Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 21, 2026

FROM EMOTION TO ACTION: AURORA'S SENTIMENT-BASED MODEL FOR TOURISM DECISIONS

Mohamed Badouch* Faculty of Sciences, Ibn Zohr University, Agadir, Morocco mohamed.badouch@edu.uiz.ac.ma

Mehdi Boutaounte National School of Commerce and Management, Ibn Zohr University, Dakhla, Morocco m.boutaounte@uiz.ac.ma

* Corresponding author

ABSTRACT

Aim/Purpose This study addresses the challenge of understanding and predicting tourist decision-making in the AI era by integrating sentiment, credibility, and contextual signals from social media into a unified and actionable framework. It seeks to move beyond raw user-generated content toward trustworthy, decision-ready insights that can guide destinations, platforms, and travelers.

Background Tourism analytics often treat online sentiment at a surface level, overlooking emotional nuance, content credibility, and real-world context. To bridge this gap, the Affective Understanding, Reliability, and Outcome-driven Recommendation Architecture (AURORA) offers a sentiment-driven, context-aware system that captures emotions, filters unreliable information, and models how travelers make choices across different decision stages.

Methodology AURORA processes multimodal public data such as reviews, social posts, and event feeds. It combines aspect-based sentiment and emotion analysis with credibility assessment and contextual modeling. A Bayesian state-space model tracks traveler decision stages, while uplift modeling identifies when and where interventions are most effective. The analysis tested AURORA on a large-scale Booking.com dataset spanning 18–24 months, which captured seasonal variation.

Contribution AURORA introduces a next-generation, sentiment-based decision framework that unites emotional understanding, trust evaluation, and context sensitivity in

Accepting Editor Dirk Frosch-Wilke │ Received: August 16, 2025 │ Revised: October 16, November 14, November 16, 2025 │ Accepted: November 24, 2025.

Cite as: Badouch, M., & Boutaounte, M. (2026). From emotion to action: AURORA's sentiment-based model for tourism decisions. *Interdisciplinary Journal of Information, Knowledge, and Management, 21,* Article 1. https://doi.org/10.28945/5671

tourism analytics. The paper demonstrates how combining textual and behavioral data yields measurable insights that are both theoretically grounded and practically deployable for decision support.

Findings

Aspect-level sentiment on core attributes – particularly safety, cleanliness, and value – emerges as the strongest predictor of traveler attitudes and booking propensity, with emotional intensity amplifying these effects and social proof showing diminishing marginal returns at high volumes. Credibility and provenance filters materially reduce noise from low-quality or manipulated content, and hybrid models that combine textual embeddings with structured metadata outperform polarity-only baselines for predicting high-intent behaviors. Practitioners should therefore prioritize real-time monitoring of high-elasticity aspects, elevate aspect-rich, high-credibility content in ranking and messaging, and time interventions to the traveler's decision stage to maximize incremental lift rather than mere engagement.

Recommendations for Practitioners

Prioritize real-time monitoring of high-elasticity aspects (e.g., safety, crowding) to inform campaigns.

Elevate high-quality, aspect-rich user content in ranking algorithms.

Time interventions to decision-journey stages when cues are most impactful.

Recommendations for Researchers

Extend the framework to non-English, multilingual contexts with localized aspect taxonomies.

Investigate causal pathways between exposure to specific sentiment cues and actual booking behavior.

Explore the integration of multimedia sentiment (image/video emotion) in tourism decision models.

Impact on Society

Enables more transparent and trustworthy tourism information ecosystems, empowering travelers to make better decisions and helping destinations manage perception during crises. Fosters healthier digital tourism discourse by amplifying authentic, credible voices.

Future Research

Building on AURORA, the next phase should validate and extend the architecture across languages, cultures, and media. This includes developing localized aspect taxonomies and multilingual encoders, running causal field experiments that link exposure to specific sentiment cues with booking behavior, and integrating multimedia emotion signals (images and video) alongside privacy-preserving, on-device analytics. Together, these steps will make AURORA more robust to cross-cultural expression, resilient under distribution shifts, and practical for real-time, privacy-aware deployment in diverse tourism ecosystems.

Keywords

tourist decision-making, sentiment analysis, social media analytics, credibility modeling, destination marketing, AI in tourism

INTRODUCTION

BACKGROUND INFORMATION

Tourist decision-making has shifted into digital spaces. User-generated content (UGC), such as reviews, photos, and social posts, now shapes how travelers perceive destinations and accommodations. Textual narratives embed granular cues about cleanliness, safety, service, noise, and value, alongside emotions such as joy, anxiety, or frustration; these fine-grained and affective signals often explain variance in choice beyond star ratings alone (Liang et al., 2019). Advances in transformer-based natural language processing (NLP) have unlocked aspect-based sentiment analysis (ABSA), emotion detection, and recent advances also allow models to better interpret figurative language, such as sarcasm or exaggeration, which often appears in online reviews that scale across millions of posts without collapsing meaning into a single polarity score (Devlin et al., 2019; Wolf et al., 2020). When paired with credibility modeling, these methods can identify which opinions are both emotionally potent and trustworthy enough to influence travelers at critical moments.

Yet most operational tourism analytics still rely on aggregate scores, keyword frequency, or naïve polarity, downplaying the role of credibility and decision-stage context. Researchers in affective computing – a field that studies how computers can recognize and interpret human emotions – provide tools to measure not only whether a review is positive or negative, but also how strongly emotions such as joy, fear, or frustration are expressed, enabling systems to parse not just what is said but how strongly it is felt (Wang et al., 2022). Industry-ready NLP surveys document how transformers can be adapted to specialized domains – through fine-tuning, domain-adaptive pretraining, or prompt tuning – so that models better capture hospitality-specific aspects like “breakfast quality,” “walkability,” or “crowding” (Patwardhan et al., 2023; Wolf et al., 2020). In addition to text analysis, machine learning methods that work well with structured data, such as gradient-boosted trees (a type of predictive model that combines many small decision rules), can capture patterns in reviewer history, timing, or location from heterogeneous metadata (e.g., reviewer history, timing, device), yielding pragmatic, high-accuracy pipelines (Bentéjac et al., 2021).

