility
5	My primary concern arose from the insistence o...	0	2	primary concern arose insistence conducting mu...
6	Good and clean hospital. The medical faciliti...	1	5	good clean hospital medical facility great goo...
7	Recently underwent a surgery for my left shoul...	1	3	recently underwent surgery left shoulder docto...
8	Over all experience was good, starting from re...	1	5	experience good starting receptionlab service ...
9	However,the services of front office (where we...	1	5	howeverthe service front office report first n...

Figure 2: Cleaned_Feedback field added to dataset

The dataset underwent several cleaning steps to prepare it for analysis:

- **Deduplication:**
  The dataset initially contained 996 records, but a small number of duplicates (19) were identified and removed. After deduplication, 977 unique records remained for analysis.
- **Feedback Cleaning and Preprocessing with Natural Language Toolkit (NLTK)**

  The Feedback field contains plaintext reviews from customers and needed to be prepared to be used as training data for topic modelling and sentiment analysis. Natural Language Toolkit (NLTK) was used to perform a series of preprocessing steps.

  First, all text was converted to lowercase to ensure consistency in word representation. Punctuation was removed to reduce noise in the data. The text was then tokenised into individual words using NLTK's word tokenizer. Common English stop words (e.g., "and", "the", "is") were removed to keep only meaningful content words. Afterwards, each remaining word was lemmatised using NLTK's WordNetLemmatizer, which reduces words to their base or dictionary form (e.g., "running" becomes "run"). This process helps standardise the vocabulary and reduce redundancy in the dataset.

  The cleaned text was stored in a new column called **Cleaned_Feedback**, which can be used for training text analysis models (Figure 2.). This preprocessing step was essential to improve the quality and relevance of the textual features extracted from the reviews.

  These steps ensured the data was consistently formatted, free of duplicate entries, and ready for further exploration and model development.

Distribution Analysis

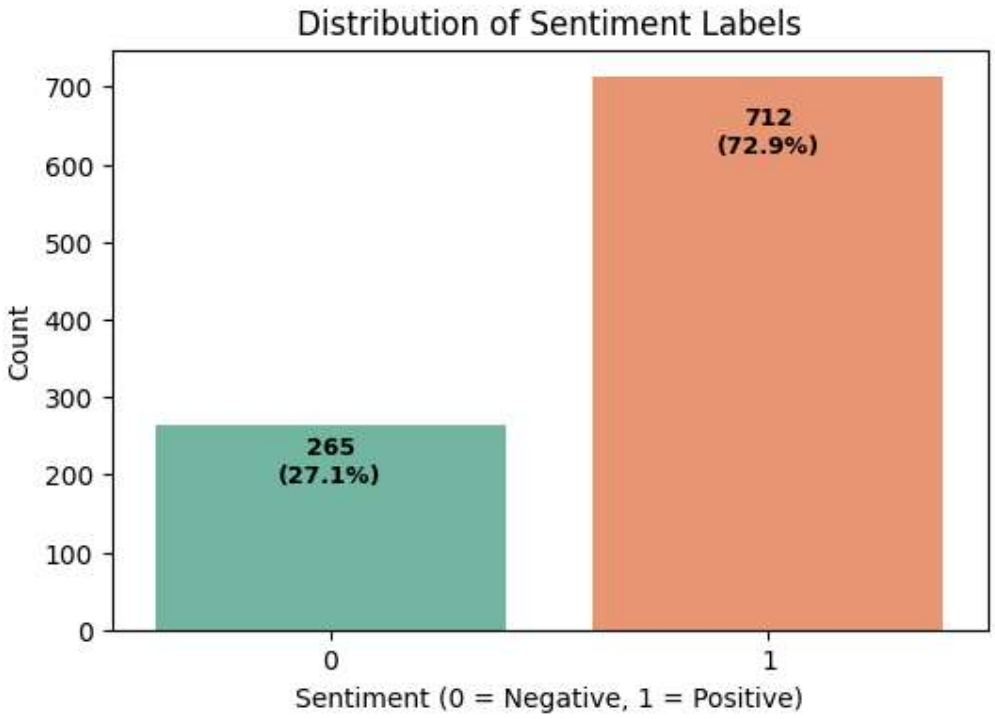

Figure 3:Sentiment Label Distribution.

