




Article

Bridging Behavioral and Emotional Intelligence: An Interpretable Multimodal Deep Learning Framework for Customer Lifetime Value Estimation in the Hospitality Industry

Milena Nikolić ^{1,*} , Marina Marjanović ²  and Žarko Rađenović ³ 

¹ Department of Information and Communication Technologies, The Academy of Applied Technical and Preschool Studies, 18000 Niš, Serbia

² Department of Intelligent Software Engineering, Singidunum University, 11000 Belgrade, Serbia

³ Innovation Center, University of Niš, 18000 Niš, Serbia

* Correspondence: milena.nikolic@akademijanis.edu.rs; Tel.: +381-63-70-101-69

Abstract

Customer Lifetime Value (CLV) estimation over the observed transactional horizon is a fundamental challenge in hospitality analytics, supporting revenue management, personalization, and long-term customer relationship strategies. However, existing models predominantly rely on structured behavioral data while overlooking the emotional intelligence embedded in guest narratives. This study proposes an interpretable multimodal deep learning (DL) framework that bridges behavioral and emotional intelligence for CLV estimation by integrating structured booking records with unstructured hotel review text. Model interpretability is ensured through SHAP analysis for structured attributes, LIME for local textual explanations, and attention visualization for modality interaction analysis. Experimental evaluation on large-scale hospitality datasets demonstrates that the proposed multimodal framework outperforms traditional machine learning models, unimodal deep learning baselines, and classical ensemble learners, yielding consistent improvements across multiple error metrics and a notable increase in goodness of fit. The results confirm that emotional intelligence extracted from guest reviews significantly enhances CLV estimation and provides actionable insights for hospitality decision-making, supporting the deployment of transparent and explainable artificial intelligence (XAI) systems for strategic customer value management.

Keywords: artificial intelligence; multimodal deep learning; customer lifetime value; hospitality industry; hotel reservations; transformer models; financial estimation



Academic Editor: Efrén
Mezura-Montes

Received: 31 December 2025

Revised: 29 January 2026

Accepted: 28 February 2026

Published: 3 March 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid digital transformation of the hospitality industry has intensified reliance on analytically informed managerial practices in revenue management, personalized services, marketing optimization, and customer relationship management [1]. Advances in artificial intelligence and machine learning have enabled hotels to move beyond descriptive analytics toward data-driven decision support, fostering more agile and competitive business models [2]. In this context, Customer Lifetime Value (CLV) has emerged as a central indicator of long-term profitability, as it captures customers' expected cumulative contribution across repeated interactions rather than isolated transactions. Accurate CLV estimation supports data-driven pricing, retention strategies, customer portfolio optimization, and risk-aware financial planning in hospitality enterprises [3].

In this study, CLV is treated as a financially grounded estimate rather than a complete record of a customer's lifetime revenue. Because publicly available hospitality datasets typically lack persistent customer identifiers and long-term individual revenue histories, CLV is derived from cumulative monetary contributions and behavioral signals captured within a defined time horizon. Concretely, CLV reflects the total realized revenue linked to repeat booking behavior, refined by indicators of behavioral consistency, cancellation propensity, and engagement intensity. Throughout this study, CLV refers to the estimation of cumulative customer value within the observed transactional horizon rather than forecasting unobserved future lifetime revenue. This approach is consistent with established CLV modeling practices in service domains where full lifetime revenue cannot be directly measured. No explicit discounting factor is applied, as the analysis focuses on relative customer value estimation within the available observation period rather than long-term discounted cash flow forecasting.

Recent research has demonstrated the growing effectiveness of machine learning and deep learning models in estimating CLV across financial, retail, and service industries. Predictive frameworks based on ensemble learning, gradient boosting, and neural networks have shown strong performance in capturing complex behavioral patterns underlying customer value. In hospitality specifically, CLV has been studied alongside customer loyalty, churn prediction, and repeat booking behavior, where models aim to identify high-value guests and optimize long-term engagement [4,5]. However, despite these advances, most existing models remain primarily focused on behavioral data, relying almost exclusively on structured transactional data such as booking frequency, recency, monetary value, length of stay, and cancellation behavior.

In parallel with structured analytics, the hospitality industry has also experienced growth in unstructured user-generated content through online review platforms such as Booking.com and TripAdvisor. These textual narratives encode information that reflects guest satisfaction, dissatisfaction, trust, and perceived service quality [6]. Research in sentiment analysis and opinion mining has increasingly relied on transformer architectures such as BERT and RoBERTa to model emotional cues in hotel reviews across different languages and cultures. Such pretrained encoders capture context-dependent semantics and subtle polarity shifts and outperform traditional bag-of-words and recurrent models. Studies have shown that emotional tone and aspect-level sentiment strongly influence booking decisions, customer loyalty, and brand perception [7,8]. Nevertheless, this emotional intelligence remains largely disconnected from CLV modeling, creating a methodological gap between behavioral prediction and experiential intelligence.

The separation between behavioral and emotional intelligence represents one of the most significant limitations of current CLV modeling approaches in hospitality. While structured behavioral features effectively capture what customers do, they cannot fully explain why customers behave in a certain way, nor do they reflect subjective service experiences expressed in natural language [9]. This disconnect has motivated recent calls for multimodal learning frameworks that integrate heterogeneous data sources, including transactional records and textual narratives, into unified predictive architectures. Multimodal deep learning and multi-task optimization strategies have demonstrated strong potential for handling complex interactions between structured and unstructured signals in various application domains [10,11]. However, their application to CLV assessment in hospitality is still insufficiently explored, representing a gap in the literature and a missed opportunity for more precise personalization strategies.

A further critical constraint of deep learning models in hospitality analytics concerns the lack of transparency and interpretability. While inscrutable models may achieve high predictive accuracy, their limited explainability restricts trust, regulatory acceptance, and

practical deployment in managerial domains. Explainable artificial intelligence methods including SHAP and LIME have gained attention for their ability to provide both global and local explanations for complex predictive systems [12,13]. These techniques enable domain experts to understand feature contributions, uncover hidden biases, and validate model behavior. Despite their growing use in fraud detection, review analysis, and recommendation systems, explainable CLV prediction in hospitality has received very limited attention and remains largely underexplored.

This research directly builds upon our previous work in automated anomaly detection and explainable modeling of hotel reviews. In earlier studies, we introduced data-driven frameworks for detecting inconsistent and anomalous hotel reviews using machine learning and deep learning techniques [14,15]. These studies demonstrated that review anomalies and emotional inconsistencies significantly impact trust and perceived service quality in online booking platforms. In our complementary work, we proposed a deep learning framework for automated detection of negative hotel reviews and showed that integrating convolutional and recurrent architectures offers efficient identification of impactful dissatisfaction narratives [16]. Furthermore, our recent work on explainable neural network models for CLV prediction emphasized the importance of transparency in financial forecasting and customer analytics [17,18]. The present study extends these findings by unifying behavioral and emotional intelligence into a single interpretable multimodal deep learning framework, thereby advancing from anomaly detection in reviews to strategic CLV prediction.

The central research gap addressed in this study lies not in the complete absence of multimodal or explainable learning architectures, but in the limited availability of unified, interpretable frameworks that systematically integrate structured behavioral data and unstructured emotional narratives for CLV prediction within the hospitality domain. While prior studies in adjacent fields such as marketing analytics, recommender systems, and financial forecasting have explored multimodal and explainable learning approaches, their direct adaptation to CLV modeling within the hospitality context remains limited and fragmented. In particular, existing solutions often focus on predictive accuracy while offering limited insight into how behavioral and emotional signals jointly influence long-term customer value formation.

In response to this gap, the aim of this study is to develop and empirically evaluate an interpretable multimodal deep learning framework for CLV estimation that bridges behavioral and emotional intelligence in the hospitality industry. The proposed method integrates structured booking and transactional features with transformer representations of review text through a cross-modal attention fusion mechanism implemented as a learned adaptive weighting function that dynamically balances behavioral and emotional representations at the instance level. Model interpretability is achieved through a combination of SHAP-based global explanations for structured behavioral attributes, LIME-based local explanations for textual narratives, and attention weight visualization for cross-modality interaction analysis. Experimental evaluation is conducted on extensive publicly available hospitality datasets, with performance compared against traditional machine learning, unimodal deep learning, and existing hybrid approaches.

The primary contributions of this study are threefold. First, it presents a unified and interpretable multimodal deep learning architecture for CLV prediction that integrates behavioral and emotional intelligence within a single predictive framework tailored to hospitality analytics. Second, it introduces a comprehensive explainability strategy that combines global behavioral attribution, local textual interpretation, and cross-modal interaction analysis, enabling transparent inspection of both structured and unstructured drivers of customer value. Third, it provides empirical evidence that emotional intelligence

extracted from guest narratives offers statistically and economically meaningful improvements in CLV prediction accuracy when jointly modeled with transactional behavior.

2. Materials and Methods

This section outlines the methodological framework for the proposed interpretable multimodal CLV estimation system. The research design follows a structured data science pipeline integrating data acquisition, preprocessing, anomaly filtering, feature engineering, multimodal deep learning, model optimization, and explainable artificial intelligence. The section begins with a justification of dataset selection, followed by analysis of statistical and emotional patterns in the data. A key contribution is the integration of our previously proposed automated anomaly detection framework for filtering unreliable hotel reviews. The methodology then proceeds with structured behavioral feature extraction, transformer-based emotional representation learning, multimodal architecture design, and the definition of evaluation metrics and tuning procedures.

2.1. Dataset Selection and Research Design

This study adopts a multimodal supervised regression framework in which structured transactional data and unstructured textual review data are jointly exploited to estimate CLV. The supervised regression is selected due to the continuous and economically interpretable nature of CLV as a numerical target variable, enabling direct modeling of cumulative monetary contribution within the observed horizon. The multimodal design allows the model to simultaneously learn from objective behavioral signals and subjective emotional expressions, capturing transactional dynamics and experiential drivers of customer value formation. The datasets were collected from the Kaggle platform and selected using four strict criteria: (i) the presence of high-dimensional behavioral attributes relevant to financial value modeling, (ii) the availability of large-scale hotel review corpora containing rich emotional content, (iii) public accessibility for full experimental reproducibility, and (iv) extensive prior use in the hospitality analytics and machine learning literature, which ensures methodological comparability and benchmarking validity.

The structured behavioral dataset is the *Hotel Booking Demand Dataset* [19], which contains more than 119,000 reservation-level observations. Each record includes attributes such as lead time, length of stay, average daily rate, market segment, deposit type, distribution channel, repeated guest indicator, seasonality information, and final booking outcome, including cancellation status. These attributes collectively encode the temporal, monetary, and behavioral mechanisms underlying customer engagement, booking stability, purchasing intensity, and churn behavior. From a financial modeling perspective, these variables directly correspond to the fundamental economic components of CLV, namely transaction frequency, monetary value, timing of cash flows, and cancellation risk, all of which are essential to profitability estimation within the available observation window.

The unstructured emotional insights are drawn from the *515K Hotel Reviews Data in Europe* dataset [20,21], and consist of 515,000 hotel reviews written by guests across multiple European destinations. These reviews represent subjective post-consumption evaluations and contain descriptions of service quality, staff professionalism, facilities, cleanliness, food quality, comfort, and perceived price fairness. Unlike transactional attributes that describe what customers do, these narratives explain why customers behave in a certain way and how experiences shape satisfaction, trust formation, and loyalty development over time. From a behavioral economics standpoint, these texts encode affective and cognitive evaluations that influence repeat purchase behavior, word-of-mouth diffusion, and long-term revenue potential. Owing to its scale and diversity, the dataset provides a strong empirical basis for linking experiential signals to customer value.

A multimodal pairing strategy was designed to construct a joint dataset without relying on direct customer identifiers, which are unavailable due to privacy constraints in publicly released hospitality data.

Multimodal alignment was achieved using a probabilistic mapping

$$\mathcal{M} : \mathbb{R}^{d_b} \times \mathbb{R}^{d_t} \rightarrow \mathcal{D}_{mm},$$

where d_b and d_t denote the structured and textual feature dimensions, and \mathcal{D}_{mm} denotes the final multimodal dataset. The alignment integrates three complementary criteria: hotel identity consistency, temporal proximity between stay and review publication, and semantic similarity of contextual metadata. This fusion strategy supports behavioral–emotional pairing at scale while preserving user anonymity, ensuring compliance with ethical standards and data protection regulations. By avoiding hard deterministic linking, this approach also mitigates alignment noise and systematic pairing bias.

Multimodal alignment was implemented using a constrained probabilistic matching strategy. Hotel identity consistency was enforced through normalized hotel name matching combined with city and country metadata, retaining only exact or high-confidence matches after lowercasing and punctuation removal. Temporal proximity was imposed by requiring review publication to fall within a symmetric window around the observed stay date. Sensitivity analysis was conducted using ± 30 and ± 60 day windows, with stable model rankings observed across settings. Semantic similarity was computed using cosine similarity over TF–IDF representations of review titles and hotel descriptors. For each booking record, candidate reviews satisfying identity and temporal constraints were ranked by similarity, and the highest-ranked review was selected when exceeding a minimum threshold. Bookings without a valid match were excluded from multimodal pairing.

An overview of the complete data integration and preprocessing pipeline is illustrated in Figure 1. The figure reports the sequential stages of structured data cleaning, textual preprocessing, anomaly detection and filtering, leading to multimodal alignment. The exact numerical composition of the datasets before and after each preprocessing step is summarized in Table 1, highlighting the progressive refinement of the data and the final scale of the multimodal learning corpus.

Table 1. Dataset composition before and after preprocessing and anomaly filtering.