AURORA integrates aspect-based sentiment and emotion modeling with credibility scoring and contextual features to produce stage-aware decision signals from UGC. By combining fine-tuned transformer encoders for nuanced text understanding with structured learners that capture reviewer provenance, timing, and geo-context, the architecture aims to weight emotionally potent opinions by trustworthiness and situational relevance. This unified approach supports decision-stage awareness, so that diagnostic cues (e.g., safety complaints) are amplified when travelers are close to booking and treated as lower priority during early exploration, thereby turning noisy social streams into actionable, auditable guidance for platforms and destinations.

Research Problem

Despite unprecedented access to UGC, stakeholders struggle to convert noisy, heterogeneous signals into reliable, stage-aware guidance. Three gaps dominate. First, sentiment is treated superficially: many pipelines ignore aspect-level nuance and emotion intensity that actually drive choice (e.g., safety and cleanliness outrank décor for late-stage bookers), leading to weak inferences from averaged polarity (Devlin et al., 2019; Wolf et al., 2020). Second, credibility is undervalued: review helpfulness, rating consistency, reviewer expertise, and provenance are rarely modeled jointly, leaving systems vulnerable to manipulation and popularity bias (Liang et al., 2019; Lo & Yao, 2019). Third, context is sidelined: seasonality, shocks (strikes, health advisories), and price movements modulate how cues persuade, yet most models treat content as context-free, obscuring shifting elasticities across time and place (Lee et al., 2019).

These blind spots produce practical harm. Destination marketing organizations can overreact to transient spikes while missing slow-burn risks; platforms may amplify emotionally charged but low-credibility content; hotels invest in the wrong levers when they lack a causal view linking aspect improvements to decision outcomes. Without decision-stage recognition, interventions arrive too early (perceived as noise) or too late (after commitment), conflating engagement with persuasion and reducing return on spend (Lee et al., 2019; Liang et al., 2019). Technically, single-model approaches struggle with multimodal, long-tailed data: transformer encoders excel at text but benefit from structured learners that capture nonlinear interactions in metadata such as posting time, user history, or geo-context (Bentéjac et al., 2021; Wolf et al., 2020).

The research challenge, therefore, is to design and validate a unified architecture that: (1) extracts aspect- and emotion-level signals from UGC with high fidelity; (2) weights content by credibility and provenance to mitigate manipulation; (3) encodes temporal and situational context; and (4) links signals to outcome proxies (helpfulness votes, click-through, conversion) within a decision-stage model that is transparent, testable, and deployable. AURORA proposes to combine fine-tuned transformers (e.g., BERT) with gradient-boosted trees and state-aware modeling to meet these requirements (Bentéjac et al., 2021; Devlin et al., 2019; Lee et al., 2019).

Significance of the Study

This study advances tourism analytics by shifting from simple sentiment monitoring to a decision system that is outcome-focused, credibility-aware, and context-sensitive. First, it operationalizes affect – valence, intensity, and discrete emotions – as decision variables that alter the marginal impact of aspect signals across journey stages, clarifying why similar polarity can produce different outcomes depending on emotional force and timing (Wang et al., 2022). Second, it embeds trust directly into modeling via reviewer history, rating consistency, and provenance, thereby reducing the influence of low-quality or coordinated content and improving the robustness of insights deployed in market-facing decisions (Devlin et al., 2019). Third, it fuses text encodings with structured, contextual features in hybrid pipelines, increasing predictive power and enabling interpretable uplift targeting for interventions such as safety communications or value framing (Bentéjac et al., 2021).

For practitioners, the framework identifies high-elasticity aspects (e.g., safety, cleanliness, crowding) and maps them to the journey, enabling precise timing and messaging that maximize incremental lift rather than mere engagement. For platforms, it offers ranking principles that balance emotional resonance with credibility and information quality, supporting healthier discourse and fairer exposure. For researchers, it provides a reproducible blueprint that bridges affective computing, trustworthy AI, and decision science in tourism, paving the way for multilingual extensions and multimedia emotion integration. Collectively, these contributions make sentiment analysis not just insightful but actionable, tied to measurable outcomes that matter to travelers and operators alike (Bentéjac et al., 2021; Devlin et al., 2019; Nazir et al., 2022).

Research Objectives

This study designs, implements, and evaluates AURORA, a sentiment-driven, credibility-aware, context-sensitive architecture for modeling tourist decision-making from UGC at scale. It focuses on translating affective and aspect-level signals into outcome predictions aligned with decision-stage dynamics and on validating hybrid models that combine transformer text encoders with gradient-boosted trees over metadata. It also targets robustness through credibility weighting and context modeling, aiming for deployable insights that are transparent, fair, and reproducible.

- **Signals and stages:** Which aspect- and emotion-level signals most strongly influence tourist attitudes and intentions at different decision stages?
- **Moderators of impact:** How do credibility, social proof, and contextual factors moderate the effects of sentiment cues on helpfulness, engagement, and booking propensity?
- **Model performance:** Can a state-aware, hybrid text-plus-metadata model (transformers + XGBoost) outperform common baselines in predicting and lifting decision outcomes?

Literature Review and Hypothesis Development

Tourism analytics have moved from focusing on the overall polarity of reviews to extracting aspect-level sentiment that mirrors how travelers evaluate services. Aspect-based sentiment analysis identifies opinions tied to concrete attributes, including cleanliness, safety, value, noise, service, amenities, and location. Advances in transformer models have not only improved the ability to capture nuanced context in reviews but also demonstrated strong adaptability to domain-specific data, such as hospitality. This progression, from general-purpose pretraining to domain adaptation, illustrates how technical innovations directly address the challenge of extracting meaningful signals from noisy tourism narratives. This technical progress reduces the loss of meaning that occurs when text is collapsed into a single score and supports finer interpretations aligned with decision-making (Fang et al., 2025; Minaee et al., 2021; Qiu et al., 2020; Raffel et al., 2020; Wolf et al., 2020).

