Chapter 3: Data Science Application Deep Survey

**Data Science Application**

The objective of this project was to identify and implement the three most effective models for extracting insights from hotel survey feedback: (1) predicting guest sentiment, (2) summarizing guest experiences, and (3) generating personalized auto-replies. Given the nuanced, short-form nature of the feedback collected, this challenge required robust Natural Language Processing (NLP) tools capable of interpreting emotional tone, intent, and contextual subtleties, particularly across varied demographics and writing styles. Although classical machine learning approaches—such as logistic regression, quadratic discriminant analysis (QDA), and decision trees—were considered, transformer-based large language models (LLMs) were ultimately selected. These models supported TF IDF, Zero Shot, few-shot and instruction-tuned classification, abstractive summarization, and contextual reply generation aligned with guest sentiment.

The models explored included Unsloth/Mistral-7B-Instruct-bnb-4bit, which was selected for its lightweight inference performance and flexible prompt design. Additionally, Unsloth/Llama-3-8B-Instruct-bnb-4bit was utilized for its enhanced semantic reasoning capabilities and strong instruction-following performance across both classification and summarization tasks. Each model was evaluated using zero-shot and few-shot prompting techniques to assess its accuracy in sentiment prediction. Gemini 1.5 Flash was also considered for its high-quality generative outputs and refined tone-matching during reply generation; however, it was excluded from final testing due to time constraints, financial limitations, and free-tier quota restrictions. Text summarization and auto-reply generation were ultimately conducted using Unsloth/Mistral-7B-Instruct-bnb-4bit. Custom prompt functions were optimized to align responses with predicted sentiment classes, ensuring that replies reflected appropriate levels of empathy, gratitude, or resolution, while summaries effectively distilled key themes from each survey in a concise, business-actionable format.

**Data Overview**

Data for this project were obtained directly from the Hyatt Regency of Green Bay’s HySat platform, a proprietary system designed to collect and manage guest satisfaction surveys. These surveys spanned a comprehensive time frame, from 2018 through May 2025, and provided a rich and representative sample of guest experiences influenced by seasonal variations, operational shifts, and demographic diversity. To construct a usable dataset, multiple survey exports were compiled and standardized, with particular emphasis on open-text fields such as "Additional Feedback on Overall Stay" and general comment sections. Incomplete and duplicate responses were systematically removed to maintain data integrity.

Textual feedback was segmented and labeled using two distinct approaches: a manually annotated sample of 100 surveys and a larger corpus of over 6,600 responses in which sentiment labels were inferred based on the Likelihood to Return (LTR) score. The LTR metric, rated on a scale from 0 to 10, was mapped to predefined sentiment categories using thresholds derived through exploratory data analysis. This mapping facilitated the identification of patterns between guest sentiment and retention behaviors, distinguishing between minor inconveniences and significant service disruptions. The enriched labeling strategy provided deeper insight into the emotional and operational factors that influenced guest loyalty and revealed key areas for improvement. The resulting cleaned and structured dataset formed the basis for NLP-driven tasks including sentiment classification, experience summarization, and auto-reply generation.

**Data Cleaning and Processing**

Data for this project was obtained directly from the Hyatt Regency of Green Bay’s HySat platform, a proprietary system designed to collect and manage guest satisfaction surveys. The surveys spanned a comprehensive period from 2018 through May 2025 and provided a rich, representative sample of guest experiences influenced by seasonal variation, operational transitions, and demographic diversity. To construct a usable dataset, multiple survey exports were merged into a single DataFrame using Pandas and were subsequently standardized, with particular attention given to open-text fields such as “Additional Feedback on Overall Stay” and general comment sections. Responses containing missing textual feedback were removed using the dropna() method. Although duplicate entries were present, no deduplication was performed, as repeated feedback entries were considered potentially informative for downstream sentiment analysis and behavioral modeling.

Text normalization procedures targeted noisy formatting and encoding inconsistencies. Common HTML artifacts, line breaks, and non-ASCII characters introduced during survey input and export were removed. Lowercasing, whitespace stripping, and formal tokenization were not employed, as early experimentation indicated negligible performance improvement when applied to transformer-based LLMs used in this study.