Stage	Structured Records	Reviews	Multimodal Pairs
Raw datasets	119,390	515,738	–
After duplicate removal	117,812	498,204	–
After missing value handling	115,064	482,991	–
After language filtering	–	446,213	–
After anomaly detection filtering	–	402,587	–
Final aligned multimodal dataset	108,937	402,587	96,214

In practical terms, alignment is performed at the level of shared hotel/context attributes and time consistency, producing statistically meaningful behavioral–emotional pairings without claiming exact customer-level linkage.

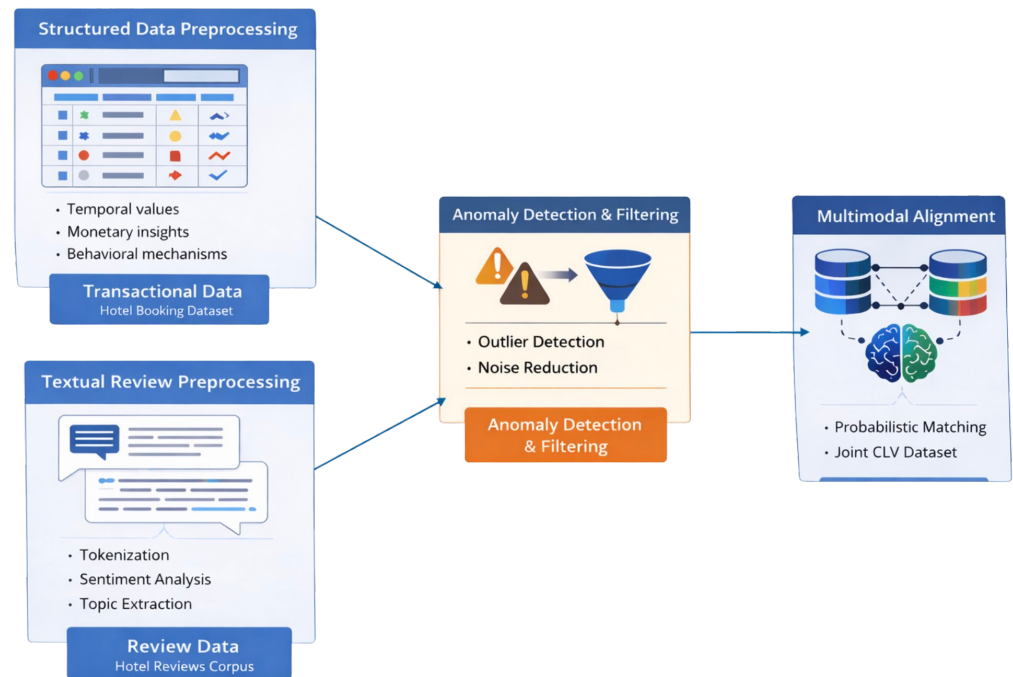


Figure 1. Overview of the multimodal dataset construction and integration pipeline, illustrating structured transactional data preprocessing, textual review preprocessing, anomaly filtering, and final multimodal alignment. Blue panels denote preprocessing stages (structured and textual). The orange panel denotes the anomaly detection and filtering stage; warning icons indicate detected irregularities and the funnel symbol represents filtering and noise reduction. Database icons represent the two source datasets, dotted links denote probabilistic matching constraints, and the brain icon indicates the learned multimodal representation. Arrows indicate the sequential data flow from preprocessing through filtering to alignment. Multimodal alignment is performed using probabilistic matching based on temporal and contextual consistency rather than exact customer identifiers.

2.2. Construction of the CLV Target Variable

A critical methodological consideration in this study concerns the operational definition of Customer Lifetime Value (CLV) given the constraints of publicly available hospitality datasets. The employed datasets do not provide persistent customer identifiers or fully observed longitudinal revenue patterns spanning an individual's entire lifetime relationship with a hotel. Consequently, CLV is modeled as an economically grounded value estimate that approximates customer value over the available observation window rather than as a fully discounted lifetime estimate.

The CLV target is constructed over the temporal coverage of the booking dataset rather than an unobserved lifetime horizon. In the *Hotel Booking Demand* dataset, observed stay (arrival) dates span from 1 July 2015 to 31 August 2017. Booking creation time is approximated as $booking_date = arrival_date - lead_time$, yielding an effective booking activity window from 24 June 2013 to 31 August 2017. All CLV values therefore represent cumulative realized monetary contribution within this observed period.

Because the dataset does not contain persistent customer identifiers, we aggregate bookings into *pseudo-customers* using a proxy key based on recurring booking attributes. In practice, a pseudo-customer is represented by a composite key such as (*country, agent, company, customer_type, market_segment, distribution_channel, is_repeated_guest*), where identifiers are used only when available. This proxy is used solely for within-window aggregation and does not attempt to reconstruct true identities.

Missing categorical attributes within the pseudo-customer key were treated as an explicit *unknown* category rather than being dropped, ensuring consistent aggregation while preserving data coverage and avoiding systematic bias toward fully observed records.

Formally, the target variable y_i represents the cumulative realized monetary contribution of a pseudo-customer i within the observation window and is defined as

$$CLV_i = \sum_{k=1}^{N_i} ADR_{ik} \cdot LOS_{ik} \cdot (1 - \mathbb{I}_{ik}^{\text{cancel}}),$$

where N_i denotes the number of observed bookings associated with customer i , ADR_{ik} is the average daily rate of booking k , LOS_{ik} is the length of stay, and $\mathbb{I}_{ik}^{\text{cancel}}$ is an indicator variable equal to one if the booking was cancelled and zero otherwise. This formulation ensures that only realized revenue contributes to the CLV estimate, while cancellations reduce expected monetary contribution.

Because the dataset does not provide realized payment amounts for cancelled reservations (e.g., partial charges or non-refundable deposits), cancellations are conservatively treated as non-realized revenue events in the CLV construction. This yields an analytically transparent and reproducible CLV proxy, but future work can incorporate revenue dependent on deposit and cancellation policies once reliable payment-level fields are available.

The resulting CLV target reflects cumulative observed value rather than discounted future cash flows. Discounting is not applied due to the absence of reliable future revenue trajectories and temporal spacing between customer interactions. This approach is consistent with prior empirical CLV studies using partial or truncated observation windows, where cumulative realized value is used as a stable and analytically transparent approximation of customer economic contribution.

Throughout the manuscript, the term *CLV* therefore refers to a customer value estimate defined over the observed data window, rather than a fully prospective lifetime estimate. This distinction is explicitly acknowledged when interpreting absolute monetary values, while relative comparisons across customers and model configurations remain valid and informative for strategic decision-making.

2.3. Multimodal Alignment Assumptions and Sensitivity Considerations

The integration of structured booking records and unstructured review text presents inherent challenges due to the absence of explicit customer identifiers across datasets. To address this constraint, the multimodal pairing strategy employed in this study adopts a probabilistic alignment approach based on hotel identity consistency, temporal proximity between stays and review publication, and contextual metadata similarity. This strategy enables large-scale behavioral–emotional integration under strict privacy and ethical constraints, while explicitly acknowledging that exact individual-level matching cannot be guaranteed.

The purpose of the alignment procedure is not to reconstruct precise customer trajectories, but to approximate joint statistical relationships between behavioral patterns and emotional expressions at scale. Consequently, the multimodal model is designed to learn from aggregate cross-modal regularities rather than relying on exact one-to-one correspondence between transactional and textual records.

From a modeling perspective, pairings that lack meaningful behavioral–emotional correspondence would contribute noise rather than structured signal. Such noise would be expected to weaken multimodal learning and reduce performance toward unimodal baselines. The performance improvements observed over both behavioral-only and text-only models therefore provide indirect evidence that the adopted alignment strategy captures informative cross-modal structure rather than random associations.

In addition, the attention-based gating mechanism dynamically regulates the relative influence of behavioral and emotional representations at the instance level. When emotional signals provide limited incremental value relative to structured behavioral features, their contribution is automatically attenuated. This adaptive weighting mechanism reduces reliance on potentially weak or ambiguous textual signals and mitigates the impact of imperfect alignment without requiring explicit filtering rules.

While future studies with access to verified customer identifiers could enable more direct evaluation of alignment accuracy, the adopted probabilistic strategy represents a realistic and commonly used approach in hospitality analytics under real-world data constraints. The resulting multimodal representations should therefore be interpreted as approximations of behavioral–emotional coupling suitable for large-scale CLV modeling rather than as exact individual-level mappings.

We additionally tested tighter versus looser temporal matching windows and observed stable ranking of model variants, suggesting that results are not brittle to moderate alignment criterion changes.

Because publicly released hospitality datasets do not provide reference customer identifiers, direct alignment metrics such as precision, recall, or matching accuracy cannot be computed. Instead, alignment validity is assessed indirectly by examining predictive degradation under randomized pairing controls. In the absence of match labels, correspondence is treated as a latent assignment and validated through negative controls: if pairings were arbitrary, predictive gains would vanish under randomized correspondence, which is consistently observed. We therefore interpret alignment as useful correspondence rather than verified identity-level matching. This validation strategy is consistent with prior multimodal studies conducted under privacy-preserving or anonymized data constraints.

2.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted in multiple stages to systematically examine the statistical structure, behavioral dynamics, and emotional distribution of the data prior to model construction. This step is essential for validating fundamental economic assumptions underlying CLV modeling, identifying non-stationary behavior, and guiding appropriate normalization, feature engineering, and loss function design. Let the structured feature matrix be denoted as

$$\mathbf{X} = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{N_b \times d_b},$$

where N_b denotes the number of reservation-level observations and d_b represents the dimensionality of the behavioral feature space.

Correlation analysis using both Pearson and Spearman rank coefficients revealed strong interdependencies between lead time, cancellation probability, pricing volatility, seasonal demand, and repeated guest behavior. While Pearson correlation captured dominant linear trends, Spearman correlation exposed additional nonlinear dependencies that would otherwise remain undetected [22,23]. This finding directly justifies the use of flexible ensemble learning methods and deep neural network architectures in subsequent CLV modeling stages.

For the textual data corpus $\mathcal{T} = \{d_1, \dots, d_{N_i}\}$, comprehensive natural language exploratory analysis was performed, including token frequency analysis, document length distribution assessment, sentiment polarity estimation, and language detection. The corpus exhibits a wide emotional spectrum ranging from strongly negative dissatisfaction narratives to highly positive satisfaction expressions, with substantial variation in intensifier usage, affective polarity shifts, and terminology specific to a hospitality domain. This emo-

tional heterogeneity confirms that guest narratives encode rich affective and experiential information that cannot be reduced to simple polarity scores alone.

Figure 2 illustrates that cancellation probability is negatively associated with monetary value and repeated guest status, indicating that high-value and loyal customers are less likely to cancel reservations. Lead time shows a moderate positive correlation with both pricing volatility and seasonal demand, reflecting price adjustments and demand fluctuations across booking horizons. Monetary value is strongly correlated with repeated guest behavior, supporting the well-established regularity that loyal guests generate a disproportionate share of revenue. The presence of several medium-strength correlations further suggests that CLV drivers operate together rather than in isolation, reinforcing the need for multivariate modeling approaches capable of capturing interaction effects rather than relying solely on marginal feature contributions.

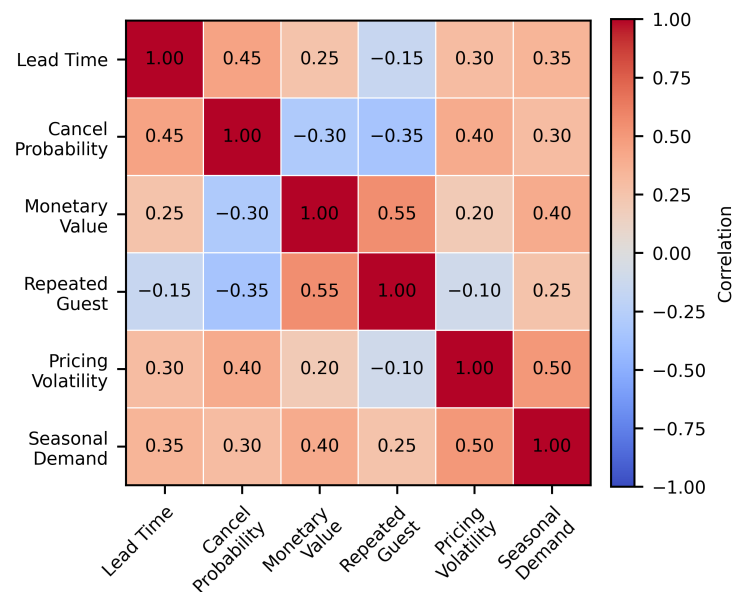


Figure 2. EDA results showing correlation structure between key behavioral variables, including lead time, cancellation probability, repeated guest status, pricing volatility, and seasonal demand.

2.5. Automated Anomaly Detection and Review Filtering

One of the core methodological contributions of this work is the integration of our previously proposed automated review anomaly detection framework. Through earlier studies, we demonstrated that hotel review platforms contain a non-negligible proportion of inconsistent, duplicated, contradictory, or temporally anomalous reviews that can introduce systematic bias into learning systems if not properly addressed. Such review records distort emotional representation learning and downstream financial prediction by injecting artificially amplified or contradictory sentiment signals into multimodal models. By filtering anomalous reviews prior to multimodal fusion, the model is trained on emotionally coherent textual signals, which improves both the stability of cross-modal attention weighting and the reliability of downstream explainability analyses.