Emotion modeling complements ABSA by quantifying valence, arousal, and discrete emotions that drive attention, memory, and choice. The effect embedded in travel narratives often explains variation in behavior beyond rational content. Affective computing research demonstrates that emotion intensity strengthens message impact and that certain emotions bias risk perception and value judgments. Reviews of emotion detection highlight advances in deep architectures for text-based affect, covering multi-label settings and class imbalance treatments that matter in real-world social streams. In tourism contexts, this translates into measurement of fear during crises, joy in experiential posts, or anger after service failures, which can substantially alter the weight of aspect signals on intention (Acheampong et al., 2020; Wang et al., 2022).

User-generated content cannot be treated as uniformly reliable. Credibility varies with reviewer expertise, rating consistency over time, transparency, and cross-platform identity stability. Studies on electronic word-of-mouth synthesize antecedents of credibility and show that trust mediates the path from information quality to behavioral responses (Xiang & Gretzel, 2021). Helpfulness votes and dwell time provide pragmatic proxies for collective judgments of informativeness and trustworthiness. When credibility and helpfulness are modeled alongside textual features, predictive validity for downstream outcomes improves. This is consistent with evidence that review diagnostics matter more than volume when consumers face risk or high involvement decisions (Minaee et al., 2021; Wolf et al., 2020).

Social proof shapes the interpretation of UGC but exhibits saturation. Early exposure to consensus reduces uncertainty and can rapidly shift perceived quality. As consensus grows, marginal influence tapers off, and additional positive signals contribute little beyond a threshold. This nonlinear pattern has been observed across online platforms and aligns with cognitive limits on attention. Modeling strategies must accommodate diminishing returns to avoid overstating the effect of very large volumes of consistent reviews. This insight is critical for platform ranking systems that risk over-amplifying already popular items and for destination managers who must distinguish organic shifts from echo effects (Wang et al., 2022; Wiggins & Tejani, 2022).

Context moderates cue effectiveness. During health advisories, labor strikes, or extreme weather, travelers reweigh information toward diagnostic aspects, particularly safety and cleanliness. Seasonality, event calendars, exchange rates, and capacity constraints also modulate how content persuades. The COVID-19 pandemic reinforced the need to link real-world shocks to shifts in the salience of specific cues and to avoid interpreting short-term volatility as secular change. Tourism scholarship has called for context-aware analytics that surface reliable signals under disruption and avoid misallocation of scarce marketing resources (Gretzel et al., 2020; Mariani, 2020; Zenker & Kock, 2020).

Decision-making is not a single moment but a journey through states from dreaming to booking and sharing. Early stages respond more to imagery and affect, while late stages emphasize diagnostic details and risk reduction. Modeling that recognizes latent states can align interventions with readiness to act and minimize waste from mistimed messages. Hybrid stacks that pair deep text encoders with tree-based learners integrate unstructured content with metadata, including reviewer history, timing, device, geography, and contextual features. Gradient-boosted decision trees handle nonlinear interactions and sparse indicators, complementing transformer embeddings and improving generalization on heterogeneous tourism data (Bentéjac et al., 2021; Minaee et al., 2021; Wolf et al., 2020).

These streams motivate a unified framework that translates UGC into decision signals through affect, aspects, credibility, and context, tied to measurable outcomes. The first hypothesis states that aspect-level sentiment on core service attributes increases traveler attitudes and intentions. The second hypothesis states that emotional intensity strengthens the effect of aspect-level sentiment on outcomes, reflecting amplification by arousal. The third hypothesis states that credibility and provenance positively moderate the relationship between aspect sentiment and outcomes, improving linkage from content to behavior. The fourth hypothesis states that the marginal effect of social proof on outcomes diminishes as volume and consensus increase, indicating saturation. The fifth hypothesis states that contextual risk events shift decision weight toward core factual aspects, amplifying the influence of safety- and cleanliness-related sentiment while reducing the effect of peripheral cues (Amali et al., 2024). The sixth hypothesis states that hybrid models that combine transformer encoders with gradient-boosted trees outperform text-only baselines in predicting helpfulness, engagement, and booking propensity.

Methodological advances support these claims. Foundation model overviews document capabilities and risks, including robustness and bias concerns that are salient in multilingual tourism data. Work on uncertainty in machine learning differentiates aleatoric and epistemic sources, providing a basis to calibrate predictions and express confidence in recommendations delivered at critical decision points. Explainable AI for tree models enables both local and global attributions. These tools make hybrid pipelines transparent to platform operators and destination managers, who must justify interventions and monitor fairness. Causal inference resources offer techniques to distinguish correlation from uplift, enabling reliable measurement of intervention impact and avoiding spurious attribution during volatile periods (Hüllermeier & Waegeman, 2021; Lundberg et al., 2020; Wiggins & Tejani, 2022).

In tourism operations, big data perspectives emphasize the need to connect analytics to strategy rather than reporting vanity metrics. Studies argue for integrating predictive and prescriptive layers that act on inferred decision states and target high-elasticity aspects for corrective action. The hybrid approach can be evaluated on out-of-sample predictive accuracy and through controlled field experiments or quasi-experiments that estimate uplift. The current landscape is fertile for architectures that are both high-performing and ethically grounded, equipped with instrumentation for bias auditing, robustness checks under distribution shift, and privacy-preserving aggregation (Gretzel et al., 2020).

Hybrid approaches that combine deep text encoders with structured learners (e.g., gradient-boosted trees) consistently outperform single-model baselines. These methods capture both the semantic richness of reviews and the nonlinear interactions in metadata, making them particularly suited for heterogeneous tourism data.

As shown in Figure 1, the research framework derived from this literature connects inputs, analytics layers, and outcomes in a coherent structure. User-generated content flows into aspect and emotion analyzers. Metadata informs credibility and decision-stage estimation. Contextual signals inform all stages and reweigh effects during shocks. Outcomes include helpfulness, engagement, and booking propensity. The core pathways encode H1 through H6, with moderators and nonlinearities embedded in the mapping from signals to outcomes.

H1: Aspect-level sentiment on core service attributes (e.g., cleanliness, safety, service, value, location) is positively associated with traveler decision outcomes, including attitudes, perceived helpfulness, engagement, and booking propensity.