Sentiment labeling followed two strategies. In the manually annotated sample of 100 responses, labels were assigned via qualitative labeling and grouped into five fixed categories: Very Positive, Positive, Neutral, Negative, and Very Negative—reflecting both emotional tone and perceived guest loyalty. For the full corpus of 6,600+ entries, sentiment labels were inferred using the Likelihood to Return (LTR) score. The LTR metric, rated on a scale from 0 to 10, was mapped to sentiment categories using thresholds derived through histogram analysis and business heuristics. The resulting mapping was as follows: “Very Positive” (score = 10), “Positive but Critical” (scores 8–9), “Mixed or Neutral” (score = 7), “Slightly Negative” (scores 5–6), “Negative” (scores 3–4), and “Very Negative” (scores 0–2). This approach facilitated correlation modeling between feedback tone and return intention, helping to distinguish minor inconveniences from impactful service disruptions.

To support NLP model deployment, only relevant columns—specifically “Additional Feedback on Overall Stay,” the LTR score, and the derived sentiment label—were retained. Personally identifiable information such as names, reservation IDs, and contact details was excluded to ensure privacy compliance. A post-labeling class distribution analysis revealed moderate imbalance in the larger dataset, with “Very Positive” responses forming the majority class. In contrast, the manually labeled test sample displayed an inverse distribution, with negative sentiment responses dominating the dataset. Although no resampling techniques were applied to address class imbalance, early trial results indicated that classification performance remained relatively stable. Preliminary comparisons between labeling strategies—one derived from Likelihood to Return (LTR) scores and another based on manual emotional annotation—revealed notable differences in prediction consistency and category sensitivity. These differences were not explored in detail within this chapter, but they informed later analysis and shaped the rationale for the selected modeling approaches. This observation also underscored the importance of aligning sentiment labeling methods with the interpretive goals of the feedback system, whether operational, emotional, or predictive in nature.

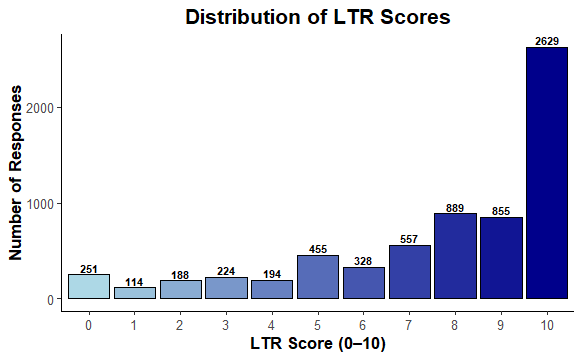
**Exploratory Data Analysis**

Exploratory data analysis (EDA) was conducted to uncover patterns, trends, and potential modeling challenges prior to training the selected NLP models. This process focused on both quantitative response metrics and qualitative textual patterns derived from the HySat survey data.

The distribution of Likelihood to Return (LTR) scores was examined using a histogram, revealing a heavily right-skewed trend within the large dataset. Most respondents rated their likelihood to return between 8 and 10, with scores of 10 comprising the single most frequent response. This pattern aligned with expectations, given that hospitality survey responses tend to lean positive due to loyalty or satisfaction bias.

#### **Figure 1**

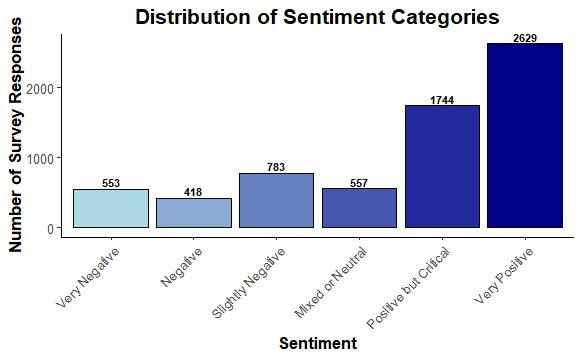
*Distribution of Likelihood to Return (LTR) Scores* This histogram visualizes the distribution of LTR scores across the full dataset, highlighting a strong skew toward higher satisfaction responses.