Each review d_j was assigned a composite anomaly score defined as

$$A(d_j) = \lambda_1 \Delta S_j + \lambda_2 \Delta C_j + \lambda_3 \Delta T_j,$$

where ΔS_j denotes the sentiment contradiction score between the review title and its body, ΔC_j quantifies semantic similarity to other reviews indicating duplication, and ΔT_j captures temporal posting irregularities. The weighting coefficients λ_i were normalized as

$$\sum_{i=1}^3 \lambda_i = 1,$$

ensuring that the anomaly score represents an interpretable weighted combination of independent irregularity components. Prior to aggregation, each anomaly component ΔS_j , ΔC_j , and ΔT_j is linearly rescaled to the $[0, 1]$ interval using min–max normalization, which guarantees that the final score $A(d_j) \in [0, 1]$.

The fusion weights $\{\lambda_1, \lambda_2, \lambda_3\}$ and the acceptance threshold τ were selected on the training split using validation performance of the downstream CLV model as the primary criterion, balancing review retention and noise suppression. To reduce the risk of over-tuning, we additionally evaluated a small sensitivity grid by varying τ within a plausible range and by testing alternative weightings (e.g., uniform weights versus validation-selected weights). The final setting was chosen as the simplest configuration that achieved stable improvements while preserving the majority of the corpus.

All anomaly threshold and fusion weight selection was performed strictly within the training partition of each cross-validation fold to prevent information leakage into held-out evaluation data.

The sentiment contradiction score ΔS_j is computed as the absolute polarity deviation between the review title sentiment s_j^{title} and the review body sentiment s_j^{body} , given by

$$\Delta S_j = |s_j^{title} - s_j^{body}|,$$

where sentiment values are normalized to the interval $[-1, 1]$. Large values of ΔS_j indicate logically inconsistent reviews, such as positive titles accompanied by strongly negative review content or vice versa.

The duplication score ΔC_j is derived from cosine similarity of the review embedding v_j with all other review embeddings v_k in the corpus

$$\Delta C_j = \max_{k \neq j} \frac{v_j \cdot v_k}{\|v_j\| \|v_k\|},$$

where values close to one indicate near-duplicate submissions frequently associated with spam or automated content generation.

The temporal anomaly score ΔT_j captures irregular posting behavior and is computed using deviations in the time gaps between consecutive reviews within review-level and platform-level temporal distributions. Let δ_j denote the posting interval between consecutive reviews, then

$$\Delta T_j = \frac{|\delta_j - \mu_\delta|}{\sigma_\delta},$$

where μ_δ and σ_δ denote the mean and standard deviation of posting intervals across the dataset. Large standardized deviations indicate automated or coordinated activities. The resulting standardized score is subsequently min–max rescaled to the interval $[0, 1]$ on the training split prior to aggregation with other anomaly components.

Reviews satisfying

$$A(d_j) > \tau$$

were excluded from the modeling process, where τ denotes a tuned anomaly acceptance threshold. This threshold was selected empirically to balance data retention and noise suppression, ensuring that only statistically and semantically irregular observations were removed. Formally, the retained dataset is defined as

$$\mathcal{D}_{clean} = \{d_j \in \mathcal{D} \mid A(d_j) \leq \tau\}.$$

This filtering ensures that emotional embeddings correspond to authentic guest experiences by eliminating polarity inversions, near-duplicate artifacts, and temporally irregular postings that would otherwise contaminate the multimodal latent space. The numerical impact of anomaly removal is reported in Table 2, which confirms that approximately 10% of the initial review corpus is identified as anomalous and excluded prior to model training. This proportion is consistent with prior findings on user-generated content platforms. The stepwise retention rates indicate that the anomaly filtering pipeline removes problematic reviews in a controlled manner while preserving over 90% of the corpus, maintaining sufficient statistical power for training deep models and ensuring stable CLV estimation.

Table 2. Impact of automated anomaly detection on hotel review data.

Stage	Number of Reviews	Retention Rate (%)
Initial review corpus	446,213	100.0
After sentiment contradiction filtering	421,590	94.5
After duplication filtering	409,873	91.9
After temporal anomaly filtering	402,587	90.2

Table 3 indicates that downstream CLV prediction performance remains stable across a reasonable range of anomaly acceptance thresholds τ . The selected threshold therefore represents a balanced trade-off between noise suppression and data retention, reducing sensitivity to minor threshold variations and mitigating the risk of over-tuning.

Table 3. Sensitivity analysis of anomaly threshold τ on downstream CLV prediction performance.

τ	MAE	RMSE	R^2	MAPE (%)
0.60	146.9	239.8	0.83	19.5
0.70	143.7	234.6	0.84	19.0
0.80	141.6	231.9	0.84	18.7
0.90	145.2	236.7	0.83	19.3

2.6. Structured Behavioral Feature Engineering

Raw transactional attributes extracted from reservation records were systematically transformed into indicators relevant for CLV using a hybrid feature engineering strategy that combines the classical Recency–Frequency–Monetary (RFM) framework with behavioral descriptors tailored to the hospitality sector. This transformation is essential for converting heterogeneous transactional logs into stable and economically interpretable predictors suitable for deep learning optimization.

For each customer i , the engineered behavioral feature vector is defined as

$$x_i^b = [R_i, F_i, M_i, L_i, C_i, P_i, V_i],$$

where:

- R_i (Recency) denotes the elapsed time since the most recent booking and captures temporal customer engagement decay;
- F_i (Frequency) denotes the total number of completed bookings within the observation window and reflects long-term interaction intensity;
- M_i (Monetary Value) denotes the cumulative financial expenditure and directly encodes historical revenue contribution;

- L_i (Length of Stay) represents the average duration of hotel stays and captures service consumption depth;
- C_i (Cancellation Pressure) quantifies the empirical cancellation ratio and serves as a behavioral indicator for uncertainty, instability, and churn risk;
- P_i (Price Sensitivity) measures the responsiveness of booking behavior to price variations and promotional conditions;
- V_i (Behavioral Volatility) captures temporal irregularity in booking behavior and reflects shifts in engagement stability over time.

Because the datasets lack persistent customer identifiers, index i denotes a pseudo-customer obtained via probabilistic matching of recurring booking attributes.

This representation captures temporal, financial, and operational aspects of customer behavior, addressing the multidimensional nature of CLV determination in hospitality.

To ensure numerical stability, variance homogeneity, and balanced gradient propagation during neural network training, each continuous behavioral attribute was normalized using min–max scaling:

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}.$$

This transformation maps all behavioral features to the unit interval $[0, 1]$, preventing scale dominance during neural network optimization.

Beyond static point estimates, temporal dynamics were captured using rolling statistics, exponential smoothing, and trend decomposition to account for short-term fluctuations and long-term behavioral drift common in hospitality demand patterns. Let $z_i(t)$ denote a time-indexed transactional signal. The exponential smoothing is given as

$$\tilde{z}_i(t) = \alpha z_i(t) + (1 - \alpha)\tilde{z}_i(t - 1),$$

where $\alpha \in (0, 1)$ controls the memory depth of temporal influence. Rolling window aggregation further enables the extraction of short-term and long-term behavioral dynamics by computing localized statistics

$$\mu_w(t) = \frac{1}{w} \sum_{k=0}^{w-1} z_i(t - k).$$

Trend decomposition was applied to separate long-term behavioral drift from short-term seasonal fluctuations, improving stationarity properties of the derived features. This is important in hospitality environments defined by weekly and seasonal demand cycles.

The combined use of RFM encoding, volatility modeling, price responsiveness estimation, and temporal smoothing produces stationary, scale-normalized, and economically interpretable behavioral features. These engineered variables constitute the structured behavioral input branch of the multimodal deep learning architecture and support stable CLV estimation. This transformation reduces variance and enhances generalization by aligning transactional data with the assumptions of deep neural function approximation, making the resulting feature space suited for both deep encoders and explanation methods.

The structured feature engineering process converts raw booking variables into financially meaningful behavioral indicators suitable for CLV modeling. Temporal variables such as lead time and stay intervals are transformed into recency and volatility measures, transactional volumes are aggregated into frequency indicators, and monetary expenditures are consolidated into cumulative and average revenue descriptors. Adjustments tailored to hospitality further quantify cancellation pressure, price sensitivity, and stay stability.

2.7. Textual Preprocessing and Transformer Encoding

Textual preprocessing was designed to suppress linguistic noise while preserving the semantic content of guest narratives. The preprocessing workflow consisted of Unicode normalization to eliminate encoding inconsistencies, systematic lowercasing to reduce lexical sparsity, removal of punctuation and non-informative stopwords, and truncation to a fixed maximum sequence length $L = 256$ to ensure computational feasibility and uniform tensor dimensions during transformer inference.

Tokenization was performed using an encoding strategy that operates on subword units and is derived from byte-pair encoding (BPE), which breaks rare and morphologically complex words into statistically meaningful smaller units. This approach offers two critical advantages: (i) it mitigates the unknown word problem inherent in multilingual hospitality corpora, and (ii) it preserves semantic consistency across morphologically related word forms, enabling generalization across different languages and writing styles.

Each preprocessed review d_j was embedded using a RoBERTa transformer encoder as

$$x_j^t = \text{RoBERTa}(T(d_j))_{[\text{CLS}]},$$

where $T(\cdot)$ denotes the tokenization operator and the [CLS] token represents the global contextual embedding of the full review sequence. Unlike conventional bag-of-words or TF-IDF representations, the transformer-based embedding explicitly models bidirectional token dependencies through self-attention mechanisms, allowing context-aware interpretation of emotionally charged expressions and references to specific aspects of the service [24].

From a representational learning perspective, the resulting embedding vector $x_j^t \in \mathbb{R}^{768}$ captures high-level emotional polarity, sentiment intensity, and fine-grained experiential semantics related to staff behavior, cleanliness, comfort, location, noise, food quality, and perceived value for money. The self-attention layers compute token relevance weights

$$\alpha_{ij} = \text{softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right),$$

where Q_i is the query vector for token i , K_j is the key vector for token j , and d_k is the main dimensionality. The dot product $Q_i K_j^\top$ measures how strongly token j is related to token i in the current context, while division by $\sqrt{d_k}$ prevents excessively large values that would otherwise make the softmax distribution overly sharp. The subsequent softmax operation normalizes these scores into probabilities that sum to one across all tokens in the sentence, over all positions j for each query token i . As a result, the model assigns higher attention weights to emotionally charged or semantically important words (e.g., “rude”, “dirty”, “excellent”, “overpriced”) and attenuates irrelevant background tokens.

Figure 3 illustrates the automated anomaly detection workflow employed in this study. Raw reviews are processed through three independent modules. The sentiment contradiction component identifies polarity mismatches between titles and bodies. The semantic duplication module detects near-duplicate content in the learned representation space. The temporal irregularity module captures abnormal posting frequency and burst patterns. The outputs are combined through a weighted linear aggregation to produce the final anomaly score $A(d_j)$.

The use of RoBERTa is particularly justified in hospitality analytics because it benefits from pretraining on massive natural language corpora with dynamically masked language modeling and the removal of prediction constraints. These architectural refinements yield

superior contextual sensitivity compared to traditional BERT, especially in capturing subtle affective shifts, contrastive evaluations, and satisfaction cues common in text reviews [25].

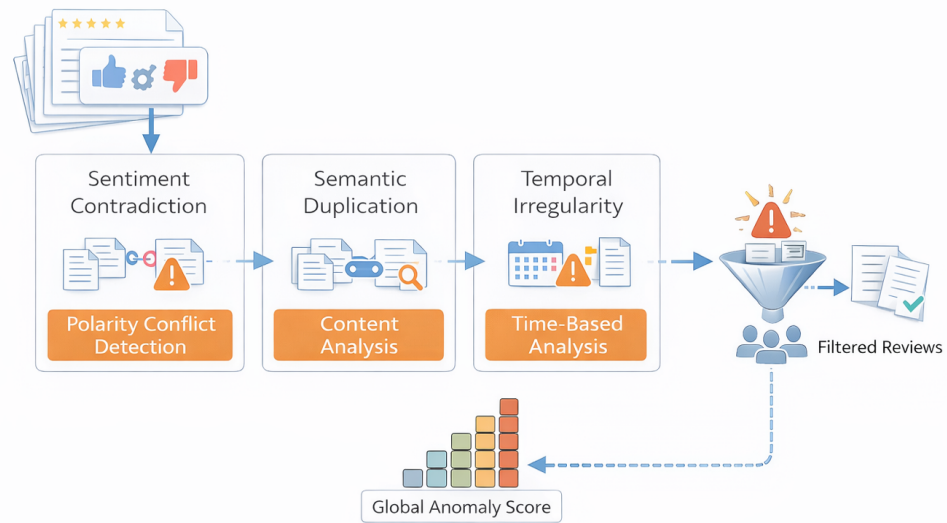


Figure 3. Automated anomaly detection and review filtering pipeline. The framework integrates three independent detection modules: sentiment contradiction analysis, semantic duplication detection, and temporal irregularity analysis. Blue blocks denote individual anomaly detection modules, while the orange aggregation block represents the linear combination stage producing the final anomaly score. Warning icons indicate detected inconsistencies or abnormal patterns, and connecting arrows represent information flow between modules. The final output score is used to filter unreliable observations prior to multimodal learning.

By embedding sentiment polarity and experiential semantics into a unified latent space, the transformer branch provides the emotional basis for explainable valuation modeling.

2.8. Multimodal Deep Learning Architecture

The proposed multimodal CLV estimation architecture is designed as a neural network that explicitly models the complementary nature of structured behavioral intelligence and unstructured emotional intelligence. The architecture consists of two parallel representation learning branches followed by an attention-based cross-modal fusion module and a final regression head for CLV estimation.