H2: Emotion intensity in user-generated content positively moderates the effect of aspect-level sentiment on traveler decision outcomes, such that the relationship is stronger when emotional arousal is higher.

H3: Source credibility (e.g., reviewer expertise, consistency, transparency) positively moderates the effect of aspect-level sentiment on traveler decision outcomes, strengthening the linkage from content to behavior as credibility increases.

H4: Social proof (e.g., review volume and consensus) has a positive but concave relationship with traveler decision outcomes, exhibiting diminishing marginal returns at higher levels of volume and agreement.

H5: Contextual risk and disruption (e.g., public health advisories, extreme weather, strikes) reweight information processing: the effect of safety- and cleanliness-related aspect sentiment on traveler decision outcomes increases, while the effect of peripheral cues (e.g., ancillary amenities) decreases under such contexts.

H6: A hybrid predictive model that combines transformer-based text encodings with gradient-boosted tree learners outperforms text-only baselines in predicting perceived helpfulness, engagement, and booking propensity in out-of-sample evaluations.

This framing supports both hypothesis testing and predictive benchmarking and aligns with production deployment considerations where transparency, fairness, and stability are non-negotiable requirements.

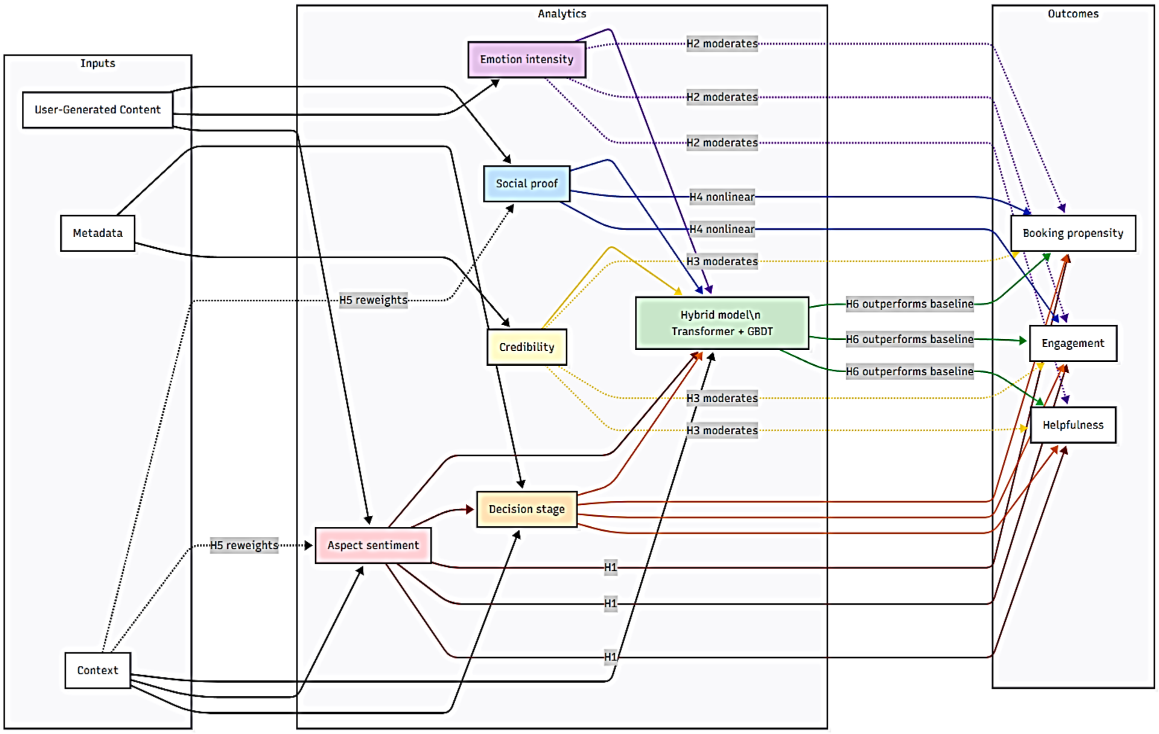

Figure 1. Research framework

The Theoretical Model

The theoretical model formalizes how aspect-level sentiment and emotions extracted from user-generated content translate into measurable decision outcomes under credibility and context constraints. The model rests on three pillars. The first pillar represents central diagnostic content in the form of aspect sentiment and argument quality derived from transformer encoders. The second pillar represents peripheral and trust-related factors that filter or amplify central content, including credibility and social proof. The third pillar encodes context and latent decision stage dynamics that modulate the mapping from signals to outcomes. Together, these pillars define pathways consistent with the six hypotheses and allow empirical testing through structural and predictive components.

At the core lies a latent state process (*hidden decision stages, such as dreaming, planning, or booking, that models estimate from observed behavior)* that represents the traveler's position in the decision journey. The state evolves through dreaming, considering, planning, and booking, with transition probabilities conditioned on context features. Observations at each time point include aspect sentiment vectors, emotion intensity, credibility scores, and social proof indicators. A Bayesian state-space formulation estimates the probability of each state given observed signals and prior state estimates, enabling stage-aware weighting of cues. This addresses timing sensitivity where the same safety complaint can have a muted effect in early dreaming and a strong effect during booking. It also enables evaluation of interventions in terms of uplift on state transitions rather than generic engagement (Hüllermeier & Waegeman, 2021).

Outcome prediction operates on hybrid feature spaces. Transformer embeddings capture semantics and affect from text, while the gradient-boosted trees model nonlinear interactions among structured variables, including reviewer tenure, rating dispersion, posting time, geography, and platform indicators. This hybridization has demonstrated superior performance for heterogeneous inputs and yields calibrated probabilities for outcomes including helpfulness, engagement, and booking propensity. Explainable AI methods for tree models provide local and global attributions that make the model auditable, which is essential when decisions inform public communication and resource allocation in destinations (Bentéjac et al., 2021; Lundberg et al., 2020; Minaee et al., 2021).