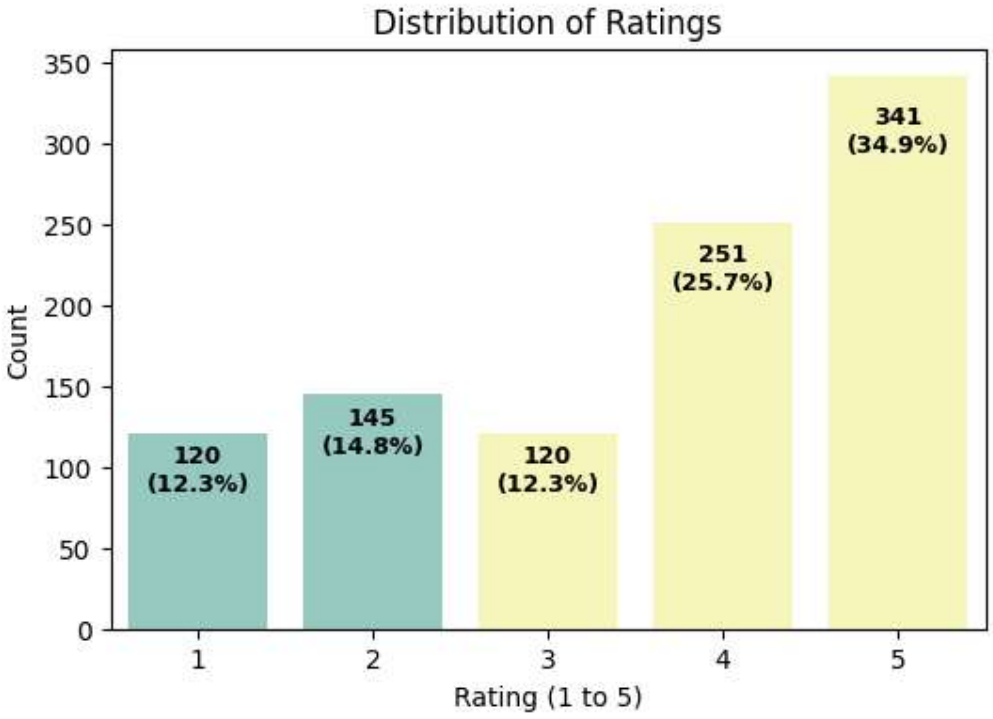

Figure 4: Ratings Distribution

The histograms in Figures 3 and 4 were used to visualise the distributions of Sentiment Label and Ratings. It was observed that significant class imbalance between negative and positive sentiment exists, with negative feedback being represented less than half as much as positive feedback. This can also be seen in the distribution of Ratings, where a much higher presence of 4 and 5 scores vs 1, 2 and 3 scores, which indicates that Ratings are highly correlated with Sentiment.