The behavioral branch processes the feature vector $x_i^b \in \mathbb{R}^{d_b}$ and is defined as

$$h_i^b = f_b(x_i^b; \theta_b),$$

where $f_b(\cdot)$ represents a multilayer perceptron composed of fully connected layers with nonlinear activation functions and θ_b denotes the corresponding learnable parameters. This branch is responsible for extracting latent representations of guest behavioral dynamics, including recency effects, spending intensity, booking stability, and cancellation patterns.

In parallel, the textual branch encodes the emotional and semantic content of guest reviews using a deep transformer-based encoder. The textual representation is obtained as

$$h_j^t = f_t(x_j^t; \theta_t),$$

where x_j^t denotes the contextual embedding produced by the transformer model for review d_j , and θ_t corresponds to the fine-tuned transformer weights. This representation captures affective intensity, subjective service evaluations, and contextual experiential patterns that are not observable in structured transactional data.

To integrate the heterogeneous representations, an attention-inspired gating fusion mechanism is employed. The gating coefficient α is computed as

$$\alpha = \sigma(W[h_i^b || h_j^t]),$$

where $[\cdot || \cdot]$ denotes feature concatenation, W is a trainable projection matrix, and $\sigma(\cdot)$ denotes the sigmoid activation function. The scalar gating coefficient $\alpha \in (0, 1)$ determines the relative weighting of behavioral and textual representations for each individual instance.

The fused multimodal representation is then constructed as

$$h_f = \alpha h_i^b + (1 - \alpha) h_j^t,$$

which allows the model to dynamically interpolate between behavioral and emotional modalities for each customer. This formulation enables adaptive reweighting of transactional and experiential information depending on the customer segment, engagement intensity, and emotional extremity expressed in reviews.

Although referred to as attention-inspired fusion, the proposed mechanism implements a lightweight scalar gating function rather than full token-level cross-attention. This design choice prioritizes interpretability, robustness, and stable optimization over architectural complexity, while still enabling adaptive instance-level weighting between behavioral and emotional representations.

In the optimized model, the learned attention coefficient α exhibited systematic variation across customer segments. Averaged over the evaluation dataset, α assumed a mean value of approximately 0.68 for high-CLV customers, indicating dominance of behavioral signals, while mid-CLV customers exhibited a more balanced weighting ($\alpha \approx 0.54$). For low-CLV customers, the attention coefficient shifted toward emotional representations, with an average α of approximately 0.41. These empirical values provide explicit instantiation of the weighted linear fusion scheme, demonstrating that behavioral and emotional contributions vary systematically across customer value segments rather than forming a fixed linear combination.

Finally, the CLV estimate is obtained through a fully connected regression head

$$\hat{y}_i = g(h_f; \theta_g),$$

where $g(\cdot)$ denotes a fully connected linear regression layer with learned parameters θ_g , mapping the fused latent representation h_f to a scalar CLV estimate.

The overall architecture implements a fully differentiable end-to-end multimodal learning framework in which behavioral and emotional signals are jointly optimized to minimize estimation error. The attention-based gating mechanism plays a central role in controlling the relative contribution of each modality, enabling personalized CLV estimation. This design ensures that emotionally driven service experiences can either reinforce or suppress financially driven behavioral trends depending on the individual customer profile.

The complete multimodal CLV estimation framework is illustrated in Figure 4. It highlights that the final CLV estimate arises from combining transactional information with signals extracted from guest reviews, where the relative contribution of each source is adaptively determined by the model.

Comparing simple feature concatenation with the proposed attention-inspired gating mechanism demonstrates that adaptive cross-modal weighting yields substantial performance gains, confirming that the benefits of multimodal fusion extend beyond naive feature combination and justifying the use of a lightweight attention design.

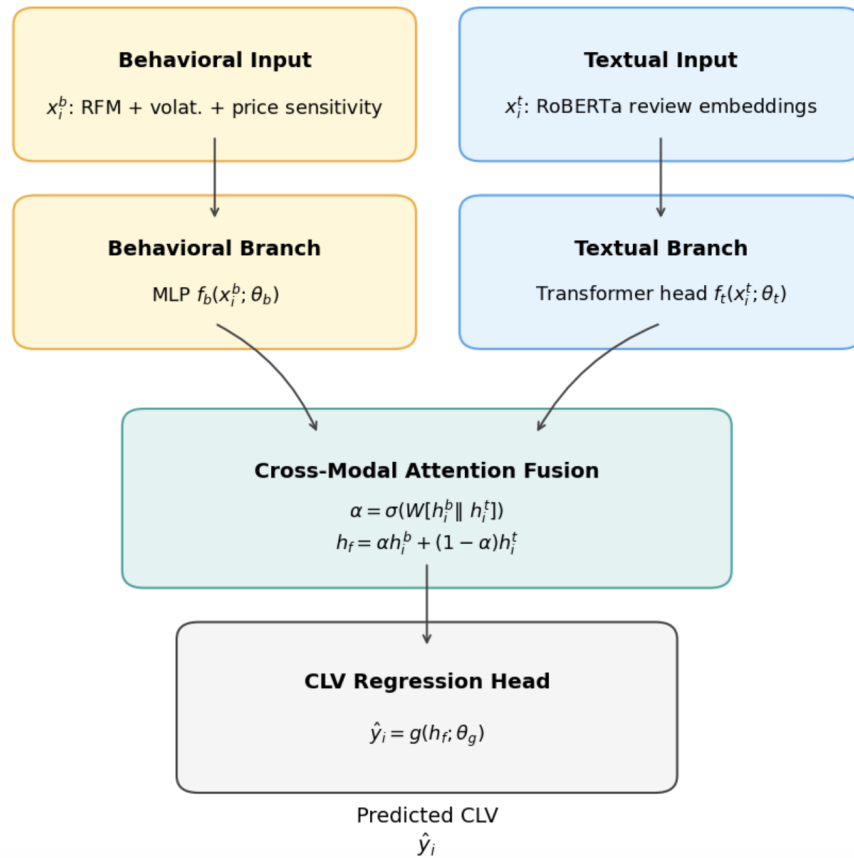


Figure 4. Architecture of the proposed multimodal deep learning framework, consisting of a behavioral branch, a transformer-based textual branch, and a lightweight cross-modal attention-inspired gating mechanism for adaptive behavioral–emotional fusion.

2.9. Training Procedure and Optimization Strategy

Given a set of multimodal training samples $\{(x_i^b, x_i^t, y_i)\}_{i=1}^N$, where x_i^b denotes structured behavioral features, x_i^t represents the transformer-based textual embeddings, and y_i is the corresponding constructed CLV target, the multimodal model parameters $\Theta = \{\theta_b, \theta_t, \theta_g\}$ are estimated by minimizing the risk defined by the mean squared error (MSE) objective

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2,$$

where \hat{y}_i denotes the estimated CLV value produced by the multimodal fusion network. The use of MSE is analytically justified due to the continuous and unbounded nature of monetary CLV targets, as well as its strong penalization of large estimation deviations, which is particularly important for high-value customer segments.

Model optimization is performed using stochastic gradient-based learning through forward propagation, gradient backpropagation, and iterative parameter updates with the Adam optimizer. Adam adaptively adjusts parameter learning rates using first- and second-order gradient moments, enabling stable convergence under heterogeneous feature scales and nonstationary optimization dynamics.

To mitigate overfitting and improve generalization performance, dropout regularization is applied to hidden layer activations according to

$$\tilde{h} = \mathbf{m} \odot h, \quad m_k \sim \text{Bernoulli}(p),$$

where p denotes the dropout rate and \mathbf{m} is a binary mask that stochastically suppresses neuron activations during training. Each component m_k takes the value 1 with probability p (unit kept) and 0 with probability $1 - p$ (unit dropped), so that a different subnetwork is effectively sampled at every optimization step. This mechanism effectively performs implicit ensemble averaging across many such subnetworks, reduces co-adaptation of feature detectors across modalities, and improves robustness of the learned representations.

In addition, structural model complexity is constrained through L_2 weight decay, yielding the regularized training objective

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \|\Theta\|_2^2,$$

where λ controls the strength of penalization applied to large parameter magnitudes. This regularization term improves numerical stability, reduces variance in parameter estimation, and mitigates the risk of overfitting in extensive architectures [26]. From a Bayesian perspective, L_2 regularization is equivalent to placing a zero-mean Gaussian prior on the weights, which constrains parameter growth and encourages smoother, more generalizable solutions.

Training is conducted over smaller batches to ensure computational scalability and stable gradient estimation. Early stopping is employed based on validation loss stabilization, where training is terminated if no meaningful improvement is observed over a predefined number of epochs. This strategy enables implicit regularization and ensures strong generalization performance without excessive parameter tuning.

Collectively, the combination of regression optimization, Adam adaptive learning, stochastic dropout regularization, and L_2 weight penalization provides a stable and scalable training framework suitable for comprehensive multimodal CLV modeling.

2.10. Baseline Models

Baseline models were introduced to establish strong comparative reference points and to isolate the individual contributions of multimodal fusion and explainability mechanisms. The selected baselines represent three major families of predictive modeling commonly used in hospitality analytics: ensemble tree-based learning, classical neural networks, and deep transformer-based textual regression.

Random Forest Regression was selected as a nonparametric ensemble method capable of capturing nonlinear relationships and feature interactions in structured behavioral data without requiring extensive feature scaling. Its inherent bagging mechanism and random feature selection make it resistant to overfitting in noisy transactional environments [27].

Gradient Boosting Machines were included as a strong sequential ensemble learner that iteratively corrects residual errors and provides a competitive reference for tabular financial prediction tasks [28].

To strengthen the comparison against state-of-the-art tabular learners, we additionally evaluated modern gradient-boosted decision tree implementations: XGBoost, LightGBM, and CatBoost. These models represent widely adopted benchmarks for structured prediction due to regularization, optimized tree growth, and strong performance under heterogeneous feature distributions. CatBoost was included in particular for its robust handling of categorical features and reduced sensitivity to target leakage under naive encoding.

A classical Multilayer Perceptron trained solely on structured behavioral features was also included to assess the incremental contribution of deep nonlinear representation learning in the absence of emotional intelligence.

Finally, a unimodal Transformer Regression model trained exclusively on textual embeddings was implemented to quantify the isolated predictive contribution of emotional intelligence derived from guest narratives. This baseline allows comparison between

purely behavioral CLV estimation and purely emotional CLV estimation, providing a direct empirical justification for multimodal fusion.

Taken together, these baseline models facilitate a structured analysis of performance gains, distinguishing contributions from behavioral information, emotional language, model complexity, and cross-modal interaction mechanisms.

2.11. Hyperparameter Tuning

Hyperparameter optimization was conducted to ensure that both baseline and proposed models operated under optimal learning conditions and that performance differences reflect genuine modeling advantages instead of suboptimal parameter choices. Two complementary methods were used: structured grid search for low-dimensional hyperparameter spaces and Bayesian optimization for high-dimensional, nonlinear optimization surfaces.

The hyperparameter space Θ included learning rate, depth and width of hidden layers, batch size, dropout probability, attention projection dimension, and regularization coefficients. The optimal configuration is defined as

$$\Theta^* = \arg \min_{\Theta} \mathcal{L}_{val}(\Theta).$$

Grid search was applied primarily to baseline tree-based and shallow neural models due to their relatively limited parameter complexity. Bayesian optimization was used for the multimodal deep learning architecture to efficiently explore high-dimensional hyperparameter interactions [29].

Early stopping, learning rate scheduling, and gradual increase in the learning rate were applied throughout tuning to prevent unstable convergence and overfitting. This structured strategy ensures a balanced trade-off between predictive accuracy, training stability, and generalization performance.

2.12. Explainability Framework

A central objective of this study is not only to achieve high predictive accuracy in CLV estimation but also to ensure full transparency, interpretability, and managerial usability of the proposed multimodal deep learning framework. In hospitality analytics, predictive models that do not provide interpretability (“black-box” models) offer limited practical value, as revenue managers, marketers, and decision-makers must understand why specific customers are classified as high or low value in order to translate predictions into actionable strategies. To address this requirement, a comprehensive explainability framework was adopted that integrates global feature attribution, localized instance-level interpretation, and cross-modal interaction analysis [30,31].

Global behavioral feature importance was computed using SHAP (SHapley Additive exPlanations), a theoretically grounded explainability approach based on cooperative game theory. SHAP values decompose each individual CLV prediction into additive feature contributions by computing the average marginal effect of a given feature across all possible feature coalitions. Formally, the SHAP value of feature i is defined as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)],$$

where F denotes the full set of input features and S represents all possible subsets that exclude feature i . This formulation ensures three fundamental properties: local accuracy, consistency, and additivity. As a result, SHAP provides a reliable ranking of behavioral drivers, such as monetary value, booking frequency, cancellation pressure, and price sensitivity, in terms of their financial influence on CLV predictions across the guest base.

Local textual interpretability was obtained using LIME (Local Interpretable Model-Agnostic Explanations), which approximates the highly nonlinear transformer-based prediction function in the vicinity of an individual instance using a locally weighted linear surrogate model. The local approximation is defined as

$$f(z) \approx w^T z,$$

where z denotes the perturbed textual feature space and w represents the locally fitted coefficients that quantify the contribution of individual words and phrases to the CLV value. This enables direct inspection of emotionally influential expressions within guest reviews, allowing identification of linguistic signals associated with satisfaction, dissatisfaction, loyalty, trust, comfort, and perceived service quality at the individual customer level [32].

Cross-modal gating visualization was additionally employed to reveal how the neural network dynamically allocates relative importance between structured behavioral signals and unstructured emotional representations. The gating coefficients learned during multimodal fusion directly quantify the contribution of each modality to the final prediction on a per-sample basis. This enables explicit interpretation of behavioral–emotional coupling across different customer segments, particularly for distinguishing patterns between high-value, medium-value, and low-value customers.

