Title: Aspect-Based Sentiment Analysis for Hotel Reviews

Project Type: Implementation Project (https://amrita-absa-hotel-reviews.streamlit.app/)

Summary

The goal of this project is to develop an Aspect-Based Sentiment Analysis (ABSA) categorization system that automatically extracts granular metadata from unstructured hotel reviews to improve decision-making transparency. Users currently rely on aggregate star ratings, a simplistic metric that obscures nuances. For example, a hotel might offer excellent service but receive a low rating due to poor location, a trade-off that became especially salient during a recent family hotel booking experience.

While platforms like TripAdvisor and Trip.com surface aspect summaries, these are often inconsistent, focusing on noisy phrases rather than standardized features. In contrast, e-commerce platforms such as Amazon present structured, decision-driving metadata like "battery life", "quality", and "functionality." This project seeks to bring similar standardization to hospitality by parsing reviews to identify high-priority categories and the sentiment associated with each, creating a richer metadata layer that reveals how guests felt about the aspects most relevant to their decisions.

The system leverages a publicly available Kaggle dataset of over 20,000 Booking.com reviews and a bottom-up facet schema spanning Location, Room, Cleanliness, Service, Facilities, Food & Beverage, Price, and Safety. A subset of 100 long reviews is manually annotated at the facet level with a three-way sentiment scheme (positive, neutral, negative) to form a ground-truth baseline. The core implementation is a BERT-based sentence-pair classifier trained in PyTorch on (review, aspect) pairs, deployed as a reusable checkpoint and integrated into a Streamlit web application that supports both single-review and batch CSV analysis.

This project directly applies the course concept of "Computationally-Created Categories," operationalizing faceted classification and user-centric information architecture. The model achieves approximately 67% accuracy with balanced F1 scores, showing that facet-level categories can reliably surface strong positive and negative signals while indicating where the schema, annotation, and data require refinement. The final output is an interactive web app where users can paste a review and see the extracted aspect–sentiment pairs.

Methodology and Implementation

Data and Annotation

This project uses a faceted representation of hotel reviews to surface the dimensions that most strongly shape information seeking and decision-making for travelers. The data come from a Kaggle dataset of over 20,000 Booking.com reviews, each containing text, rating, and metadata. To enable practical annotation, I calculated token counts per review, sorted them by length, and randomly sampled 100 from the top 200 most verbose entries. This subset serves as a gold standard for defining the aspect schema and training and evaluating the model, focusing on rich narratives where trade-offs such as great service but poor location are expressed.

From reading these 100 reviews, I built a bottom-up aspect schema reflecting how guests naturally describe their experiences. The final facets capture the major decision factors, grouped into core categories:

- Location
  - Attractions
  - Public transportation
  - Neighborhood
  - Local convenience
- Room
  - Size
  - Comfort
  - Lighting
  - Layout
  - Furnishing
  - Ambiance
  - Temperature control
  - Noise
  - In-room amenities
- Cleanliness
  - Room
  - Bathroom
  - Common areas
  - Bedding & towels
  - Odor
  - Overall hygiene
- Service
  - Front-desk
  - Friendliness
  - Professionalism
  - Responsiveness
  - Language ability
  - Housekeeping
  - Luggage assistance

    - Concierge
- Facilities
    - Hardware condition
    - Age
    - Elevator access
    - Wheelchair accessibility
    - WiFi
    - Renovations
    - Gym
    - Pool
    - Sauna
- Food & beverage
    - Breakfast
    - Restaurant
    - Room service
    - Variety
    - Dietary inclusive
- Price
    - Value for money
    - Affordability
    - Additional fees
    - Discounts
- Safety
    - Hotel security
    - Neighborhood safety
    - Fire plan

Each facet is annotated using a three-way sentiment scheme: positive, neutral, or negative. This structure strikes a balance between interpretability and expressiveness while remaining feasible for manual labeling. When mixed opinions appear within a facet, I assign the sentiment a typical reader would infer. One limitation of this scheme is that it compresses genuinely mixed experiences into a single label, which a future iteration could address with a dedicated "contradictory" category.

The annotation design applies principles of faceted classification and user-centric information architecture. The facets are intentionally broad enough to cover key decision drivers without overwhelming users. Instead of tagging every phrase, the schema produces digestible summaries like "Service: negative, Location: positive", that convey nuanced trade-offs at a glance. Fixing the facet set enables consistent cross-hotel comparisons and structured queries, such as Which hotels rate high on Safety but low on Service?. These are nsights unattainable through unstructured text alone.

Aspect Based Sentiment Analysis

The project implements an Aspect-Based Sentiment Analysis (ABSA) system that converts unstructured hotel reviews into structured, decision-level insights. A BERT-based sentence-pair classifier associates each review with the standardized facets to impose interpretability.

The training dataset is organized in a long format, where each record corresponds to a (review, aspect) pair labeled positive, neutral, or negative. This structure teaches the model to assess sentiment at the facet level rather than reducing a review to a single label. Preprocessing follows BERT conventions, including Unicode normalization, lowercasing, whitespace cleaning, and control character removal. Review–aspect pairs are encoded using BertTokenizerFast, with the review as the first segment and the aspect name as the second.