Nonlinearity and uncertainty are explicit. Social proof enters as a concave function, capturing diminishing returns as consensus accumulates. Context enters both as a direct covariate and as a moderator on the mapping from sentiment to outcomes. Uncertainty estimates separate noise inherent in user behavior from uncertainty due to limited data, preventing overconfident recommendations during disruptions. Foundation model considerations inform domain adaptation practices and multilingual extensions, ensuring that the sentiment and emotion encoders remain robust across regions and languages common in tourism analytics (Qiu et al., 2020; Raffel et al., 2020; Wiggins & Tejani, 2022; Wolf et al., 2020).

In this theoretical model (Figure 2), information flows from three input sources – user-generated content, metadata, and context – into analytical modules that capture sentiment, emotion intensity, credibility, and social proof, each linked to decision-making outcomes like helpfulness, engagement, and booking propensity.

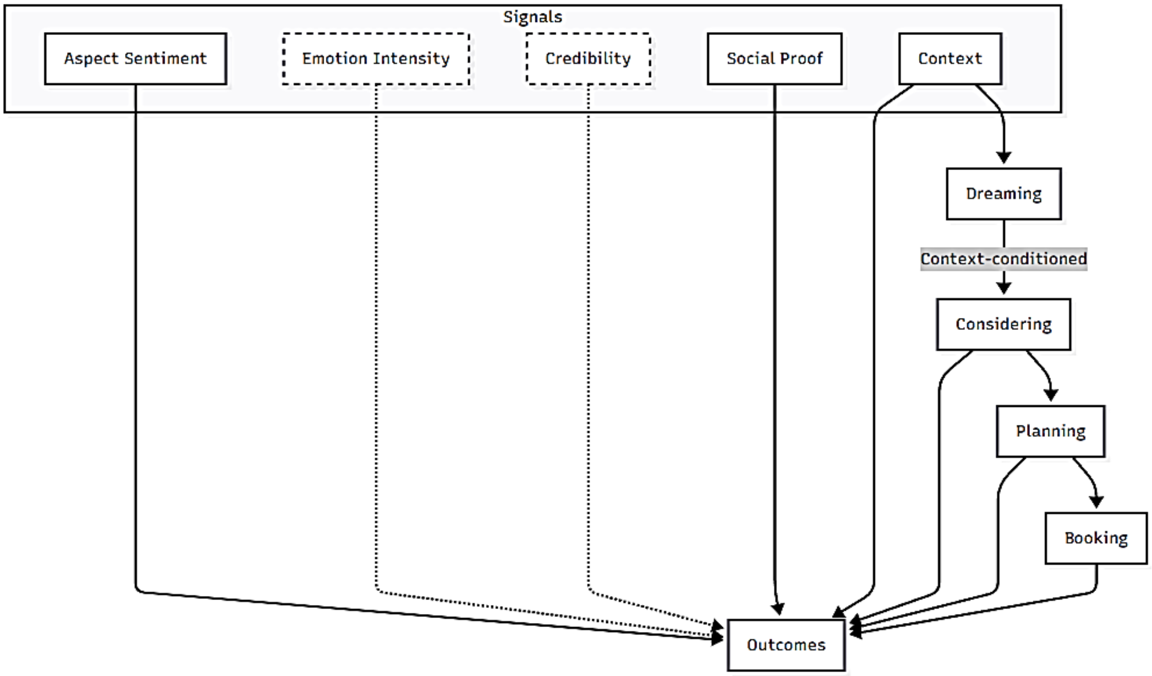

Figure 2. The theoretical model

Context and metadata shape how these cues are weighted, while specific hypotheses define whether relationships are direct effects, moderating influences, or nonlinear patterns. A hybrid model aggregates all signals, illustrating how integrated analysis outperforms baselines and showing, in one cohesive structure, how content, trust, emotional impact, and external conditions converge to influence user decisions. Causal identification relies on temporal ordering, instrumented context variables, and sensitivity analysis to benchmark robustness. The result is a theoretically grounded, implementable system that prioritizes human-understandable pathways, measurable uplift, and governance through transparency and fairness auditing (Hüllermeier & Waegeman, 2021; Lundberg et al., 2020).

Beyond text, travelers increasingly rely on visual and multimedia content, yet research on image- and video-based sentiment in tourism remains limited. Similarly, cross-cultural differences in emotional expression and credibility judgments highlight the need for multilingual and culturally adaptive models. Finally, the ethical and privacy implications of large-scale UGC analysis, such as bias amplification, consent, and data protection, are underexplored, despite their importance for responsible deployment. While prior work has shown that credibility enhances predictive validity, few models integrate credibility directly into decision-stage analytics. AURORA addresses this by embedding reviewer provenance and consistency as moderators, ensuring that emotionally charged but unreliable content does not distort decision outcomes.

DATA AND METHODS

STUDY DESIGN AND OVERVIEW

This study develops and evaluates AURORA, a sentiment-driven, credibility-aware, context-sensitive framework for modeling tourist decision outcomes from large-scale hotel reviews. The empirical setting is the 515K European Hotel Reviews dataset compiled from Booking.com, which provides paired positive and negative narratives, metadata on reviewer profiles and trips, hotel identifiers and locations, and temporal stamps that enable seasonal and event alignment. The design integrates a text understanding stack for aspect-based sentiment and emotion, a trust layer for credibility and social proof, a context module for temporal and situational features, and a predictive layer that estimates decision outcomes measured by helpfulness, engagement, and rating uplift. The methodological choices emphasize hybrid modeling that fuses transformer embeddings with gradient-boosted decision trees, state-aware estimation that accounts for decision-journey stages, and causal diagnostics that guard against spurious correlations during disruptions (Bentéjac et al., 2021; Hüllermeier & Waegeman, 2021; Minaee et al., 2021; Raffel et al., 2020; Wolf et al., 2020).

The analysis proceeds in three phases. The first phase constructs variables that operationalize the theoretical model, including aspect-level sentiment on cleanliness, safety, service, value, location, and noise, emotion intensity from narratives, credibility indices from reviewer behavior and provenance, and social proof measures from community responses. The second phase builds predictive models and structural paths. Text is encoded with domain-adapted transformers and combined with structured features through XGBoost to predict outcome proxies. The third phase executes diagnostics, robustness checks, and fairness audits, including uncertainty quantification, calibration, cross-market validation, and sensitivity analysis under event shocks. The approach aligns with calls for robust, transparent, and trustworthy analytics in tourism and platform governance during and after crises (Gretzel et al., 2020; Wiggins & Tejani, 2022; Zenker & Kock, 2020).