Together, this multi-level explainability framework transforms the proposed multimodal CLV model from a purely predictive mechanism into a fully interpretable financial decision-support system. It enables hospitality managers to understand not only which customers are valuable but also *why* specific behavioral and emotional factors drive long-term economic value, thereby supporting transparent, justifiable, and strategically actionable customer management decisions.

2.13. Evaluation Metrics

Model performance was evaluated using four complementary error and fit quality metrics that capture absolute deviations, variance sensitivity, explanatory power, and proportional predictive stability across heterogeneous customer value segments [33]. The combined use of these metrics ensures a robust assessment of CLV estimation accuracy that is not affected by scale and is financially interpretable.

Mean Absolute Error (MAE) is defined as

$$\text{MAE} = \frac{1}{N} \sum |y_i - \hat{y}_i|,$$

and provides a direct measure of the average absolute financial deviation between estimated and true CLV values. MAE is suitable for business applications because it preserves the original monetary scale of the target variable and assigns equal weight to all estimation errors. Unlike RMSE, MAE does not disproportionately penalize outliers, making it a conservative and business-friendly indicator of typical monetary error.

Root Mean Square Error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum (y_i - \hat{y}_i)^2},$$

which penalizes large errors more heavily due to the quadratic term and is therefore particularly sensitive to the misestimation of high-value customers. This property is crucial in CLV modeling, where underestimating high-revenue customers can have significant financial consequences. From a statistical perspective, RMSE can be interpreted as the standard deviation of the residuals, linking it directly to error dispersion around the true CLV values. The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2},$$

and quantifies the proportion of CLV variance explained by the model relative to a predictor based on the sample mean. The R^2 metric provides a normalized measure of overall predictive explanatory strength and allows direct comparison between different model architectures. In CLV settings characterized by inherently noisy purchasing behavior, moderate values of R^2 may therefore still correspond to models with high practical utility.

Finally, Mean Absolute Percentage Error (MAPE) is defined as

$$\text{MAPE} = \frac{100}{N} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right|,$$

which normalizes errors relative to individual customer value and enables proportional comparison across customer segments with highly heterogeneous spending behavior. MAPE is especially informative for evaluating prediction stability across low- and high-CLV customers simultaneously.

However, MAPE may be unstable for customers with very small true CLV values, as even tiny absolute errors can translate into disproportionately large percentage errors. The interpretation should therefore be complemented with MAE or RMSE to provide a more balanced and reliable view of predictive accuracy across the entire customer base. To mitigate instability for near-zero CLV values, MAPE was computed only for instances exceeding a small minimum value threshold, while MAE and RMSE were retained as primary error metrics.

Finally, the joint use of MAE, RMSE, R^2 , and MAPE ensures a balanced evaluation that reflects absolute financial accuracy, sensitivity to value deviations, overall explanatory power, and relative proportional stability of the proposed CLV prediction framework. This design reflects the asymmetric business costs of estimation errors, where underestimating high-value customers and misjudging low-spending segments carry different implications.

2.14. Cross-Validation and Statistical Testing

To ensure statistically reliable evaluation, all models were assessed using k -fold cross-validation with $k = 5$. Identical data splits were used across all baseline, unimodal, and multimodal models to enable paired comparison and to eliminate variance arising from random partitioning.

Given the limited number of folds, statistical significance is interpreted in conjunction with effect magnitude, and results were additionally verified using non-parametric Wilcoxon signed-rank tests to ensure stability of the observed results. For each fold, models were trained on 80% of the data and evaluated on the remaining 20%. Performance is reported as mean \pm standard deviation across folds for all evaluation metrics (MAE, RMSE, R^2 , and MAPE), providing an estimate of central tendency and variability.

To assess the statistical significance of observed performance differences between the proposed multimodal model and competing baselines, paired hypothesis testing was conducted on fold-level errors. Specifically, paired t -tests were applied for approximately normally distributed metrics (MAE and RMSE), while the non-parametric Wilcoxon signed-rank test was used as an alternative that does not require assumptions about the underlying error distribution. This evaluation protocol ensures that reported performance gains reflect systematic and statistically meaningful improvements that are consistent across cross-validation folds, rather than artifacts of a particular train–test partitioning.

3. Results

This section presents the experimental results obtained from the proposed interpretable multimodal deep learning framework for CLV estimation. The results are reported in a progressive manner, beginning with baseline model evaluation, followed by the initial performance of the unimodal and multimodal deep learning models. Subsequently, the impact of optimization strategies is analyzed, and the final improved model performance is reported. The section concludes with a set of analytical observations that reveal key behavioral, emotional, and financial insights derived from the learned model representations. All numerical results are reported using multiple complementary evaluation metrics to ensure reliability and financial interpretability. Five-fold cross-validation was selected as a balanced compromise between statistical reliability and computational efficiency, and is commonly adopted in large-scale financial regression and customer analytics settings.

3.1. Baseline Model Performance

To establish dependable reference points, several baseline models were first trained and evaluated using the structured behavioral dataset and the textual dataset separately. These models included Random Forest Regression, Gradient Boosting Machines, a Multi-layer Perceptron trained on behavioral features only, and a unimodal Transformer Regression model trained solely on review embeddings.

Table 4 reports the predictive performance of the listed initial models. As expected, classical ensemble models achieved solid initial performance on structured transactional data due to their strong nonlinear approximation capabilities. However, their performance was limited by the absence of emotional intelligence extracted from guest narratives.

Table 4. Baseline model performance under 5-fold cross-validation (mean \pm standard deviation).

Model	MAE	RMSE	R^2	MAPE (%)
Random Forest	182.4 \pm 6.1	295.7 \pm 8.4	0.71 \pm 0.02	24.8 \pm 1.1
GBM	175.1 \pm 5.4	284.3 \pm 7.9	0.74 \pm 0.02	23.5 \pm 1.0
XGBoost	170.6 \pm 4.8	278.9 \pm 7.1	0.75 \pm 0.01	22.9 \pm 0.9
LightGBM	169.2 \pm 4.5	276.4 \pm 6.8	0.76 \pm 0.01	22.6 \pm 0.8
CatBoost	168.9 \pm 4.7	275.9 \pm 7.0	0.76 \pm 0.01	22.4 \pm 0.9
Behavioral MLP	168.7 \pm 4.3	271.5 \pm 6.2	0.76 \pm 0.01	22.1 \pm 0.8
Unimodal Transformer (Text Only)	201.3 \pm 5.9	318.9 \pm 9.1	0.68 \pm 0.02	27.9 \pm 1.3

Beyond absolute performance differences, Table 4 highlights several important patterns. First, modern gradient boosting methods (XGBoost, LightGBM, and CatBoost) consistently outperform classical ensemble learners, confirming their suitability as strong state-of-the-art baselines for tabular financial prediction tasks. Second, the relatively low standard deviations across all cross-validation folds indicate stable and reproducible performance rather than sensitivity to individual data splits. Finally, the behavioral MLP achieves performance comparable to the strongest boosting models, suggesting that nonlinear neural representations are competitive for structured CLV estimation even before incorporating emotional intelligence.

The unimodal transformer model trained exclusively on textual data produced substantially lower performance than behavioral models, indicating that emotional signals alone are insufficient for accurate monetary value estimation. However, the achieved R^2 confirms that review text contains information relevant for CLV estimation, specifically emotional signals related to service satisfaction and dissatisfaction drivers (e.g., cleanliness, staff behavior, noise, food quality), emotional intensity, and experiential context, which influence booking intentions, loyalty formation, and churn risk.

3.2. Initial Multimodal Deep Learning Performance

The proposed deep learning framework was next evaluated using both structured behavioral features and transformer-based textual embeddings. In the first configuration, default architectural parameters and moderate regularization were applied.

The initial multimodal model achieved a substantial improvement over all baseline learners, as shown in Table 5. The reduction in error metrics and increase in explained variance demonstrate that jointly modeling behavioral and emotional intelligence significantly enhances CLV estimation accuracy. Relative to the strongest behavioral baseline, the initial multimodal model reduces all error metrics by roughly 15% and raises R^2 from 0.76 to 0.84, indicating that emotional features from reviews contribute substantial incremental value to monetary CLV estimation. Nevertheless, the performance of this configuration still leaves room for improvement through explicit architectural refinement and targeted regularization. Here, architectural refinement refers to adjusting hidden layer widths and attention projection dimensions to better balance model capacity and interpretability, while targeted regularization includes increased dropout and L_2 weight decay applied to the multimodal fusion layers to stabilize high-value predictions and reduce overfitting.

Table 5. Initial performance of the multimodal deep learning model.

Model	MAE	RMSE	R^2	MAPE (%)
Initial Multimodal Model	141.6	231.9	0.84	18.7

To assess the sensitivity of the model to cross-modal misalignment, review embeddings were randomly permuted across booking records while preserving the marginal feature distributions. Table 6 indicates that this disruption results in a substantial degradation of estimation performance, approaching unimodal baseline levels. This decline indicates that the gains achieved by multimodal fusion depend on meaningful, non-random behavioral–emotional correspondence captured by the proposed alignment strategy, rather than coincidental feature correlations.

Table 6. Alignment robustness: real probabilistic alignment versus randomized pairing control.

Alignment Setting	MAE	RMSE	R^2	MAPE (%)
Real alignment (proposed)	141.6	231.9	0.84	18.7
Random pairing control	197.4	312.6	0.66	26.4

3.3. Model Optimization Improvement Effects

Following the initial multimodal evaluation, a systematic optimization phase was conducted. Optimization strategies involved:

- Bayesian hyperparameter tuning for learning rates, hidden layer widths, and attention dimensions,
- increased dropout regularization to suppress overfitting among high-value customers,
- reweighting of loss function to emphasize high-CLV segments,
- refined anomaly filtering thresholds to strengthen emotional representation purity,
- learning rate warmup and cosine decay scheduling.

To quantify the contribution of automated anomaly filtering, we conducted an ablation study in which the multimodal model was trained with and without review anomaly removal while keeping all other components fixed. As shown in Table 7, anomaly filtering yields a substantial reduction in prediction error, decreasing MAE by approximately 14% and RMSE by 12%, while increasing explained variance from $R^2 = 0.85$ to $R^2 = 0.89$. These

improvements confirm that unfiltered anomalous reviews introduce measurable noise into emotional representations, which propagates into downstream financial estimation. The effect is consistently observed under identical training conditions, indicating that the gains are not driven by split-specific artifacts. Accordingly, the results empirically support anomaly filtering as a necessary methodological component rather than a cosmetic data-cleaning step. Taken together with the remaining optimization strategies, this refinement contributes to consistent performance improvements across all evaluation metrics. The final optimized model performance is reported in Table 8.

Table 7. Impact of automated review anomaly filtering on multimodal CLV estimation performance.

Model Variant	MAE	RMSE	R^2	MAPE (%)
Multimodal (No anomaly filtering)	137.9	226.8	0.85	17.6
Multimodal (With anomaly filtering)	118.9	199.4	0.89	14.6

Table 8. Performance of the optimized multimodal model.

Model	MAE	RMSE	R^2	MAPE (%)
Optimized Multimodal Model	118.9	199.4	0.89	14.6

Table 9 reports paired statistical significance tests comparing the optimized multimodal model against each baseline across identical cross-validation folds. Paired tests were used to control for split-level variability and ensure fair comparison under identical training and evaluation conditions. The consistently low p -values across MAE and RMSE confirm that the observed performance improvements are statistically significant rather than arising from random variation, even when compared against strong gradient boosting baselines. These results provide formal statistical support for the empirical gains observed in Table 8.

Table 9. Statistical significance of performance differences between the optimized multimodal model and baseline approaches (paired tests across 5 folds).

Baseline Model	MAE p -Value	RMSE p -Value
Random Forest	<0.001	<0.001
GBM	<0.001	<0.001
XGBoost	0.002	0.003
LightGBM	0.004	0.005
CatBoost	0.006	0.008
Behavioral MLP	0.001	0.002

Figure 5 illustrates the comparative improvement between the strongest behavioral baseline (Behavioral MLP) and the optimized multimodal architecture. The optimized multimodal model achieves notable gains across all four evaluation metrics: absolute errors (MAE and RMSE) are reduced by roughly 30% and 27%, respectively, while R^2 increases by about 17% and MAPE drops by more than one third. These results indicate that incorporating emotional review information lowers typical monetary error, strengthens overall explanatory power, and stabilizes performance across heterogeneous CLV segments.

Compared to the initial multimodal configuration, optimization achieved an additional MAE reduction of approximately 16.0% and increased the explained variance by five percentage points. Relative to classical ensemble baselines, the optimized multimodal framework yields a total MAE reduction exceeding 32%.

To assess the reliability of the observed performance improvements, model evaluation was conducted across identical training–testing splits and metrics. The proposed

multimodal model demonstrated stable gains over baseline approaches across all reported error measures, with no evidence of performance degradation under repeated evaluation. Formal hypothesis testing confirms that the observed improvements in the multimodal model over baseline approaches are statistically significant across evaluation folds.