Training uses PyTorch’s BertForSequenceClassification with three output classes. Optimization employs AdamW with a linear warm-up–decay schedule, class-weighted cross-entropy to handle imbalance, and gradient clipping for stability.

Throughout development, Gemini AI accelerated debugging by detecting syntax and reference errors in real time, freeing attention for higher-level design work such as learning-rate tuning and rebalancing class weights.

The trained model and tokenizer are released on Hugging Face and integrated into a Streamlit application via the absa_model.py module. The app’s predict_aspect_sentiments() function handles inference: cleaning input, encoding each (review, aspect) pair, and returning sentiment predictions with confidence scores.

This stage operationalizes the concept of computationally-created categories: using machine learning to scale human annotation into structured metadata, making each hotel review not only evaluative but also explanatory showing not just if a hotel was good, but why.

User Interface

The user-facing layer of this project is a Streamlit web application designed to make the hotel review aspect-based sentiment model tangible and easy to explore. The interface is organized into two main views, a “Single review” page for interactive exploration and a “Batch CSV analysis” page for running the model over many reviews at once. On the single-review page, a user can paste a hotel review into a large text box and click “Analyze review” to trigger the ABSA pipeline. The page then displays an “Aspect breakdown” table that surfaces each high-priority category along with the predicted sentiment label and confidence.

The batch view is aimed at more systematic exploration. Here, users can upload a CSV with a review_text column, specify how many rows to process, and then run ABSA across that subset with a single button click. This supports scenarios like quickly scanning sentiment patterns across a large corpus or assisting annotation tasks, while still using the same standardized aspect schema the project is built around.

To build this interface, I relied on AI assistance. I used Gemini to generate an initial Streamlit template based on the following prompt: “Write Streamlit code for a two-tab app with one page for pasting a single hotel review and one page for uploading a CSV of reviews, then display a table of aspect-level sentiment.” The generated scaffold handled the overall layout, tab structure, and file-upload components. From there, I customized the code to better match my preferences, including the dark theme, section titles, button labels, and how the aspect breakdown table is rendered. For someone with essentially no front-end experience, this workflow was extremely helpful even though the template did not fully match my ideal design, it allowed me to start from a working prototype and focus on wiring in the ABSA model and refining the interaction to highlight the decision-making trade-offs that motivated the project.

Evaluation & Future Work

The model's performance was evaluated on a held-out test set of 160 aspect instances using standard classification metrics and a confusion matrix. Overall accuracy reached 66.9%, indicating that the system is able to correctly predict facet-level sentiment for approximately two-thirds of the labeled examples. The macro and weighted F1 scores are closely aligned (0.657 and 0.658 respectively), suggesting that the model's performance is reasonably balanced across classes despite label imbalance.

At the class level, the model performs strongest on positive and negative sentiment. The positive class achieves a precision of 0.689, recall of 0.810, and F1 score of 0.745 (support = 63), while the negative class reaches 0.702 precision, 0.750 recall, and an F1 score of 0.725 (support = 44). These results indicate that when the model predicts positive or negative sentiment, it is typically correct, and it recovers most of the truly positive or negative cases.

In contrast, the neutral class is noticeably weaker, with precision 0.590, recall 0.434, and an F1 score of 0.500 (support = 53). The confusion matrix shows that many truly neutral examples are misclassified as either positive or negative, reflecting the inherent difficulty of distinguishing neutral language from subtly opinionated statements. This pattern is consistent with the small training set and the three-way labeling scheme, where borderline cases are common and often context-dependent.

These results suggest that the model is already useful for highlighting strong positive and negative signals at the aspect level, which supports the project's goal of making decision-relevant trade-offs more transparent. However, they also underscore the need for further improvements.

Limitations and Future Work

The current system demonstrates the feasibility of automated facet-level sentiment classification for hotel reviews but has several limitations that affect scope and generalizability. The most significant is dataset size: with only 100 manually labeled reviews, the model cannot fully capture the linguistic diversity of how guests describe service, cleanliness, location, and other facets. It performs reliably on familiar phrasing but may falter on edge cases, nuanced trade-offs, or uncommon styles. Expanding the annotated corpus to several thousand reviews would likely improve accuracy and robustness, especially for sparse facets like Safety and Facilities.

A second limitation is the three-way sentiment scheme (positive, neutral, negative), which compresses more complex feelings into a single label. Many reviews contain both satisfaction and disappointment within the same aspect, and a future “contradictory” label could better represent these ambivalent experiences.

The use of a single annotator also introduces potential bias, as individual judgments shape tone and boundary decisions. Moving to a multi-annotator setup and reporting inter-annotator agreement (for example, Cohen’s or Fleiss’ Kappa) would strengthen reliability. Looking ahead, scaling the dataset, refining the label scheme, adding collaborative annotation, and using active learning to target ambiguous cases would move the system toward a more robust, domain-adaptive ABSA framework.

Relevant Links:

- Dataset:
  https://www.kaggle.com/datasets/thedevastator/booking-com-hotel-reviews/data
- Final Implementation:
  https://amrita-absa-hotel-reviews.streamlit.app/
- Reference: M. Hoang, O. Alija Bihorac, and J. Rouces, “Aspect-based sentiment analysis using BERT,” in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, Association for Computational Linguistics, 2019.