Visual Inspection of Cleaned Feedback

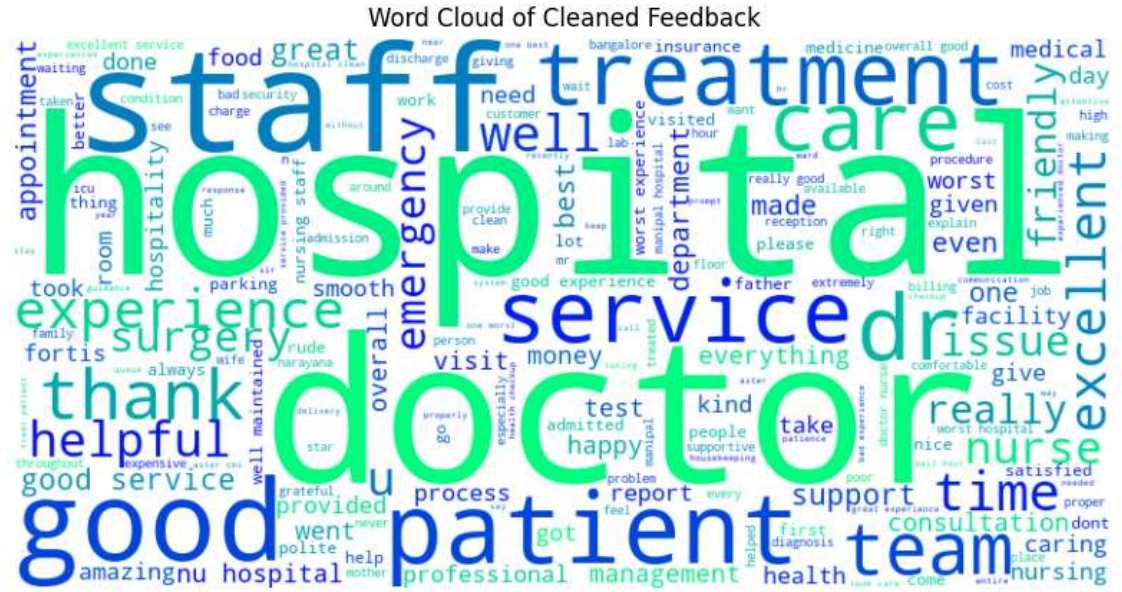

Figure 5: Word Cloud of Cleaned Feedback

Figure 5 shows a word cloud, which was used to provide a quick visual overview of the most common themes and keywords present in the 'Cleaned_Feedback' column. Prominent terms such as "*doctor*", *"staff"*, and *"hospital"* suggest that much of the feedback is centred around medical personnel and healthcare services. Additionally, the presence of positive words like *"friendly"*, *"excellent"*, and *"helpful"* aligns with the overall sentiment distribution of the dataset, which contains most positive reviews. Despite this, some less frequent but still visible negative terms such as *"worst"*, *"issue"*, and *"problem"* can also be observed, indicating that negative experiences are present, though less dominant.

Topic Modelling

LDA Model Training Summary

Topic Number	Top 20 Words
0	doctor, staff, patient, worst, hospital, experience, appointment, wait, time, management, money, good, treatment, helpful, went, hour, bad, consultation, dont, health
1	dr, doctor, treatment, staff, good, service, thanks, team, experience, nurse, nursing, care, excellent, caring, hospital, surgery, amazing, happy, helped, emergency
2	hospital, good, service, staff, doctor, patient, experience, care, excellent, treatment, nurse, nu, great, worst, best, facility, overall, emergency, team, really
3	staff, really, report, patient, issue, lot, good, time, health, medicine, care, service, rude, parking, gave, got, come, team, professional, communication

Figure 6: List of Topic Keywords

	Feedback	Sentiment Label	Ratings	Cleaned_Feedback	Dominant_Topic
0	Good and clean hospital. There is great team o...	1	5	good clean hospital great team doctor good ot ...	2
1	Had a really bad experience during discharge. ...	1	5	really bad experience discharge need sensitive...	0
2	I have visited to take my second dose and Proc...	1	4	visited take second dose process really smooth...	3
3	That person was slightly clueless and offered...	1	3	person slightly clueless offered one package g...	3
4	There is great team of doctors and good OT fac...	0	1	great team doctor good ot facility	2

Figure 7: Dominant_Topic field added to dataset

Latent Dirichlet Allocation (LDA) was used to find common themes in the customer feedback data. The 'Cleaned_Feedback' column was first transformed into a document-term matrix using CountVectorizer, which removed common English stop words and ignored very rare or very frequent words to reduce noise (max_df=0.95, min_df=2). LDA was then applied with 4 topics, which was chosen to give a manageable number of clear, interpretable themes.

For each topic, the top 20 keywords were listed to help understand what the topic is about as seen in Figure 6. Each feedback entry was assigned a dominant topic, based on the topic with the highest probability in that document and is seen in Figure 7. This made it possible to group reviews by topic and explore patterns such as average rating and sentiment for each group.

The results helped identify key themes in the feedback, including positive experiences with staff and service, as well as common complaints such as waiting times or poor communication. Grouping the feedback in this way gives useful insight into what customers value and where improvements may be needed.