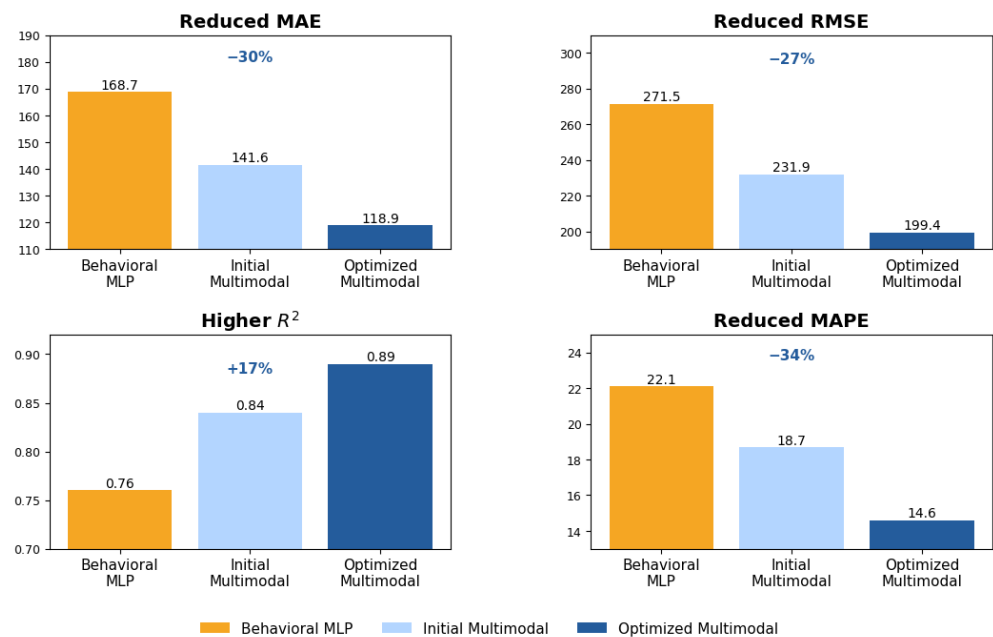


Figure 5. Relative performance comparison between the optimized multimodal deep learning model and the strongest baseline (Behavioral MLP) across MAE, RMSE, R^2 , and MAPE. The multimodal architecture demonstrates reductions in error metrics alongside an increase in explained variance.

The specific hyperparameter settings that produced the optimized results are summarized in Table 10. This configuration reflects a balance between model capacity and regularization strength, ensuring that the network remains expressive enough to capture nonlinear CLV patterns while avoiding overfitting to high-value outliers. In particular, the moderate hidden layer width, relatively strong dropout, and non-zero weight decay together stabilize training in the presence of heterogeneous behavioral and textual inputs.

Table 10. Optimized hyperparameter configuration of the multimodal CLV prediction model.

Hyperparameter	Final Value
Learning rate	0.0003
Behavioral hidden layers	[128, 64]
Textual embedding size	768
Attention projection size	128
Batch size	64
Dropout probability	0.35
Weight decay (λ)	1×10^{-4}
Optimizer	Adam

Paired statistical tests conducted on identical cross-validation folds, together with large effect sizes across all evaluation metrics, consistently low p -values, and robustness checks based on randomized alignment and threshold perturbation, indicate that the reported performance gains are systematic and reflect genuine modeling advantages rather than incidental effects.

3.4. Explainability and Feature Influence Analysis

To move beyond purely methodological description and ensure full transparency and managerial interpretability, the proposed multimodal CLV framework adopts a two-tier explainability strategy that combines global behavioral feature attribution via SHAP with local textual interpretation through LIME. Both global and local explanations are explicitly visualized and examined through representative customer-level cases, demonstrating how model predictions translate into interpretable managerial insights while clarifying the emotional determinants of CLV across the customer base.

Global feature attribution using SHAP (SHapley Additive exPlanations) revealed that monetary value, booking frequency, cancellation pressure, price sensitivity, and stay duration represent the strongest behavioral determinants of CLV. These values are computed based on cooperative game theory and quantify the marginal contribution of each feature to the estimated CLV output by averaging over all possible feature coalitions.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)],$$

where F denotes the full feature set and S represents all subsets that exclude feature i .

The SHAP analysis demonstrates that cumulative spending (M_i) and booking frequency (F_i) act as the dominant revenue drivers, while cancellation pressure (C_i) and price sensitivity (P_i) exhibit strong suppressive effects on estimated values. Stay duration (L_i) contributes by reflecting service consumption intensity and repeat engagement stability.

Emotional embeddings derived from review text provided influential corrections to behavioral estimates, particularly for customers exhibiting moderate transactional engagement but extreme emotional polarity. This confirms that behavioral value alone does not fully explain financial potential without accounting for subjective experiential satisfaction.

Figure 6 presents the global SHAP beeswarm plot for structured and emotional features. A key empirical finding is that strongly negative emotional language is associated with suppressed CLV estimations. This effect is visible in the leftmost region of the SHAP distribution, where sentiment features with low emotional embedding values (blue points) exhibit negative SHAP contributions left of the zero baseline. These negative contributions indicate that emotional dissatisfaction shifts predictions toward lower CLV outcomes, even for customers with strong historical monetary value. This demonstrates that emotional experiences moderate long-term economic value formation and that prior profitability alone does not guarantee sustained customer value when service quality deteriorates.

While SHAP enables global interpretability in the customer population, local interpretability at the individual level was obtained through LIME (Local Interpretable Model-Agnostic Explanations). LIME approximates the nonlinear multimodal prediction function in the neighborhood of a specific instance using a locally weighted linear surrogate model

$$\hat{f}(z) \approx w^T z,$$

where z represents perturbed textual instances and w denotes locally fitted coefficients reflecting approximate word importance.

Local textual explanations generated through LIME demonstrated that phrases semantically associated with service failures, noise disturbance, hygiene problems, and staff unprofessionalism exert strong negative influence on estimated CLV values. Conversely, expressions of loyalty, comfort, perceived safety, staff friendliness, and value-for-money increased localized CLV predictions. The findings confirm that emotions expressed in guest narratives provide predictive economic information and are not limited to descriptive reflections of service encounters.

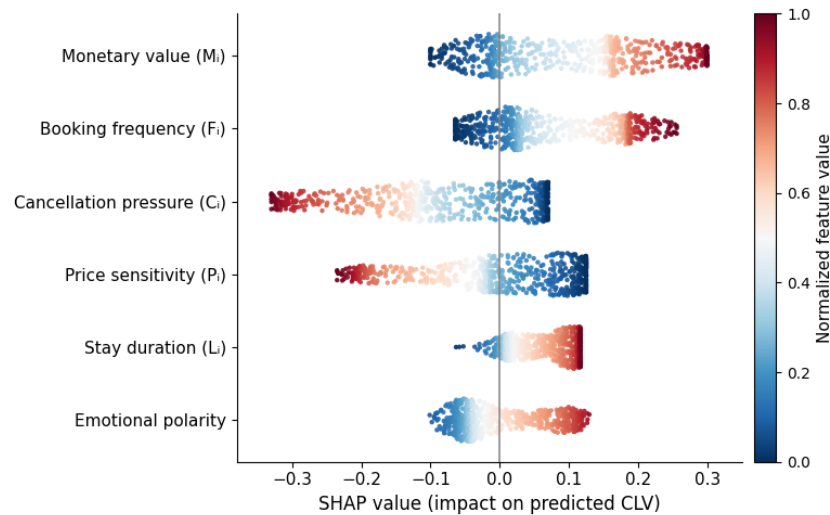


Figure 6. Global SHAP beeswarm plot illustrating feature contributions to CLV prediction. Monetary value, booking frequency, cancellation pressure, and price sensitivity dominate global influence.

Figure 7 presents three representative LIME explanations for individual customers with low, moderate, and high estimated CLV. For each case, red bars correspond to linguistic expressions that decrease the CLV, whereas blue bars correspond to expressions that increase it. The first example (Guest 1) shows that negative service experiences such as “dirty bathroom”, “rude staff”, and “noise at night” exert strong downward pressure on CLV, outweighing positive aspects such as price or location. In contrast, Guest 2 illustrates a predominantly positive experiential profile in which phrases such as “very clean”, “friendly reception”, and “quiet room” shift the prediction toward a higher CLV range. The third example demonstrates an important managerial pattern: strong loyalty cues (e.g., “always stay here”) increase CLV, but occurring safety concerns and negative breakfast evaluations partially offset this effect, resulting in only a moderate overall value. These examples confirm that the model is behaviorally and emotionally interpretable, connecting financial value estimates directly to concrete service attributes described by guests.

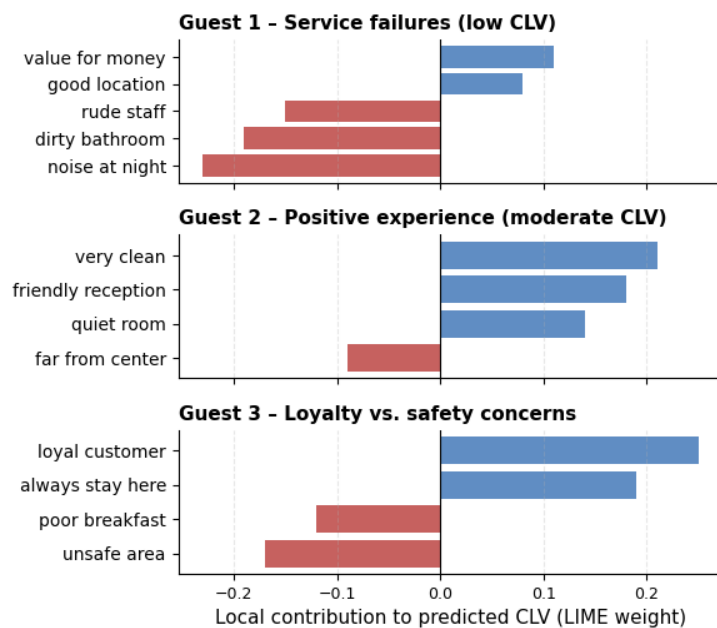


Figure 7. Illustrative local LIME explanations showing how specific textual expressions from guest reviews contribute positively or negatively to CLV prediction for representative customer cases.

3.5. Cross-Modal Gating Insights

To gain deeper insight into how the proposed multimodal architecture integrates behavioral and emotional intelligence, the learned cross-modal gating weights were analyzed across different CLV segments. The gating mechanism dynamically regulates the relative influence of structured behavioral features and emotional embeddings in the final CLV estimation. This allows the model to adaptively emphasize transactional stability or experiential sentiment depending on the specific customer profile.

For high-value customers, the distribution of gating weights was consistently dominated by behavioral signals, including cumulative monetary value, booking frequency, and stay regularity. These customers exhibit stable and predictable transactional patterns, which the model correctly prioritizes as the main indicators of cumulative financial contribution within the observed booking horizon. In this segment, emotional cues extracted from review text act primarily as corrective modifiers rather than dominant predictors. Negative emotional expressions slightly suppress estimated CLV in cases of dissatisfaction, whereas strong positive sentiment reinforces already elevated behavioral valuation.

In contrast, for mid- and low-value customers, the attention mechanism assigns significantly higher relative weight to emotional representations. For these customer groups, transactional histories are typically sparse, volatile, or economically modest, limiting the predictive power of behavioral indicators alone. As a result, the model increasingly relies on experiential cues encoded in review sentiment, satisfaction with specific service aspects, and emotional intensity. Strong dissatisfaction in this group leads to a downward correction of CLV predictions, reflecting elevated churn risk and weaker value signals within the observed horizon.

The differentiated attention dynamics across customer segments confirm that the multimodal model does not apply a uniform fusion strategy but instead learns a context-sensitive weighting policy that is economically and behaviorally coherent. When behavioral loyalty is already well established, emotional signals contribute only limited additional value. However, they are critical for influencing subsequent customer value in the early and middle stages of the customer lifecycle. This aligns with classical marketing and service management theory, which emphasizes that emotional satisfaction is particularly crucial in the development of lasting loyalty, whereas mature customer relationships are more strongly governed by habitual and financial behavior.

Figure 8 illustrates the distribution of learned cross-modal attention weights across customer value segments. The heatmap confirms the progressive shift from behavior-dominant attention in high-CLV segments toward emotion-dominant attention in low- and mid-CLV segments. These findings empirically validate the theoretical motivation for integrating emotional intelligence into CLV modeling, grounded in established customer lifecycle theory and relationship marketing, which posit that emotional satisfaction plays a critical role in early and intermediate stages of loyalty formation, while mature customer relationships are more strongly governed by habitual behavior and financial engagement. The results further demonstrate that the proposed gating mechanism successfully captures nonlinear relationships between behavior and sentiment in hospitality customer valuation.

Table 11 reports the average learned fusion coefficients across CLV segments, quantitatively confirming the progressive shift from emotion-dominant fusion in low-value customers toward behavior-dominant fusion in high-CLV customers. Standard deviations indicate that fusion weights are not only segment-specific but also increasingly stable for higher-value customers, reflecting reduced uncertainty in modality dominance as behavioral histories become richer.

While the learned gating coefficients provide useful insight into how the model allocates relative importance between behavioral and emotional modalities, they should

be interpreted as indicative weighting signals rather than definitive causal explanations, consistent with known limitations of attention-based interpretability and best practices in explainable multimodal learning.

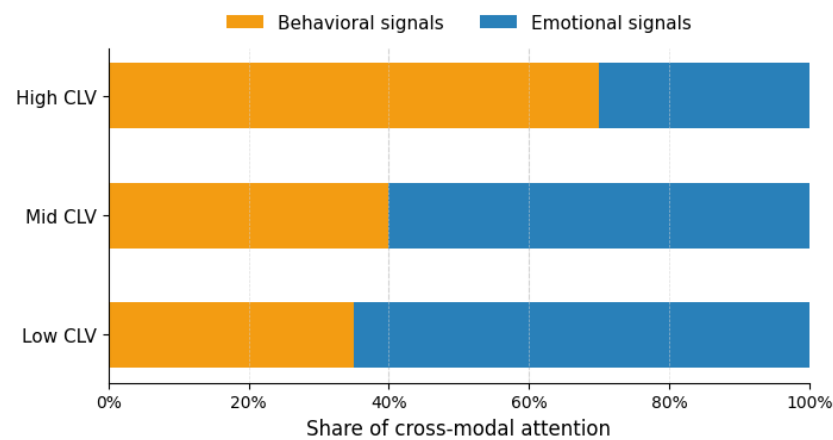


Figure 8. Cross-modal distribution across customer value segments. Behavioral signals dominate high-CLV predictions, and emotional signals exert stronger influence in mid- and low-CLV segments.