Sentiment label frequencies were analyzed for both the manually labeled test dataset and the LTR-inferred corpus. In the smaller manual sample, negative sentiments dominated, offering a more critical lens for evaluating classifier performance. In contrast, the larger LTR-based dataset showed a majority of responses classified as “Very Positive” or “Positive but Critical,” highlighting natural class imbalance resulting from numerical rating skew.

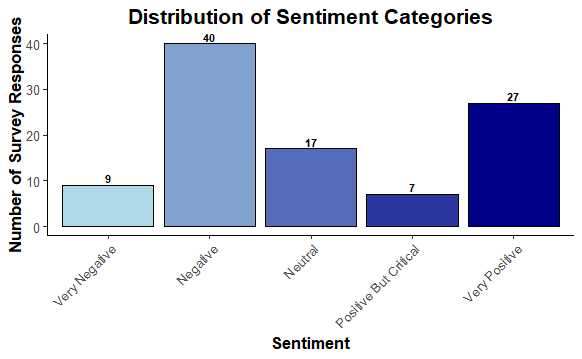
#### **Figure 2**

*Sentiment Category Frequency — LTR-Inferred Dataset* Bar chart representing sentiment labels derived from Likelihood to Return scores across the full corpus, revealing dominant positive categories and class imbalance.

****

#### **Figure 3**

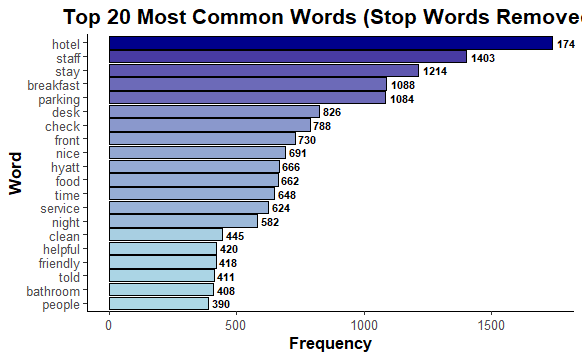
*Sentiment Category Frequency — Manually Labeled Dataset* Bar chart showing sentiment class distributions from the manually annotated test set, which emphasizes negative and mixed responses to evaluate classifier nuance.

****

A keyword frequency analysis was performed to identify common terms associated with guest satisfaction and dissatisfaction. Positive feedback frequently included terms such as "clean," "friendly," "convenient," and "quiet," while negative reviews centered around words like "waited," "dirty," "unpleasant," and "never again." These linguistic insights guided prompt engineering and tone-matching logic for downstream summarization and reply generation.

#### **Figure 4**

*Top 20 Most Common Words in Feedback (Stop Words Removed)* Horizontal bar chart showcasing the most frequent meaningful words used across all survey responses.

****

Demographic or seasonal segmentation was not included in this phase of analysis, as the primary modeling objective was to evaluate language-based sentiment patterns using natural language processing techniques. Since the selected NLP models operated solely on textual input, metadata such as guest age, stay type, or visit date was not required to generate sentiment predictions or summaries.

However, this exclusion does not imply that such metadata lacks analytical value. Prior studies suggest that demographic and seasonal factors may influence sentiment expression, linguistic tone, or feedback length. Integrating metadata-aware models may offer richer subgroup-specific insights or expose hidden bias in classification results. Future iterations may incorporate hybrid approaches that combine structured and unstructured data to further refine predictive accuracy and business relevance.