Interpretation and Evaluation of Topics

Figure 8: Topic Word Clouds

Dominant_Topic	Feedback Count	% Positive	% Negative	Avg Rating	1	2	3	4	5
0	195	52.3	47.7	2.85	38 (19.5%)	55 (28.2%)	38 (19.5%)	27 (13.8%)	37 (19.0%)
1	229	83.4	16.6	4.03	16 (7.0%)	22 (9.6%)	10 (4.4%)	73 (31.9%)	108 (47.2%)
2	447	78.3	21.7	3.70	48 (10.7%)	49 (11.0%)	53 (11.9%)	136 (30.4%)	161 (36.0%)
3	106	65.1	34.9	3.28	18 (17.0%)	19 (17.9%)	19 (17.9%)	15 (14.2%)	35 (33.0%)

Figure 9: Topic-Level Statistical Summary

To interpret the LDA-generated topics, several steps were taken to provide both visual and numerical insight. The word clouds in Figure 8 were created for each topic using the top keywords, helping to visually highlight dominant themes. Then, topic-level summaries seen in Figure 9 were generated, showing feedback counts, sentiment proportions, average ratings, and the distribution of star ratings. This combination of visual and tabular data allows for clearer understanding of the concerns and sentiments associated with each topic.

The findings for each topic are presented on the next page.

Topic	Theme	Top Words	Sentiment	Avg. Rating	Rating Spread	Interpretation	Insights
0	Frustration and Poor Experience	doctor, staff, patient, worst, appointment, wait, bad, hour, consultation, money	52.3% positive / 47.7% negative	2.85	67.2% in 1–3 stars	Reviews mention delays, poor service, and dissatisfaction with consultations and value.	Waiting times and administrative coordination need improvement.
1	Exceptional Clinical Care	dr, doctor, treatment, care, thanks, nurse, surgery, excellent, happy, amazing	83.4% positive (strongest)	4.03	~80% in 4–5 stars	Highly positive feedback on doctors, nurses, and successful treatments; emotionally impactful experiences.	Reflects best-in-class healthcare delivery.
2	General Satisfaction with Hospital	hospital, service, doctor, patient, excellent, care, team, facility	78.3% positive	3.70	Even spread, leaning 4–5 stars	Captures overall trust and satisfaction with hospital quality, without intense emotion.	General praise; a broader, less specific version of Topic 1.
3	Communication & Process Issues	staff, really, report, issue, time, rude, parking, communication, health	65.1% positive	3.28	Mixed, slight lean to 4–5 stars	Highlights non-clinical issues like rude staff, parking trouble, and poor communication.	Signals service delivery and interpersonal issues.

Figure 10: Interpretation of Topics

Topic Quality Summary

- **Separation**: Topics are well-separated across clinical, logistical, and interpersonal concerns.
- **Number of Topics**: 4 is appropriate for ~1000 reviews as more topics risk diluting meaning, and fewer may merge distinct issues.
- **Emerging Themes**: Praise for doctors and care (Topics 1 & 2); dissatisfaction with process and support (Topics 0 & 3).

Distribution Analysis of Topics

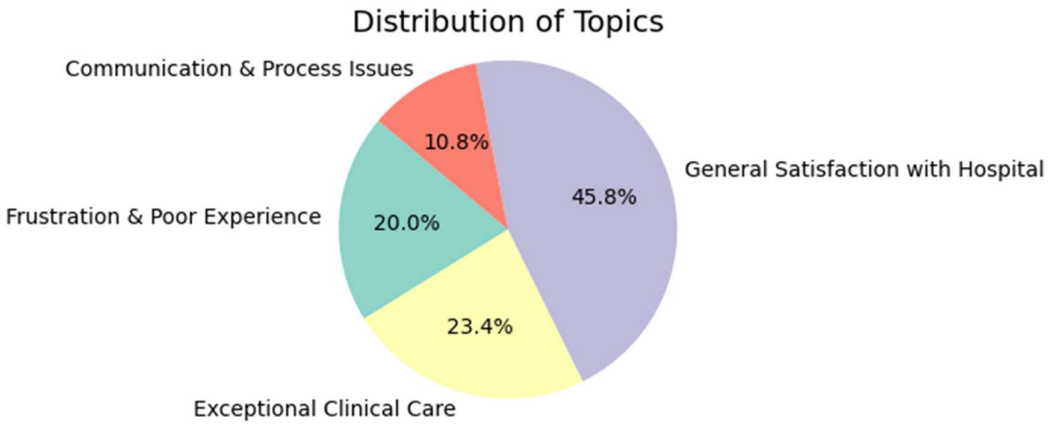

Figure 11: Distribution of Topics

Interpretation of Topic Distribution

The topic distribution shows that most reviews are positive, with General Satisfaction with Hospital (45.8%) and Exceptional Clinical Care (23.4%) making up nearly 70% of the feedback. This aligns with what we know about the dataset, which has a strong majority of positive sentiment and high ratings. However, *Frustration & Poor Experience* (20%) and *Communication & Process Issues* (10.8%) still represent a notable share of concerns, mostly around delays, staff behaviour, and service delivery, highlighting key areas for improvement despite overall positive experiences.