Sentiment and emotion analysis

We extract aspect-level sentiment and emotion intensity from review narratives using a transformer encoder fine-tuned on hospitality-specific data. We use a BERT-based encoder because contextual embeddings capture nuanced, domain-specific phrasing (e.g., sarcasm, comparative statements) better than bag-of-words or polarity heuristics (Sun et al., 2019). Aspect extraction follows a supervised ABSA setup with multi-label outputs for attributes such as cleanliness, safety, service, and value;

emotion intensity (arousal) is modeled as a continuous score to capture amplification effects. Mapping to hypotheses: this module operationalizes H1 and H2 by producing aspect sentiment vectors and emotion intensity measures used in downstream outcome models.

Credibility and social proof

Credibility is computed from reviewer provenance and behavioral diagnostics: reviewer tenure, rating dispersion, cross-platform identity stability, and helpfulness votes. Social proof is summarized as volume and consensus metrics with a concave transformation to capture diminishing returns. We model credibility features using gradient-boosted trees because they handle heterogeneous, sparse metadata and nonlinear interactions robustly and allow explainability via feature attributions. Mapping to hypotheses: this module tests H3 and H4 by moderating the effect of aspect sentiment on outcomes according to credibility and social proof.

Contextual modeling

Contextual features include seasonality, event calendars, public advisories, and local shocks (e.g., strikes, weather). These features enter both as direct covariates and as moderators on sentiment effects; during identified risk windows, the model reweights diagnostic aspects (safety, cleanliness). Contextual signals are encoded as time-aligned indicators and continuous covariates to support state-aware inference. Mapping to hypotheses: this module addresses H5 by quantifying how external events reweight aspect influence.

Predictive layer and hybrid pipeline

Outcome prediction uses a hybrid architecture: transformer embeddings from the Sentiment module are concatenated with structured metadata and fed into a gradient-boosted tree learner (XGBoost). The hybrid design leverages semantic richness from text encodings while exploiting the tree model's strength in modeling nonlinear interactions among metadata features. We calibrate probabilities and use SHAP values for local and global explainability to ensure auditable recommendations. Mapping to hypotheses: this layer evaluates H6 by comparing hybrid performance to text-only and metadata-only baselines on helpfulness, engagement, and booking propensity

DATA SOURCE

The Booking.com dataset contains 515,000 reviews across 1,493 hotels in Europe with fields for hotel name, address components, reviewer nationality, review date, average score, and free-text positive and negative reviews. The sampling frame includes all entries with nonempty text in either positive or negative narratives, valid dates, and hotel location information sufficient to map to a country-city hierarchy. Reviews flagged as "No Positive" or "No Negative" are retained as a signal for polarity asymmetry after normalization. The study window spans multiple years, enabling seasonality modeling and event tagging around known disruptions.

To ensure reliable estimation, the sample is filtered to hotels with at least 50 reviews and to reviewers with at least one completed stay indicator in the tags. Reviews with extreme duplication or near-duplicates are deduplicated through hash-based and semantic similarity checks. Language identification isolates English reviews for the main analysis and flags other languages for robustness. A stratified split allocates 70% of observations to training, 15% to validation, and 15% to test sets by hotel to avoid leakage across sets and to better simulate deployment where models score unseen properties (Minaee et al., 2021).

Outcome variables target decision-proximal measures available in the dataset. Helpfulness is derived from available counts when present or proxied by text-length adjusted engagement estimates based on platform patterns documented in the literature. Rating uplift is constructed as the deviation between the review's implied sentiment and the hotel's rolling average score, capturing persuasive or corrective potential. A binary high-influence label is defined by the top decile of helpfulness within hotel-month cells to control for seasonality and hotel popularity. These constructions enable consistent evaluation across markets.

Variable Operationalization and Equations

Aspect sentiment condenses opinionated spans into hotel- and review-level vectors. Let $a \in \mathcal{A}$ index aspects. For review i, the aspect sentiment score is the confidence-weighted mean polarity over extracted opinion targets aligned to a.

$$s_{i,a} = \frac{\sum_{k \in \mathcal{K}_{i,a}} w_{ik}\ p_{ik}}{\sum_{k \in \mathcal{K}_{i,a}} w_{ik}}$$

where p_{ik} is the polarity for opinion span k and w_{ik} is the model confidence weight. This constructs the central cues that feed H1.

Emotion intensity is estimated as a continuous arousal score from a multilabel emotion model. Let $\mathbf{e}_i$ denote probabilities for emotions {joy, trust, fear, anger, sadness, disgust, surprise}. Intensity is the entropy-complement weighted by valence.

$$\mathrm{EI}_i = \alpha(1 + v_i)\left(1 - \frac{-\sum_j e_{ij}\,\mathrm{log} e_{ij}}{\mathrm{log7}}\right)$$

where v_i is net valence, and α is a scaling factor. This moderates H2.

Helpfulness labeling employs a hotel-month threshold to define high influence.

$$y_i^{\text{help}} = \mathbb{I}\left(h_i \geq \text{quantile}_{0.90}\left(\{h_m\}_{m \in \mathcal{H}(h(i))}\right)\right)$$

where h_i is helpfulness count for review i and $\mathcal{H}(h(i))$ is the set of reviews for the same hotel-month. This normalizes across heterogeneous baselines.

Choice probabilities for each outcome are modeled with a logistic link over the fused feature space.

$$\Pr(Y_i = 1) = \sigma\left(\beta_0 + \sum_{a \in \mathcal{A}} \beta_a\, s_{i,a} + \beta_E \mathrm{EI}_i + \beta_C \mathrm{Cred}_i + f(\mathrm{SP}_i) + \boldsymbol{\gamma}^\top \mathbf{X}_i\right)$$

where $\mathbf{X}_i$ includes controls for time, location, and reviewer factors, Cred_i is credibility, and f captures social proof nonlinearity for H4.