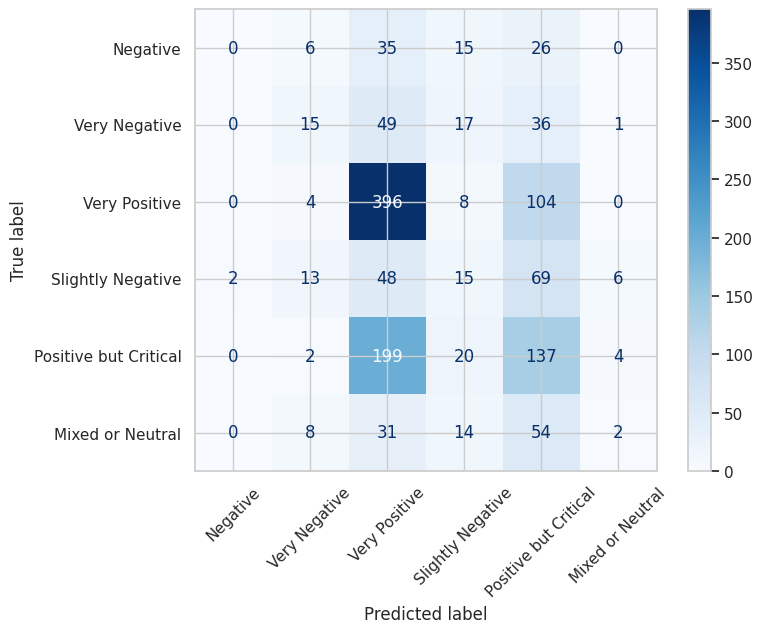
**Model development, Deployment and Evaluation**

The modeling pipeline focused on classifying guest sentiment from survey responses using a combination of traditional machine learning and large language model (LLM) techniques. Two distinct approaches were implemented across both the large LTR-inferred corpus and the smaller manually labeled dataset: a TF-IDF-based logistic regression classifier and prompt-based zero-shot/few-shot classification using instruction-tuned LLMs. Both methods were applied to two datasets — one large corpus inferred from Likelihood to Return (LTR) scores, and one smaller manually annotated test set. For traditional modeling, logistic regression was trained on TF-IDF feature vectors extracted from the preprocessed text. Text cleaning included lowercasing, tokenization, and stopword removal, with hyperparameter tuning focused on optimizing regularization strength and n-gram ranges. Training and validation splits followed an 80/20 structure across all experiments.

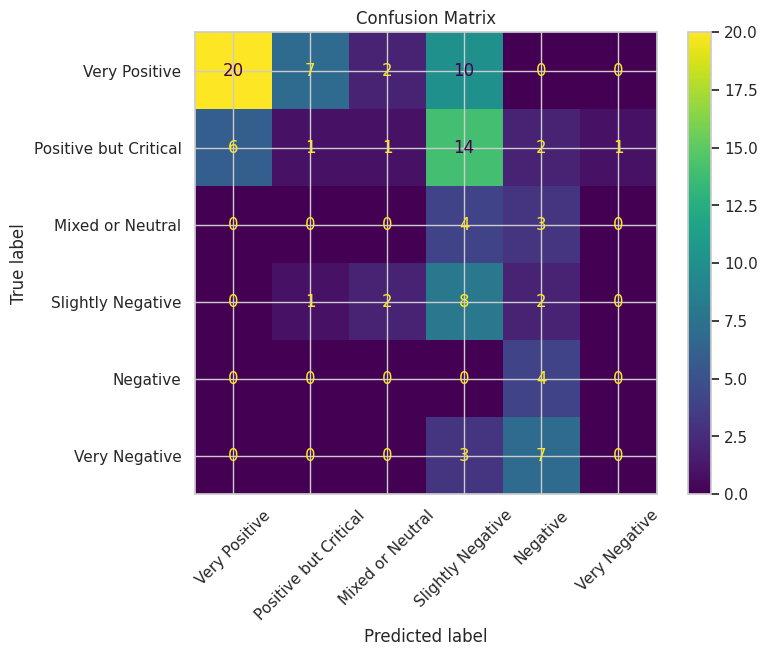
In parallel, multiple configurations of prompt-based LLMs were deployed using quantized models such as LLaMA-3 8B and Mistral-7B via the unsloth framework. Both zero-shot and few-shot prompting strategies were tested. Few-shot prompts were dynamically generated from the training data, and later refined with manually selected examples to improve classification performance on edge cases — especially for nuanced categories like "Positive but Critical" and "Mixed." Prompt engineering emphasized strict class boundaries and caution when interpreting polite but critical language. Hyperparameter tuning for these models included adjustments to batch size, token length, and instruction strength through wording and sentiment definitions.