Alignment Between Topic Distribution and Sentiment Balance

The distribution of topics derived from LDA closely reflects the known sentiment balance of the dataset, which consists of approximately 70% positive and 30% negative feedback. Topics representing positive experiences - such as *General Satisfaction with Hospital* (45.8%) and *Exceptional Clinical Care* (23.4%) - together account for around 69.2% of the dataset. In contrast, themes associated with negative experiences - *Frustration & Poor Experience* (20%) and *Communication & Process Issues* (10.8%)- make up the remaining 30.8%. This close alignment suggests that the topic modelling process has effectively captured the underlying sentiment structure of the reviews, reinforcing the validity and coherence of the identified themes.

Sentiment Analysis

XGBoost + TF-IDF Model Training Summary

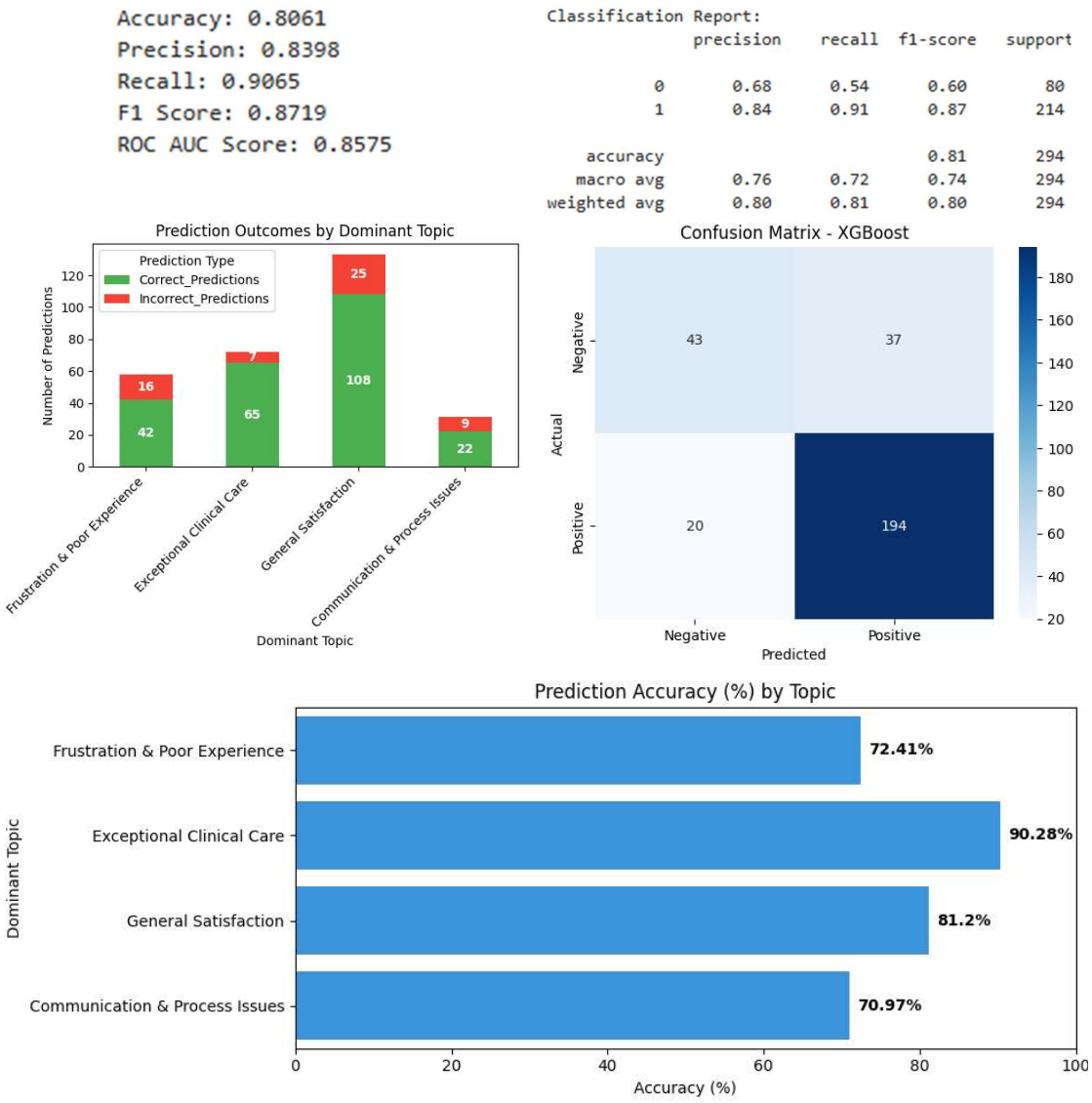

Figure 12: Results from XGBoost + TF-IDF Model

Sentiment Analysis Model Training Summary

An XGBoost (Extreme Gradient Boosting) classifier was trained to perform binary sentiment classification (positive vs negative) on the pre-processed hospital review data. The textual feedback was first vectorised using TF-IDF, capturing important word features while reducing noise. A stratified train-test split (70:30) ensured that both positive and negative sentiments were proportionally represented in each set.

Model Evaluation

The sentiment analysis model shows strong overall performance, with an accuracy of 80.6%, precision of 83.98%, recall of 90.65%, and an F1 score of 87.19%. The high recall and F1 score for the positive class (1) indicate the model is effective at correctly identifying positive feedback. The ROC AUC score of 85.75% further suggests that the model has solid discriminative power overall.

However, a closer look at the confusion matrix and class-specific metrics reveals a significant issue: the model struggles with identifying negative feedback (class 0). Out of 80 actual negative reviews, only 43 were correctly classified, while 37 were incorrectly predicted as positive (false positives). This leads to a recall of just 54% for the negative class, meaning nearly half of the negative feedback goes undetected.

This imbalance is critical in a healthcare context, where identifying negative experiences is especially valuable for service improvement. The model's tendency to misclassify negative reviews stems in part from the dataset's inherent class imbalance - with positive feedback being far more frequent than negative. The weighted average scores, while high, can therefore obscure performance issues on the minority class.

To better understand model performance, prediction accuracy across the four dominant topics were analysed. The model performed best on Topic 1 ("Exceptional Clinical Care") and Topic 2 ("General Satisfaction with Hospital"), achieving 90.28% and 81.20% accuracy respectively, likely due to their clear and consistent positive language. In contrast, Topic 0 ("Frustration and Poor Experience") and Topic 3 ("Communication & Process Issues") saw lower accuracies (72.41% and 70.97%), reflecting the challenges of classifying more complex or mixed feedback. These differences highlight the model's stronger performance on positive sentiment and its limitations in handling subtler, critical reviews.

