PDAN8411 POE

Makabongwe Lwethu Sibisi

ST10145439

Table of Contents

MODEL SELECTION 2

Selected Models: 2
Tool Selection Justification 2
Comparison with Alternatives 2

DATASET SELECTION 3

Data Needed for LDA and Sentiment Analysis: 3
Data Quality Considerations & Common Pitfalls: 3
Dataset Suitability 4

ANALYSIS PLAN: 5

EDA: 5
Feature selection 6
Model Training 7
Interpret and evaluate model 8

CONDUCT YOUR ANALYSIS 9

EVALUATE MODEL 9

WRITE A REPORT 9

BIBLIOGRAPHY 10

Model Selection

Selected Models:

I will use VADER for sentiment analysis and LDA for identifying broad areas of concern. VADER is a rule-based tool that assigns sentiment scores using a built-in lexicon and handles negations, intensifiers, and punctuation. It requires no training data and is optimized for short, informal texts like reviews (Miller, 2024). LDA works by identifying recurring word patterns in the text and grouping them into topics. It performs well on moderate-sized datasets and produces interpretable results in the form of word clusters (GeeksforGeeks, 2025).

Tool Selection Justification

Firstly, the dataset only has around 900 reviews. VADER does not need labelled data or training, so it avoids the risk of overfitting and matches the dataset's size well. Secondly, both VADER and LDA are easy to implement and fast to run. VADER is plug-and-play, and LDA with TF-IDF vectorization allows me to extract topics without a complex training process (Wijono, 2024). Thirdly, both tools output results that are easy to explain to non-technical stakeholders. VADER gives a compound sentiment score between –1 and +1, and LDA provides topics defined by their most frequent words.

Comparison with Alternatives

The dataset size of 900 entries made supervised models like logistic regression or SVM risky due to potential overfitting. These models would demand cross-validation and parameter tuning. Deep learning models such as BERT, despite their linguistic depth, require GPUs, longer training, and extensive tuning, likely offering limited extra benefit for our data. While NMF or BERTopic were alternative topic models, LDA's simplicity and interpretable outputs on smaller datasets made it the preferred choice (Daly, 2023).

Dataset Selection

Data Needed for LDA and Sentiment Analysis:

- Text Data:
  Unstructured, free-form text is required for Latent Dirichlet Allocation (LDA) to extract topics. The "Feedback" column in the hospital reviews dataset provides this, allowing the model to identify recurring themes and issues expressed by patients.

- Sentiment Labels:
  For sentiment analysis, labelled data indicating whether the review is positive, negative, or neutral is ideal. The dataset's "sentiment" column fulfils this requirement, enabling both supervised learning and validation of sentiment predictions.

- Ratings:
  Numeric or categorical ratings, such as the dataset's star ratings, offer an additional quantitative measure of satisfaction. These can be used to segment feedback and correlate with both topics and sentiment, providing deeper insights into patient experiences.

Data Quality Considerations & Common Pitfalls:

When preparing text for LDA and sentiment analysis, high data quality is essential. We need to ensure completeness; every review should have meaningful feedback, sentiment labels, and ratings, as missing fields make analysis unreliable. Consistency in labelling is also critical; misaligned sentiment and star ratings (like "positive" with low stars) create noise and prevent trustworthy results. Don't overlook text cleanliness either; typos, irrelevant content, and extra punctuation distort both topic models and sentiment classifiers, so deduplication, spelling correction, and normalization are necessary. Furthermore, class imbalance (too many positive or negative reviews, for example) can bias models; addressing this through resampling, weighting, or augmentation is crucial for reliable sentiment detection.

Finally, short or vague feedback may lack the richness needed for topic extraction or sentiment, making it necessary to filter reviews to include only informative content. Proactively addressing these issues ensures the analysis generates accurate and actionable insights for the client. (Collibra, 2023)

Dataset Suitability

The hospital reviews dataset is ideal for both LDA and sentiment analysis. The "Feedback" column offers the unstructured text needed for topic modelling. Meanwhile, the "sentiment" and "rating" columns allow for sentiment classification and the segmentation of reviews by satisfaction level. This setup helps separate feedback into positive and negative groups, and it allows for the identification of the most common issues within each rating.

The dataset contains around 900 entries, which is a suitable size for meaningful analysis; it's large enough to create a robust model but not so big that it becomes resource intensive. By thoroughly cleaning and validating the data, I can ensure the results are reliable and directly relevant to the client’s goal of understanding and addressing patient concerns.

Analysis plan:

EDA:

(Adam, 2025)

Feature selection

I will systematically select features through the following steps:

1. **Identify Core Variables**

   I will designate Sentiment Label as the binary target (0=negative, 1=positive), with Feedback as the primary predictor. The Ratings column will serve as a validation feature to check sentiment-label consistency.