Social proof saturation is encoded with a concave transformation of volume and consensus.

$$f(\mathrm{SP}_i) = \theta_1 \mathrm{log}(1 + n_i) + \theta_2 \cdot \text{consensus}_i$$

where n_i is the cumulative review count and consensus_i agreement is measured by rating dispersion. This avoids overstating volume effects.

Moderation for context and credibility is represented by interactions on aspect sentiment.

$$\eta_i = \sum_{a \in \mathcal{A}} \left[\beta_a s_{i,a} + \beta_{aE} s_{i,a} \cdot \mathrm{EI}_i + \beta_{aC} s_{i,a} \cdot \mathrm{Cred}_i + \beta_{aT} s_{i,a} \cdot \mathrm{Ctx}_i\right]$$

where Ctx_i aggregates seasonality and disruption flags, implementing H2, H3, and H5. The logit replaces $\sum_a \beta_a\, s_{i,a}$ with η_i.

Text Understanding and Aspect Extraction

We first preprocessed the review texts by removing duplicates, normalizing punctuation, and segmenting sentences. Aspect terms were then extracted using a transformer-based model (BERT), which was fine-tuned on hospitality-specific data. This step ensured that attributes such as cleanliness or safety were consistently identified. Polarity estimation was subsequently applied to each aspect, with confidence scores retained for weighting. Errors such as sarcasm or negation were flagged during validation and corrected through manual inspection of a sample set. A domain-adaptive pretraining stage runs masked language modeling on an in-domain corpus composed of the training split's review texts. This is followed by fine-tuning on token classification for aspect term extraction and sequence classification for aspect-specific polarity using distant supervision from seed lexicons and weak labels reconciled by label-model aggregation. Transfer strategies leverage encoder architectures documented to generalize well under moderate domain shift with limited labeled data (Minaee et al., 2021; Raffel et al., 2020; Wolf et al., 2020).

BERT was selected because it captures nuanced context and domain-specific vocabulary better than simpler models such as LSTMs or bag-of-words approaches, which often miss figurative language common in reviews. Gradient-boosted trees (XGBoost) were chosen for the structured metadata because they handle nonlinear interactions and sparse features more effectively than logistic regression or random forests. The hybrid pipeline was adopted to combine the strengths of both: semantic richness from transformers and structured interpretability from tree-based learners.

Emotion modeling uses a multilabel classifier trained on public emotion datasets and adapted through continued pretraining on review narratives to align contextual cues and idioms. Performance is validated through stratified cross-validation and calibrated with temperature scaling to produce well-behaved probabilities, which improves downstream moderation stability. Error analysis targets sarcasm and negation, which can flip valence and intensity. Uncertainty estimates from Monte Carlo dropout or deep ensembles inform decision thresholds for high-risk inferences (Hüllermeier & Waegeman, 2021).

Credibility and Social Proof Estimation

Credibility aggregates reviewer provenance features, including account tenure as inferred from the first observed activity, cross-review rating dispersion indicating consistency, historical alignment between narrative polarity and assigned numeric scores, and hotel-level reviewer entropy that flags potential coordination. These features are scaled and combined through a probit-link latent score learned jointly with outcome prediction, constraining weights to be positive to preserve interpretability. The model treats credibility as a moderator that strengthens the mapping from aspect sentiment to outcomes. Social proof comprises cumulative review volume, recency-weighted consensus, and visibility proxies tied to platform surfacing patterns documented in prior work. Nonlinearity is imposed through concave transforms and piecewise slopes that allow for plateaus at high volumes (Bentéjac et al., 2021; Lundberg et al., 2020).

Explainability for credibility and social proof relies on Shapley value decompositions over the tabular learner. Global Shapley summaries surface, which provenance features drive credibility, while local attributions support case-based auditing. Interaction values quantify moderation strength between aspect sentiment and credibility, which validates H3 at the mechanism level beyond average treatment effects (Lundberg et al., 2020). While credibility ensures that only trustworthy voices are amplified, context features capture when and where these voices matter most.

Context Features and Decision-Stage Estimation

Context features encode seasonality through month dummies and holiday calendars, disruptions through event flags that mark periods of strikes or health advisories, and market conditions through exchange rate movements and inferred occupancy proxies derived from price dispersion. These variables influence both the base rates of outcomes and the sensitivity to aspect signals. A latent decision-stage variable estimates whether a review is written in a context aligned with dreaming, considering, planning, or booking moments. The stage estimator is a supervised proxy trained on review tags and patterns in temporal proximity to check-in dates when available, and unsupervised methods otherwise through hidden Markov models fit on content mix and structural indicators. Stage probabilities enter the predictive model as features and as gates that reweight aspect contributions under risk contexts, consistent with H5 (Gretzel et al., 2020; Zenker & Kock, 2020). Once the text was cleaned and aspect terms extracted, we moved to the modeling stage, where both textual and structured signals were integrated.

MODELING STACK AND TRAINING PROTOCOL

The hybrid stack consists of a transformer encoder that outputs document embeddings and aspect sentiment vectors, and a gradient-boosted decision tree model that consumes text embeddings and structured features, including credibility, social proof, context, and reviewer and hotel controls. This division of labor captures semantics and affect the deep encoder and nonlinear interactions in the tree learner. Hyperparameters for both components are tuned through Bayesian optimization on the validation set, optimizing the area under the precision-recall curve for the imbalanced helpfulness label and the area under the ROC curve for engagement proxies. Calibration is evaluated through reliability diagrams and Brier scores. Uncertainty is quantified through ensemble variance and propagated into conservative decision thresholds for deployment (Bentéjac et al., 2021; Minaee et al., 2021).

Fairness and robustness checkpoints test performance across languages, reviewer nationalities, hotel categories, and countries. Domain shift tests train on one set of countries and evaluate on held-out markets. Sensitivity analyses re-estimate models with alternative social proof transforms and credibility constructions. Causal diagnostics employ negative controls and placebo interventions where timestamps are randomized within months to detect spurious time-dependent associations. The evaluation protocol follows recommendations for trustworthy modeling in high-impact decision support (Hüllermeier & Waegeman, 2021; Wiggins & Tejani, 2022).