Model Retraining with Class Weighting

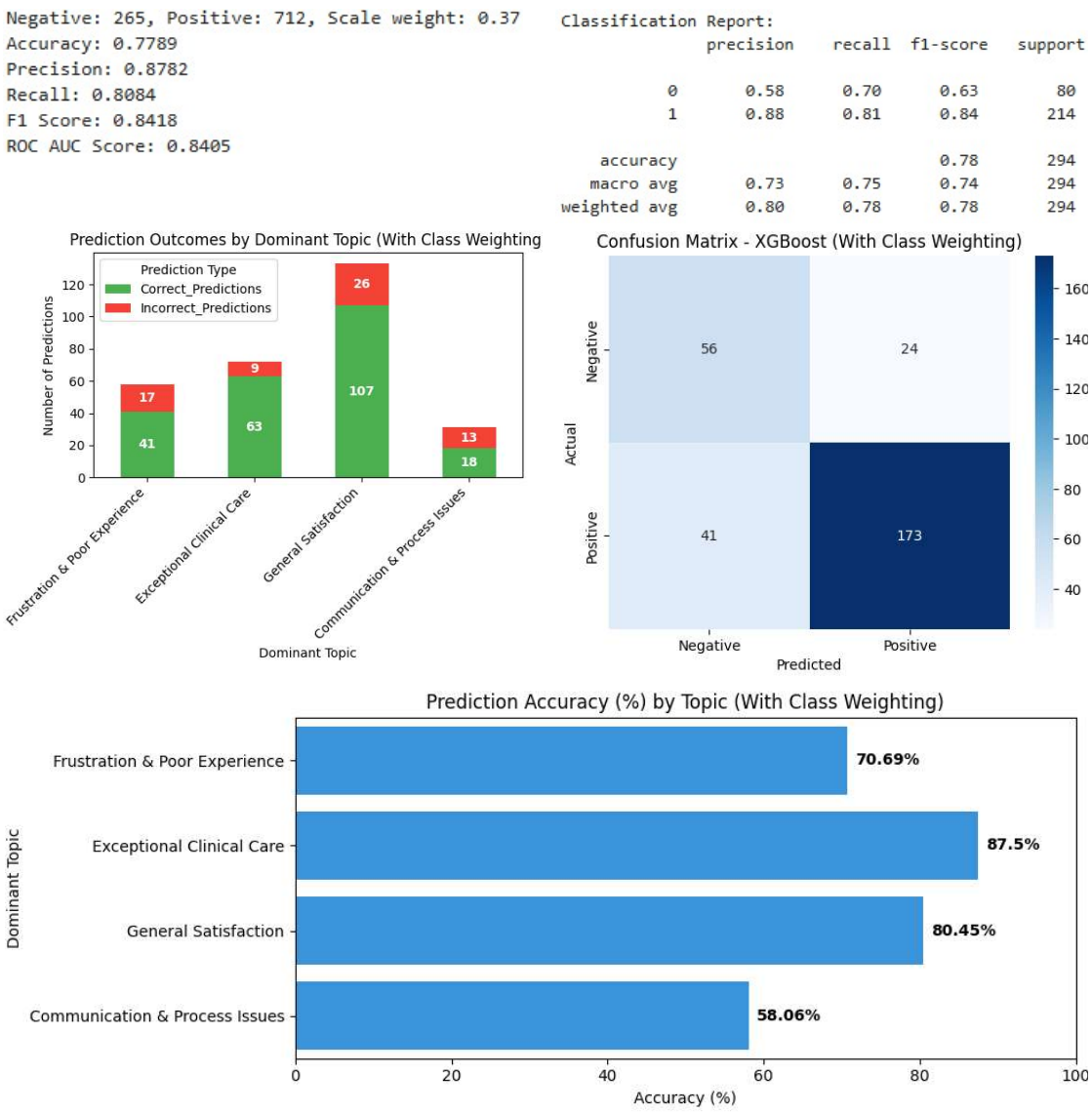

Figure 13: Results from XGBoost + TF-IDF Model with Class Weighting

Sentiment Analysis Model Retraining with Class Weighting

In this version of the sentiment analysis model, class weighting was introduced to address the imbalance between positive and negative feedback in the dataset. The dataset was found to have significantly more positive reviews than negative ones, which can bias the model toward overpredicting the majority class as was seen in the previous version. To compensate for this, the *scale_pos_weight* parameter in the XGBoost classifier was set based on the ratio of negative to positive samples. Specifically, this value was calculated using the formula:

$$scale\ weight\ = \frac{negative\ class\ count}{positive\ class\ count}$$

This adjustment rebalances the loss function during training, encouraging the model to place greater emphasis on correctly identifying the minority (negative) class. The goal of this step is not necessarily to increase overall accuracy, but to produce a model that is fairer and more attentive to underrepresented negative feedback, which is often crucial in applications like healthcare review analysis.

Model Evaluation (Implemented Class Weighting)

After applying class weighting to address the class imbalance (265 negative vs. 712 positive reviews), the model achieved 77.9% accuracy, 87.8% precision, 80.8% recall, 84.2% F1 score, and an ROC AUC of 84.1%. Compared to the unweighted model (80.6% accuracy), this approach improved the model's sensitivity to negative feedback, recall for the negative class increased from 54% to 70%, and true negatives rose from 43 to 56. However, this came with a trade-off: true positives decreased, and overall accuracy slightly dropped.