2. **Engineer Text Features**

   I will generate clean_feedback through preprocessing: lowercasing, stopword removal, and lemmatization. From this, I'll create TF-IDF vectors using optimal parameters (max_df=0.95, min_df=2) to filter overly common and rare terms. Latent topics will be derived via LDA topic modeling to uncover thematic patterns.

3. **Encode Sentiment Features**

   For VADER analysis, I'll convert raw feedback to sentiment scores, mapping compound scores ≥0.05 to positive sentiment. All sentiment labels will be numerically encoded (negative=0, positive=1) for model compatibility.

4. **Validate Feature Relevance**

   I'll verify feature quality through:

   - Topic-sentiment correlation analysis
   - Rating-sentiment consistency heatmaps
   - TF-IDF vector diagnostics (inspecting key terms per topic)

5. **Optimize for Model Requirements**

   I'll tailor features to each model:

   - LDA: Processed TF-IDF vectors from clean_feedback
   - Logistic Regression: TF-IDF vectors without further filtering
   - VADER: Raw Feedback to leverage its specialized lexicon

6. **Eliminate Redundant Features**

I'll remove uninformative features like the null Unnamed: 3 column and low-variance terms through TF-IDF thresholds.

7. **Finalize Through Validation**

   I'll confirm feature effectiveness by comparing model performance metrics (accuracy, precision, recall) and thematic alignment in visualizations. This ensures features directly support both theme discovery and sentiment classification objectives.

Model Training

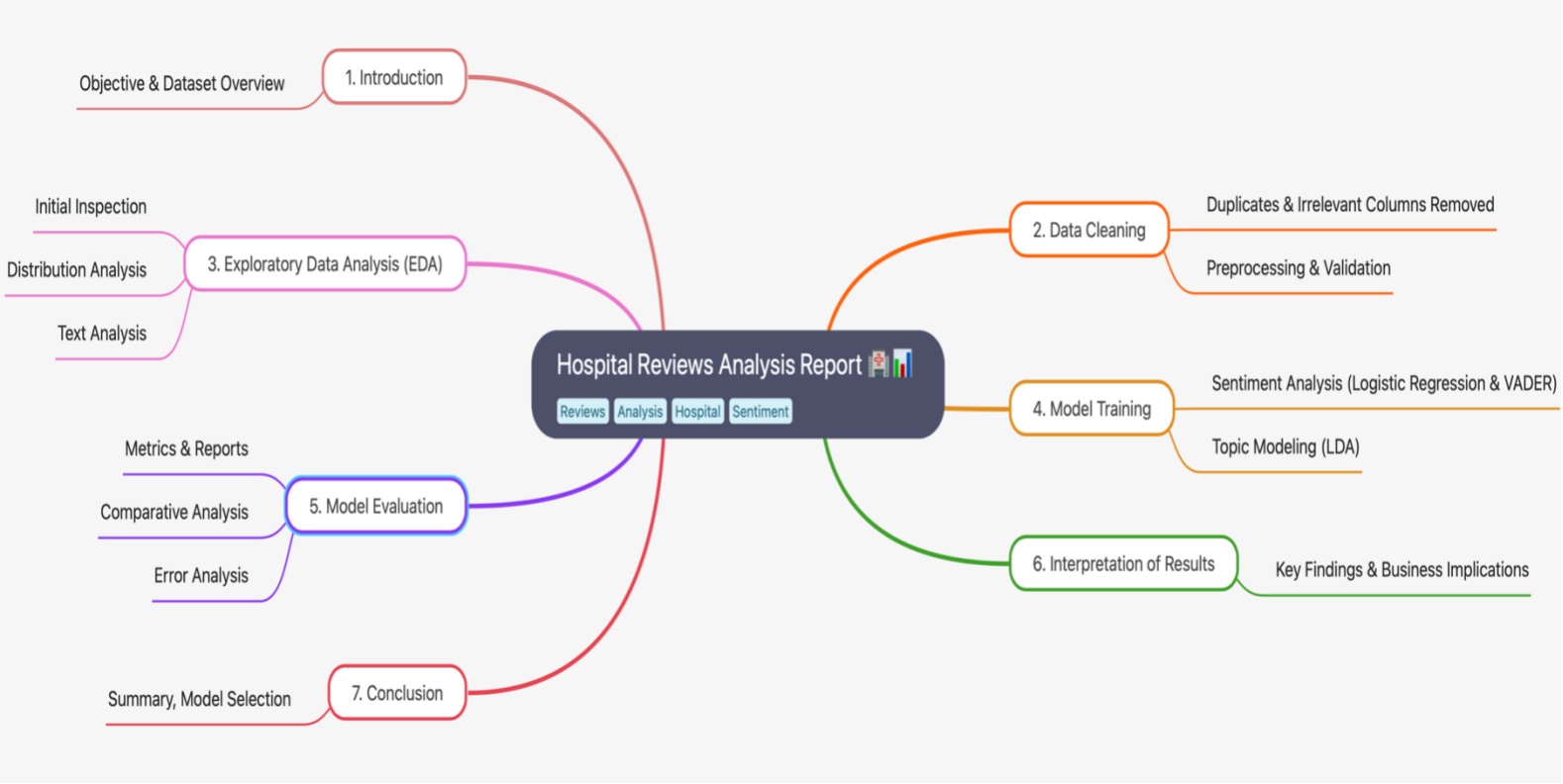
I will implement VADER sentiment analysis without traditional model training since it uses a fixed lexicon and rule-based system rather than learnable parameters. The process involves:

1. **Initialization**
   I will instantiate VADER's pre-configured analyser, which contains validated sentiment intensity mappings for over 7,500 lexical features, including slang and emoticons.
2. **Rule Application**
   For each review, I will generate a compound sentiment score (-1 to +1) using VADER's grammatical rules that account for:
   - Intensity modifiers ("extremely good")
   - Negations ("not helpful")
   - Contrastive conjunctions ("but")
3. **Classification**
   I will map compound scores to sentiment labels using standard thresholds:
   - Positive: score ≥ 0.05
   - Negative: score ≤ -0.05
4. **Validation**
   Finally, I will compare VADER's classifications against pre labelled sentiment to evaluate performance, bypassing conventional train-test splits as the model requires no parameter optimization. This approach leverages VADER's linguistic expertise while eliminating training complexity. (Miller, 2024)

Interpret and evaluate model

I will evaluate model performance using accuracy, precision, recall, and F1-score, with particular emphasis on negative recall since identifying patient complaints is critical for healthcare improvement. I'll generate classification reports and confusion matrices to compare VADER and Logistic Regression, manually analysing misclassifications to detect limitations in handling sarcasm or medical jargon. Performance will be validated against star ratings and topic assignments to ensure business relevance, prioritizing models that achieve >80% negative recall while maintaining overall F1-score >0.85 for actionable insights.

Write a report



Conduct Your Analysis

Please find Reponses within the GitHub repo. The following YouTube videos were used as reference for the code.

https://github.com/VCDN-2025/pdan8411-poe-just-makab.git

3a) (Keith, 2020)

https://www.youtube.com/watch?v=78ut-S-QOEQ&list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4

b) (Science, 2024)

https://youtu.be/aL21Y-u0SRs?si=_AOJbJE3nXgkGYKp

c) (Dylan, 2024)

https://www.youtube.com/watch?v=jB1XMrv_dCA

Evaluate model

https://github.com/VCDN-2025/pdan8411-poe-just-makab.git

Write a report

Please refer to the Report Document.

Bibliography

Collibra, 2023. *The 6 data quality dimensions with examples*. [Online]
Available at: https://www.collibra.com/blog/the-6-dimensions-of-data-quality
[Accessed 26 June 2025].

Miller, I., 2024. *VADER sentiment analysis*. [Online]
Available at: https://hex.tech/templates/sentiment-analysis/vader-sentiment-analysis/
[Accessed 26 June 2025].

GeeksforGeeks, 2025. *Sentiment Analysis using VADER - Using Python*. [Online]
Available at: https://www.geeksforgeeks.org/python/python-sentiment-analysis-using-vader/
[Accessed 26 June 2025].

Wijono, W., 2024. *Practical Guide to Topic Modeling with Latent Dirichlet Allocation (LDA)*. [Online]
Available at: https://medium.com/data-science/practical-guide-to-topic-modeling-with-lda-05cd6b027bdf
[Accessed 26 June 2025].

Daly, Q., 2023. *Step-by-Step NMF Example in Python*. [Online]
Available at: https://medium.com/@quindaly/step-by-step-nmf-example-in-python-9974e38dc9f9
[Accessed 26 June 2025].

Adam, 2025. *Exploratory data analysis to unveil patterns in a car insurance dataset*. [Online]
Available at: https://eyowhite.com/exploratory-data-analysis-to-unveil-patterns-in-a-car-insurance-dataset/
[Accessed 24 April 2025].

Miller, I., 2024. *VADER sentiment analysis*. [Online]
Available at: https://hex.tech/templates/sentiment-analysis/vader-sentiment-analysis/
[Accessed 26 June 2025].

Science, R. &. M. D., 2024. *Hands-On Machine Learning: Logistic Regression with Python and Scikit-Learn*. [Online]
Available at: https://youtu.be/aL21Y-u0SRs?si=_AOJbJE3nXgkGYKp
[Accessed 24 June 2025].

Dylan, D. w., 2024. *Find the Most COMMON Words with Python WordCloud*. [Online]
Available at: https://www.youtube.com/watch?v=jB1XMrv_dCA
[Accessed 24 June 2025].

Keith, M., 2020. *Lwarn Exploratory Data Analysis(EDA) in Python*. [Online]
Available at:
https://www.youtube.com/playlist?list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4
[Accessed 24 June 2025].