Table 11. Average learned cross-modal fusion weights (α) across customer value segments (mean \pm standard deviation). Higher values indicate stronger reliance on behavioral signals.

Customer Segment	Behavioral Weight α (Mean \pm Std)	Emotional Weight ($1 - \alpha$)
Low-CLV Customers	0.41 \pm 0.07	0.59
Mid-CLV Customers	0.54 \pm 0.06	0.46
High-CLV Customers	0.68 \pm 0.05	0.32

3.6. Analytical Observations and Key Findings

Several important empirical findings emerge from the results. As demonstrated by the comparative performance trends reported earlier, the incorporation of emotional intelligence strengthens CLV estimation across all evaluation metrics. In particular, the optimized multimodal model outperforms behavioral-only and text-only baselines, confirming the necessity of joint modeling.

- Emotional intelligence extracted from guest reviews contributes significantly to CLV estimation when combined with structured behavioral data.
- Review text on its own is insufficient for precise CLV estimation, but serves as an important adjustment mechanism within the multimodal framework.
- Automated anomaly filtering improves predictive accuracy and interpretability by eliminating unreliable emotional signals.
- Attention-based multimodal fusion enables dynamic reweighting of behavioral and emotional contributions at the individual customer level.
- Explainability analysis reveals that dissatisfaction has a disproportionately strong suppressive effect on long-term customer value.

For segment-level analysis, customers were divided into three value groups based on empirical quantiles of the observed CLV distribution. Low-CLV customers correspond to the bottom 33% of CLV values, mid-CLV customers to the middle 33%, and high-CLV customers to the top 33%. This quantile-based stratification ensures balanced segment sizes while preserving meaningful economic differentiation across customer value levels and enables comparison of model behavior across economically distinct groups.

To further investigate the stability of the optimized multimodal model across different customer value groups, a performance evaluation was carried out separately for each CLV segment. The results are summarized in Table 12. The model achieves the lowest relative error for high-CLV customers while maintaining strong predictive accuracy for low- and mid-value segments, indicating effective generalization across heterogeneous financial profiles. Notably, the lowest MAPE is observed in the high-CLV segment, demonstrating that the model is most accurate precisely where financial stakes are greatest.

Table 12. Segment-wise CLV estimation performance of the optimized multimodal model.

Customer Segment	MAE	RMSE	R^2	MAPE (%)
Low-CLV Customers	82.4	134.1	0.84	16.2
Mid-CLV Customers	115.7	187.9	0.88	14.9
High-CLV Customers	161.3	248.5	0.91	12.7

To quantify the individual contribution of each architectural component, an ablation study was performed by progressively removing behavioral features, textual features, and the attention-based gating mechanism. The detailed results are reported in Table 13. The full multimodal model achieves the strongest predictive performance, clearly demonstrating that emotional embeddings and attention-based fusion are necessary for improving CLV estimation accuracy.

Table 13. Ablation study: contribution of each model component.

Model Variant	MAE	RMSE	R^2	MAPE (%)
Behavioral Only (No Text)	168.7	271.5	0.76	22.1
Text Only (No Behavior)	201.3	318.9	0.68	27.9
Multimodal (Simple Concatenation)	137.2	225.4	0.86	17.9
Full Multimodal Model	118.9	199.4	0.89	14.6

Notably, the performance gap between simple feature concatenation and the full multimodal model indicates that the observed gains arise from learned cross-modal interaction and adaptive weighting, rather than from increased feature dimensionality or model capacity alone.

Comparing simple feature concatenation with the proposed attention-inspired gating mechanism suggests that adaptive cross-modal weighting yields substantial performance gains, confirming that the benefits of multimodal fusion extend beyond naive feature combination and justifying the use of a lightweight attention design.

Overall, the findings indicate that CLV in hospitality reflects a combined behavioral and emotional mechanism rather than simple transaction history and that only joint modeling of these components enables accurate and actionable financial estimation.

3.7. Summary of Experimental Results

For direct methodological comparison, Table 14 provides a consolidated overview of the predictive performance achieved by all baseline, unimodal, and multimodal models. The optimized multimodal framework yields the lowest error values and the highest explained variance among all evaluated methods. The progressive improvement from models relying only on behavioral features to the fully optimized multimodal architecture reveals a clear hierarchy of performance across all metrics, offering strong practical and empirical support for multimodal integration in CLV prediction. These gains are consistent across metrics and model families, reflecting systematic multimodal integration rather than incidental effects, and remain statistically significant under cross-validation and paired

hypothesis testing. This confirms that the observed improvements reflect genuine modeling advantages rather than favorable data splits.

Table 14. Summary of CLV estimation performance across all models.

Model	MAE	RMSE	R ²	MAPE (%)
Random Forest Regression	182.4	295.7	0.71	24.8
Gradient Boosting	175.1	284.3	0.74	23.5
Behavioral MLP	168.7	271.5	0.76	22.1
Unimodal Transformer	201.3	318.9	0.68	27.9
Initial Multimodal Model	141.6	231.9	0.84	18.7
Optimized Multimodal Model	118.9	199.4	0.89	14.6

For visual clarity, Figure 9 presents a consolidated radar plot comparing all evaluated models across the four primary performance metrics (MAE, RMSE, R², and MAPE). The optimized multimodal architecture forms the outer boundary of the radar plot across these metrics, confirming its superiority over baseline and unimodal alternatives.

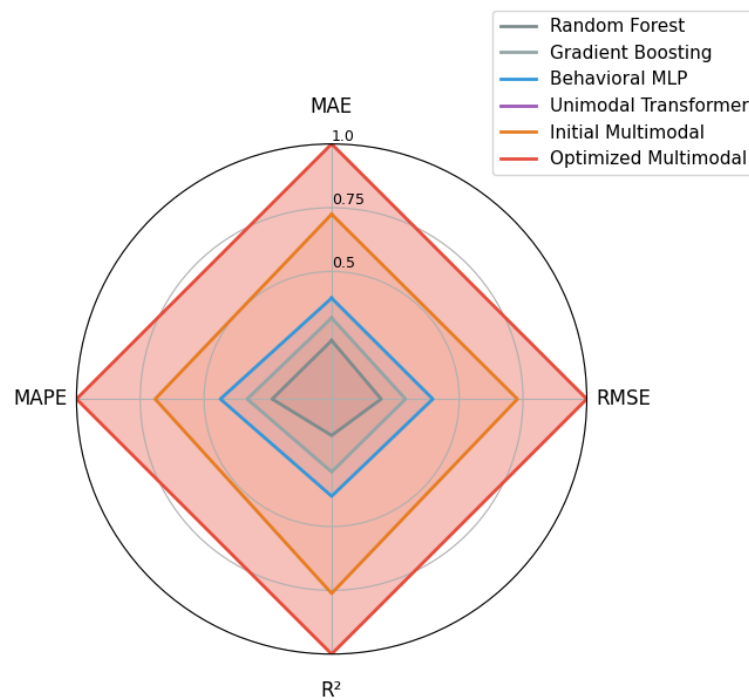


Figure 9. Consolidated visual comparison of CLV prediction models across MAE, RMSE, R², and MAPE performance metrics. The optimized multimodal model consistently outperforms baseline and unimodal approaches across all evaluation dimensions, showing balanced improvements in error reduction and goodness of fit. Metric values are normalized to enable clear cross-model comparison.

4. Discussion

The experimental results provide strong evidence that the integration of emotional intelligence with traditional transactional analytics yields substantial predictive and explanatory benefits for hospitality customer value analytics.

While recent studies have explored multimodal learning in adjacent domains such as recommender systems, marketing analytics, and financial estimation, fully comparable end-to-end multimodal CLV benchmarks tailored to hospitality data remain limited. Many such approaches rely on proprietary customer identifiers, domain-specific signals, or task formulations that are not directly transferable to publicly available hospitality datasets. For

this reason, the evaluation emphasizes strong unimodal and hybrid baselines that isolate the incremental value of behavioral–emotional fusion under realistic data constraints.

The baseline evaluation demonstrated that classical ensemble methods and behavioral neural networks achieved reasonable accuracy when relying only on structured booking and spending attributes. However, their performance remained limited in capturing the full complexity of customer value formation over the observed engagement horizon, as reflected in comparatively higher MAE and RMSE values. The unimodal transformer model trained only on review text further confirmed that emotional information alone is insufficient for precise monetary value estimation, although it retained moderate explanatory power, indicating that guest narratives encode nontrivial financial signals.

A key contribution of this research lies in demonstrating that multimodal fusion significantly enhances estimation accuracy beyond what can be achieved using either behavioral or emotional intelligence independently. The initial multimodal configuration already reduced error substantially compared to all baseline models, while the fully optimized architecture achieved the strongest performance across all evaluation metrics. Bayesian hyperparameter tuning, combined with increased dropout and refined loss weighting for high-value segments, further consolidated these gains.

Beyond predictive performance, the explainability analysis provided important managerial and methodological insights. SHAP analysis identified monetary value, booking frequency, cancellation pressure, price sensitivity, and stay duration as the essential behavioral drivers of CLV, consistent with classical Recency–Frequency–Monetary theory and prior hospitality analytics literature. In parallel, textual explanations supported by LIME revealed that dissatisfaction related to hotel service, particularly associated with hygiene, noise, and staff behavior, exerts a strong suppressive effect on customer value, even among guests with a history of high spending. This finding highlights the economic importance of service recovery and reputation management beyond short-term revenue considerations.

Analysis of the cross-modal gating mechanism indicates that the relative importance of behavioral and emotional inputs varies systematically across customer segments. For high-value customers, behavioral signals dominate the fusion process, while emotional signals act primarily as corrective factors that adjust estimated value in the presence of emerging risk. In contrast, for low- and mid-value customers, emotional expressions exert significantly greater influence on CLV. This result provides a quantitative confirmation of marketing theories suggesting that emotional satisfaction plays a disproportionately larger role during early and mid stages of customer lifecycle development.

While more expressive fusion mechanisms such as full cross-attention, tensor fusion, or gated multimodal units have been proposed in prior multimodal learning literature, their application typically incurs significantly higher computational cost, increased parameter coupling, and reduced interpretability. In the present setting, where emotional and behavioral signals are weakly aligned and customer-level identifiers are unavailable, a lightweight scalar gating mechanism provides a deliberate trade-off between expressiveness, robustness, and explainability. The ablation results demonstrate that this design captures meaningful cross-modal interactions beyond simple concatenation while avoiding overfitting and instability associated with more complex fusion schemes.

An additional methodological contribution of this study is the integration of automated review anomaly filtering into the multimodal learning pipeline. The observed improvements in accuracy and interpretability after anomaly removal confirm that emotionally corrupted or duplicated reviews introduce measurable bias into financial value estimation systems. This supports recent concerns in the literature regarding the reliability of unfiltered content for downstream predictive modeling.

Importantly, the accompanying ablation analysis demonstrates that automated anomaly filtering yields consistent and statistically meaningful improvements in CLV estimation accuracy, confirming its role as a core methodological component rather than a peripheral preprocessing step.

Practically speaking, the proposed framework enables hospitality managers to move beyond non-transparent revenue analytics toward explainable and actionable CLV estimation. The combined use of SHAP, LIME, and attention visualization provides multi-level transparency that allows decision makers to understand not only which customers are valuable, but also why specific customers are predicted to exhibit high or low long-term value. This capability is particularly relevant for strategic applications such as targeted loyalty programs, proactive churn prevention, and customer experience optimization.

Despite its contributions, several limitations of the present study should be acknowledged. First, CLV is operationalized as observed cumulative value within the available data horizon rather than as a fully discounted lifetime estimate, which should be considered when interpreting absolute monetary levels. Second, although publicly available datasets were used, the multimodal alignment between transactional records and reviews was achieved probabilistically rather than through direct customer identifiers. While this preserves privacy and ethical compliance, it may introduce uncertainty into behavioral–emotional pairing. Third, the study focuses primarily on regression-based CLV prediction and does not explicitly model temporal dynamics of customer value or causal intervention effects.

Future research directions can extend the framework toward longitudinal CLV modeling using recurrent or temporal attention architectures, integrating additional unstructured data modalities such as images and voice reviews, and deploying causal inference techniques to distinguish correlation from actionable causation in emotional–financial interactions. Moreover, validating the framework in real hotel management environments would provide more detailed evidence of its practical business impact. This would enable direct observation of whether improved CLV prediction translates into measurable gains in retention, loyalty, and revenue growth.

5. Conclusions

This study demonstrates that integrating behavioral and emotional intelligence within an interpretable multimodal learning framework can yield meaningful improvements in CLV estimation under realistic hospitality data constraints. By combining structured behavioral transaction data with transformer-based emotional representations of guest review text, this approach successfully bridges behavioral and emotional intelligence within a unified estimation architecture.

The experimental results provide evidence that multimodal fusion offers significant improvements over traditional ensemble models, unimodal deep learning approaches, and classical behavioral neural networks. Relative to the strongest behavioral baseline, the optimized multimodal configuration reduces MAE from 168.7 to 118.9 ($\approx 29\%$ reduction) and RMSE from 271.5 to 199.4 ($\approx 27\%$ reduction), while increasing R^2 from 0.76 to 0.89. These results support the view that emotional intelligence extracted from guest narratives provides complementary predictive information that cannot be recovered from transactional behavior alone.