Topic-level accuracy declined slightly across the board, most notably for Topic 3 ("Communication & Process Issues"), which dropped from 71% to 58%, indicating that weighting may have introduced more confusion in nuanced or mixed-feedback cases.

In summary, while class weighting improved fairness by reducing false positives and enhancing detection of negative sentiment, it slightly compromised general performance. In the case where we need to perform this type of analysis on feedback from Hello Peter, we have ratings to work with, and thus the next step that will be taken to improve performance will be to include the Ratings feature in the model.

Model Retraining with Ratings Feature Included

```
Negative: 265, Positive: 712, Scale weight: 0.37
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
ROC AUC Score: 1.0
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	80
1	1.00	1.00	1.00	214
accuracy			1.00	294
macro avg	1.00	1.00	1.00	294
weighted avg	1.00	1.00	1.00	294

Figure 14: Results from XGBoost + TF-IDF + Ratings with Class Weighting

Incorporating Ratings to Improve Model Accuracy

Following the previous stages of model training using only text data, the next step in improving predictive performance was to include an additional feature - the numerical Ratings field. This decision was based on the goal of aligning the model with the real-world context of the Hello Peter platform, which includes both textual feedback and user-provided star ratings. Since this proof-of-concept dataset uses Google Maps reviews, which also include numerical ratings, it made practical and methodological sense to leverage that information during model training.

The ratings field provides valuable sentiment signals that complement the textual content. By combining TF-IDF features from the cleaned feedback with the corresponding rating values, the model can better capture the nuanced relationship between numerical satisfaction levels and expressed sentiment. This integrated approach allows the model to generalise more effectively, especially for future application to Hello Peter reviews where both text and ratings are available. This retraining step marks a deliberate shift toward building a more context-aware and accurate sentiment classifier, while still addressing class imbalance through weighted learning.

Model Evaluation (Implemented Ratings Feature)

After incorporating the ratings feature into the sentiment analysis model, the classifier achieved perfect scores across all evaluation metrics (Accuracy, Precision, Recall, F1 Score, and ROC

AUC = 1.0). While this might seem like an ideal outcome, such flawless performance is extremely rare in real-world data science applications, especially with noisy, subjective data like customer reviews. This raises valid concerns about potential overfitting or feature dominance and in this case, the possibility that the ratings field alone is driving the predictions, rather than the textual feedback.

Given the very high correlation typically observed between user-provided ratings and sentiment (e.g., low stars often correspond with negative sentiment), it's plausible that the model is relying heavily, or even entirely, on the ratings field to determine the sentiment label. This would undermine the value of the text analysis component and reduce the generalisability of the model, especially when applied to platforms or scenarios where ratings might be missing, inconsistent, or unreliable.

Conclusion: Ratings vs Feedback in Sentiment Classification

To better understand the predictive power of each input, another sentiment analysis model was trained using only the ratings feature. This was done to isolate the impact of ratings on sentiment classification, and to assess whether ratings alone, without textual input, could accurately reflect a reviewer's emotional tone.

The results from the Ratings-only model **(also achieving perfect performance across all evaluation metrics)** highlight a key insight: numerical ratings are highly predictive of sentiment labels, and in this dataset, they may serve as a clearer indicator of a customer's emotional stance than the written feedback itself. While this level of accuracy is unlikely to generalise perfectly across all datasets, it strongly suggests that ratings alone often encapsulate the sentiment behind a review in a way that is easier for machine learning models to interpret.

In contrast, the textual feedback appears to play a more valuable role in uncovering *why* that sentiment exists, as demonstrated in the topic modelling section, where themes like clinical care, wait times, and communication breakdowns were extracted. Thus, feedback is better suited for thematic exploration and categorising concerns, whereas ratings are more efficient for sentiment classification within the bounds of this proof of concept.

These findings support a two-pronged approach to real-world feedback analysis: use *ratings* as the primary driver for automated sentiment detection and rely on *feedback text* for qualitative insights and concern categorisation. In future applications, especially with larger and more diverse datasets, the integration of both inputs may yield more robust sentiment models, but for now, each source has a clearly distinct and complementary strength.

End