Models were deployed and evaluated within a cloud-based Google Colab environment, allowing for flexible experimentation across both traditional and large language model architectures. Predictions were generated locally using Python-based pipelines, enabling direct control over input handling, prompt design, and classification logic. These pipelines were intentionally designed to be modular and scalable, supporting future deployment to cloud-based systems such as RESTful APIs or real-time chatbot applications, where guest feedback could be analyzed and responded to automatically. Throughout development, invalid model predictions were logged and excluded, contributing to iterative improvements in prompt engineering. Evaluation metrics included precision, recall, F1-score, and confusion matrix analysis.

**Figure 5**: *Confusion Matrix — TF-IDF Model on LTR-Inferred Dataset*



A baseline TF-IDF logistic regression model trained on the large LTR-inferred corpus achieved a weighted average F1-score of **0.37**, with strongest performance observed in polarized sentiment classes such as “Very Positive” and “Very Negative.” As illustrated in *Figure 4*, the confusion matrix reveals high precision for dominant categories, while mid-range sentiments like “Mixed or Neutral” and “Slightly Negative” were frequently misclassified. These results reflect the model’s sensitivity to class imbalance and its reliance on surface-level lexical cues, which may limit its ability to capture nuanced emotional tone.

**Figure 6**: *Confusion Matrix — Zero-Shot LLM on Manual Dataset* 

In contrast, zero-shot LLMs tested on the LTR-inferred dataset yielded an average F1-score of approximately 0.32, though they exhibited a tendency to overpredict moderate positivity in ambiguous reviews.

Incorporating few-shot prompts with manually selected edge cases and stricter instructional logic led to measurable improvements in precision for mid-range categories such as "Neutral" and "Positive but Critical." The final LLM handled different types of guest feedback evenly — whether positive, negative, or mixed — without getting stuck or making overly confident guesses. It also worked well with the parts of the system that summarize reviews and write thoughtful replies that match the mood of the guest.

In addition to classification, the project included a suite of text summarization models and tone-aware auto-reply generators. For summarization, three approaches were tested: a Hugging Face pre-trained summarizer, a zero-shot prompting model, and a few-shot prompting model that leveraged curated examples to produce more context-sensitive outputs. Each approach was evaluated for its ability to generate concise, emotionally aligned summaries of guest feedback while preserving core themes and hospitality-specific language. To support both summarization and business intelligence efforts, a custom theme extraction function was developed to identify recurring issues within each LTR-derived sentiment category. These thematic clusters provided operational insight and helped focus hotel management’s attention on areas most frequently associated with dissatisfaction or praise. For example, “Mixed or Neutral” responses surfaced themes such as *cold/discomfort/poor quality*, *deterioration*, and *unreasonable fees or poor service*. “Negative” and “Very Negative” sentiments highlighted frustrations like *odor*, *unsatisfactory rooms*, *dirty conditions*, and *poor service quality*, while “Slightly Negative” feedback consistently raised issues with *digital key frustration* and *water-heating inconvenience*. In contrast, “Very Positive” reviews emphasized strengths such as *cleanliness*, *food variety*, and *pool maintenance*.

Auto-reply generation was implemented via a separate prompting pipeline, designed to generate polite, context-aware responses mapped to each sentiment category. This system used a predefined sentiment-to-tone schema to ensure appropriate levels of empathy, assurance, and gratitude. Replies to negative feedback emphasized accountability and support, while responses to positive feedback reinforced loyalty and appreciation. Although sentiment classification results were not directly used to trigger summarization or reply generation, the consistency in tone and theme across components contributed to a unified guest engagement strategy.

**Conclusion**

Spanning all stages, the data science lifecycle, from data integration and cleaning to model development, deployment, and evaluation, laid the foundation for extracting sentiment, identifying thematic patterns, and generating context-aware responses from hotel guest feedback. These methodological components went beyond producing performance metrics; they enabled operational insight, supported scalable engagement strategies, and facilitated automated communication tools aligned with hospitality service goals. With the pipeline implemented, the results are presented in the following chapter.