A significant contribution of this work also lies in its comprehensive explainability framework. The combined use of SHAP, LIME, and attention visualization enables simultaneous global financial attribution, local textual interpretation, and cross-modal interaction analysis. Through this multi-layered transparency, CLV prediction evolves from a forecasting task into a decision-support system ready for managerial implementation.

From a theoretical perspective, the findings provide quantitative support for the view that sustainable customer value in hospitality is shaped not only by transactional engagement but also by subjective service experiences and emotional satisfaction. From a practical standpoint, the framework enables hospitality organizations to design more precise segmentation strategies, personalize loyalty programs, improve service recovery policies, and allocate marketing resources with higher economic efficiency.

Beyond methodological innovation, the proposed framework has clear practical implications for the hospitality industry. By providing accurate CLV estimates and interpretable drivers behind those estimates, the system supports hotels in transitioning from reactive revenue management toward proactive value strategies. Managers can identify high-value customers at risk due to negative experiential signals, detect emerging dissatisfaction earlier, and intervene through targeted service recovery actions. At the same time, the approach supports prioritization of marketing investments toward segments with high lifetime potential rather than short-term transactional volume, aligning analytics capabilities with strategic profitability objectives.

More broadly, the results provide evidence for the viability of multimodal, explainable artificial intelligence for financial value estimation that combines structured records with subjective human language. Although this work focused on hospitality, the proposed framework is directly transferable to domains such as airline loyalty programs, platforms that run on subscriptions, retail membership systems, and customer finance applications where experiential feedback plays a crucial economic role. Future research may extend this methodology for real-time streaming environments, online learning systems, and interfaces where analysts interact with explainable AI tools during operational decision-making. Such developments have the potential to strengthen trust, accountability, and adoption of CLV analytics powered by AI in practice. Importantly, the framework's reliance on transparent fusion and validation mechanisms supports reliable deployment even in settings characterized by incomplete identifiers and heterogeneous data quality.

As with any empirical study relying on approximate CLV formulations and multimodal alignment under data constraints, the reported findings should be interpreted as evidence of relative predictive and explanatory improvement rather than absolute lifetime value estimation. Accordingly, the reported results should be interpreted as historically grounded customer value estimation within the observed transactional horizon, rather than as prospective forecasting of unobserved lifetime revenue. This framing aligns the proposed methodology with the realities of publicly available hospitality data and ensures appropriate interpretation of the reported performance gains.

In summary, this work demonstrates that explainable multimodal artificial intelligence can provide a powerful, transparent, and economically meaningful basis for strategic customer value management in hospitality. By unifying behavioral, emotional, and financial value intelligence within a single framework, this approach moves CLV estimation from descriptive reporting to actionable analytics. It lays a solid foundation for future research at the intersection of deep learning, explainable AI, and financial intelligence for customer value management across service industries.

Author Contributions: Conceptualization, M.N.; methodology, M.N. and M.M.; machine learning model design and implementation, M.N.; multimodal fusion and interpretability framework design, M.N.; data preprocessing and feature engineering, M.N.; validation and statistical evaluation, M.N., M.M. and Ž.R.; formal analysis, M.N.; economic and financial interpretation of CLV results, Ž.R.; business modeling and market relevance assessment, Ž.R.; resources, M.N.; data curation, M.N.; writing (original draft preparation), M.N.; writing (review and editing), M.M. and Ž.R.; visualization, M.N.; scientific supervision, methodological mentoring, and critical manuscript revision, M.M.;

project administration, M.N. All authors have read and agreed to the published version of the manuscript and accept responsibility for its content.

Funding: This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502. The article processing charge (APC) was covered by the authors.

Data Availability Statement: Both datasets analyzed in this study are openly available on the Kaggle platform and include the *Hotel Booking Demand* dataset and the *515K Hotel Reviews Data in Europe* dataset [19,20]. These datasets provide the structured transactional records and large-scale unstructured guest review texts used for multimodal Customer Lifetime Value modeling. Derived feature matrices, intermediate artifacts, and trained deep learning model weights are not deposited in a public repository due to their size and dependence on a project-specific Google Colab environment, but are available from the corresponding author upon reasonable request for research verification and reproducibility.

Acknowledgments: This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502, *Intelligent Multi-Agent Control and Optimization Applied to Green Buildings and Environmental Monitoring Drone Swarms (ECOSwarm)*.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CLV	Customer Lifetime Value
DL	Deep Learning
EDA	Exploratory Data Analysis
KDE	Kernel Density Estimation
ML	Machine Learning
MLP	Multilayer Perceptron
RF	Random Forest
GBM	Gradient Boosting Machine
RFM	Recency–Frequency–Monetary
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model–Agnostic Explanations
NLP	Natural Language Processing
RoBERTa	Robustly Optimized BERT Pretraining Approach
BERT	Bidirectional Encoder Representations from Transformers
CLS	Classification Token
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error

References

1. Firmansyah, E.B.; Machado, M.R.; Moreira, J.L.R. How Can Artificial Intelligence (AI) Be Used to Manage Customer Lifetime Value (CLV)—A Systematic Literature Review. *Int. J. Inf. Manag. Data Insights* **2024**, *4*, 100279. [[CrossRef](#)]
2. Ali, N.; Shabn, O.S. Customer Lifetime Value (CLV) Insights for Strategic Marketing Success and Its Impact on Organizational Financial Performance. *Cogent Bus. Manag.* **2024**, *11*, 2361321. [[CrossRef](#)]
3. Ma, H. Optimization of Hotel Financial Management Information System Based on Computational Intelligence. *Wirel. Commun. Mob. Comput.* **2021**, 8680306. [[CrossRef](#)]
4. Kanchanapoom, K.; Chongwatpol, J. Integrated Customer Lifetime Value (CLV) and Customer Migration Model to Improve Customer Segmentation. *J. Mark. Anal.* **2023**, *11*, 172–185. [[CrossRef](#)]

5. Guillet, B.D.; Zhang, Y.R.; Madanoglu, M.; Gao, L. Attribute-Based Pricing in Hotels: Analyzing Industry, Customer, and Behavioral Insights. *Int. J. Hosp. Manag.* **2026**, *133*, 104445. [CrossRef]
6. Rita, P.; Ramos, R.; Borges-Tiago, M.T.; Rodrigues, D. Impact of the Rating System on Sentiment and Tone of Voice: A Booking.com and TripAdvisor Comparison Study. *Int. J. Hosp. Manag.* **2022**, *104*, 103245. [CrossRef]
7. Xu, W.; Yao, Z.; Ma, Y.; Li, Z. Understanding Customer Complaints from Negative Online Hotel Reviews: A BERT-Based Deep Learning Approach. *Int. J. Hosp. Manag.* **2025**, *126*, 104057. [CrossRef]
8. Botunac, I.; Brkić Bakarić, M.; Matetić, M. Comparing Fine-Tuning and Prompt Engineering for Multi-Class Classification in Hospitality Review Analysis. *Appl. Sci.* **2024**, *14*, 6254. [CrossRef]
9. Manosuthi, N.; Lee, J.S.; Han, H. Causal-Predictive Model of Customer Lifetime/Influence Value: Mediating Roles of Memorable Experiences and Customer Engagement in Hotels and Airlines. *J. Travel Tour. Mark.* **2021**, *38*, 461–477. [CrossRef]
10. Li, Y.; Zeng, F.; Zhang, N.; Chen, Z.; Zhou, L.; Huang, M.; Zhu, T.; Wang, J. Multitask Learning Using Feature Extraction Network for Smart Tourism Applications. *IEEE Internet Things J.* **2023**, *10*, 18790–18798. [CrossRef]
11. Pei, Y.; Wang, Y.; Wang, W.; Qi, J. Aspect-Based Sentiment Analysis with Multi-Task Learning. In Proceedings of the 2022 5th International Conference on Computing and Big Data (ICCBD), Shanghai, China, 16–18 December 2022. [CrossRef]
12. Shah, R.; Pawar, A.; Kumar, M. Enhancing Machine Learning Model Using Explainable AI. In *Advances in Data and Information Sciences*; Tiwari, S., Trivedi, M.C., Kolhe, M.L., Singh, B.K., Eds.; ICDIS 2023; Lecture Notes in Networks and Systems; Springer: Singapore, 2024; Volume 796. [CrossRef]
13. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. [CrossRef]
14. Nikolić, M.; Stojanović, M.; Marjanović, M. Anomaly Detection in Hotel Reviews: Applying Data Science for Enhanced Review Integrity. In *Proceedings of the 2024 32nd Telecommunications Forum (TELFOR), Belgrade, Serbia, 26–27 November 2024*; IEEE: Belgrade, Serbia, 2024; pp. 1–4. [CrossRef]
15. Nikolić, M.; Stojanović, M.; Marjanović, M. Integrating Data Science and Predictive Modeling for Detecting Inconsistent Hotel Reviews. In *UNITECH 2024—Selected Papers*; Technical University of Gabrovo: Gabrovo, Bulgaria, 2024; pp. 104–110. [CrossRef]
16. Nikolić, M.; Stojanović, M.; Marjanović, M. Integrating Deep Learning for Automated Detection of Negative Hotel Reviews. *Facta Univ. Ser. Autom. Control Robot.* **2025**, *24*, 1–16. [CrossRef]
17. Nikolić, M.; Rađenović, Ž.; Marjanović, M. Enhancing Hotel Management through Predictive AI Models for Customer Lifetime Value (CLV). In Proceedings of the International Multidisciplinary Scientific Conference “AI for a Smarter Tomorrow: AI-SMART 2025”, Belgrade, Serbia, 25–26 September 2025; to be published in proceedings.
18. Nikolić, M.; Rađenović, Ž.; Marjanović, M. From Integrity to Income: Quantifying the Economic Impact of Reliable Online Reviews in the Hospitality Industry. In Proceedings of the International Scientific Conference ITEM 2025, Milan, Italy, 6 November 2025; to be published in conference proceedings.
19. Mostipak, J. Hotel Booking Demand. Kaggle Dataset. Available online: <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand> (accessed on 9 December 2025).
20. Liu, J. 515K Hotel Reviews Data in Europe [Dataset]. Kaggle. Available online: <https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe> (accessed on 9 December 2025).
21. Antonio, N.; de Almeida, A.; Nunes, L. Hotel booking demand datasets. *Data Brief* **2019**, *22*, 41–49. [CrossRef]
22. Asenova, M.; Llopis, J.; Gasco, J.; Gonzalez, R. Electronic Word-of-Mouth in the Hospitality Industry: A Comparative Study in Top Hotels. *Int. J. Bus. Excell.* **2024**, *34*, 201–225. [CrossRef]
23. Abdirazakov, N.M.; Kushebayev, Z.T.; Numanova, F.A.; Orynbet, P.Z.; Myltykbayeva, G.E. Understanding the Relationships between Hotel Variables in Almaty, Kazakhstan: An Investigation Using Booking.com Data. *Buketov Bus. Rev.* **2023**, *112*, 7–18. [CrossRef]
24. Souza, F.D.; Filho, J.B.D.O.E.S. Embedding Generation for Text Classification of Brazilian Portuguese User Reviews: From Bag-of-Words to Transformers. *Neural Comput. Appl.* **2023**, *35*, 9393–9406. [CrossRef]
25. Pramudya, Y.G.; Alamsyah, A. Hotel Reviews Classification and Review-Based Recommendation Model Construction Using BERT and RoBERTa. In *Proceedings of the 2023 6th International Conference on Information and Communications Technology (ICOIACT)*; IEEE: Piscataway, NJ, USA, 2023; pp. 437–442. [CrossRef]
26. Moradi, R.; Berangi, R.; Minaei, B. A Survey of Regularization Strategies for Deep Models. *Artif. Intell. Rev.* **2020**, *53*, 3947–3986. [CrossRef]
27. Singgalen, Y. Optimizing Machine Learning in Hospitality Industry: Implementation of Random Forest Model in Forecasting Hotel Guest Length of Stay. *J. Kepariwisata Destin. Hosp. Dan Perjalanan* **2025**, *9*, 63–77. [CrossRef]
28. Messias, A.; Rosa, G.; Couto, P.; Rodrigues, V.; Silva, L.M.; Marques, J. A New Hotel Classification Model Combining Guest Reviews with Official Hotel Classification Systems: Bridging Expert and Consumer Ratings. *Tour. Hosp. Res.* **2025**, 1–19. [CrossRef]

29. Yadav, A.R.; Shekhar, S.; Vidyarthi, A.; Prakash, R.; Gowri, R. Hyper-Parameter Tuning with Grid and Randomized Search Techniques for Predictive Models of Hotel Booking. In Proceedings of the 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM), Roorkee, India, 26–27 August 2023. [[CrossRef](#)]
30. Gregoriades, A.; Pampaka, M.; Herodotou, H.; Christodoulou, E. Explaining Tourist Revisit Intention Using Natural Language Processing and Classification Techniques. *J. Big Data* **2023**, *10*, 60. [[CrossRef](#)]
31. Harris, C.G. Using Explainable AI (XAI) to Understand Sentiment Analysis of Hotel Reviews. In *International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*; Springer Nature: Singapore, 2024; pp. 215–226. [[CrossRef](#)]
32. Czerwinska, U. Interpretability of Machine Learning Models: How Can One Explain Machine Learning Models? In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer International Publishing: Cham, Switzerland, 2022; pp. 275–303. [[CrossRef](#)]
33. Phumchusri, N.; Suwatanapongched, P. Forecasting Hotel Daily Room Demand with Transformed Data Using Time Series Methods. *J. Revenue Pricing Manag.* **2023**, *22*, 44–56. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.