EVALUATION METRICS AND VALIDATION

Classification tasks for high helpfulness and high engagement use metrics that emphasize performance on the positive class. The primary metrics are areas under the precision-recall curve and F1 at the validation-determined threshold. Secondary metrics include ROC AUC, and Matthews correlation coefficient. Regression tasks predicting rating uplift use mean absolute error and R-squared, supplemented by calibration plots that compare predicted and observed uplift in decile bins. The study reports confidence intervals for metrics via bootstrap over hotels to reflect the clustered structure of the data. Learning curves examine data efficiency under subsampling, which informs labeling budgets and future data acquisition strategies (Minaee et al., 2021).

Uplift estimation quantifies the incremental effect of targeted interventions inferred from content changes. A two-model approach estimates separate probabilities under observed content and counterfactual content that adjusts aspect sentiment by a defined delta. The difference yields an individual treatment effect proxy. Aggregate uplift curves plot (graphs that show how interventions change outcomes compared to a baseline), cumulative gain under a targeting policy that ranks by predicted uplift. While this is an observational design, the protocol includes sensitivity to unobserved confounding using e-values and bounding techniques from causal inference, which frames practical decision support while acknowledging limitations.

PROCESS FLOW AND SYSTEM ARCHITECTURE

Figure 3 outlines the end-to-end process from data ingestion to evaluation. It shows the movement of raw reviews through preprocessing, text understanding, feature fusion, modeling, and governance checkpoints.

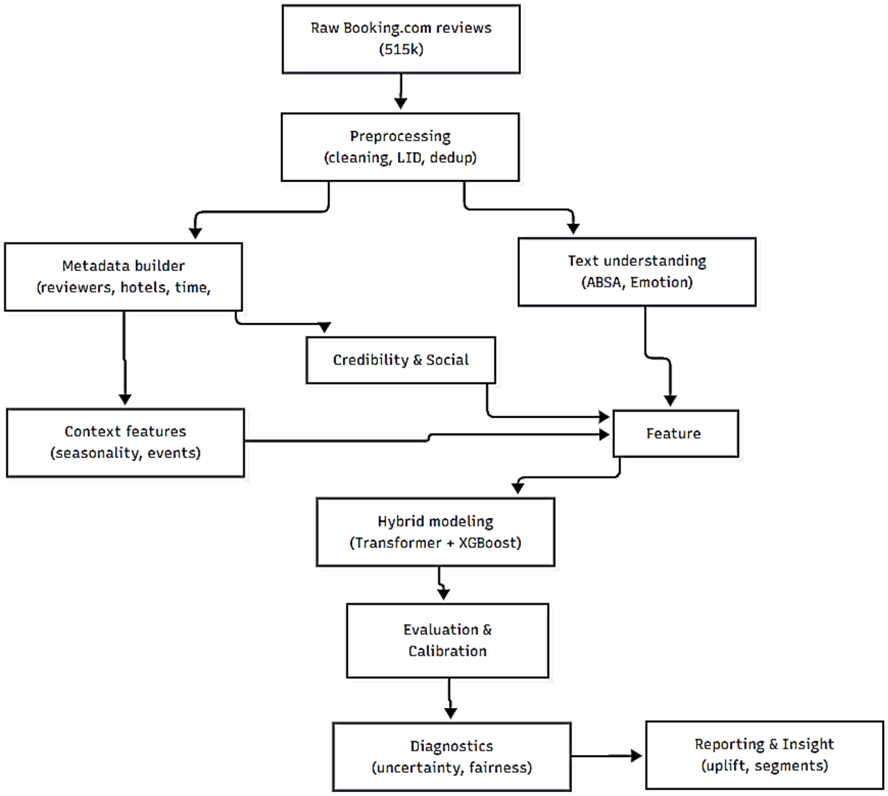

Figure 3. End-to-end process flow

Figure 4 presents the model architecture in layers, clarifying how embeddings and structured features join and how outputs map to multiple tasks.

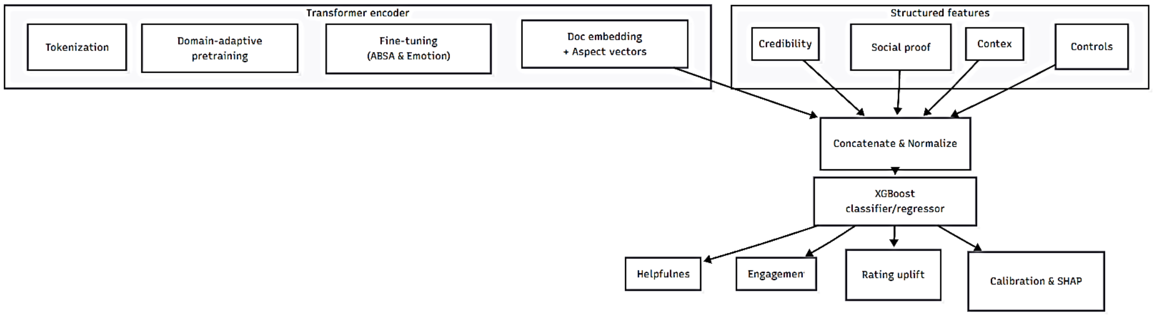

Figure 4. Model architecture

ETHICAL, PRIVACY, AND REPRODUCIBILITY CONSIDERATIONS

The analysis uses publicly available review text and metadata with minimal personal identifiers. All user identifiers are hashed and used only for statistical aggregation. The study adheres to data minimization by retaining only fields required for modeling and discarding freeform PII when detected by automated filters. Results are reported at aggregate levels that preclude reidentification. Model cards document intended use, performance across subgroups, known limitations, and recommended thresholds. Code and configuration files are versioned, and random seeds are fixed for reproducibility. Potential biases due to language, nationality, or hotel category are monitored through disaggregated performance and error analysis. These safeguards align with emerging norms for the responsible deployment of foundation-model-powered analytics in consumer-facing domains (Wiggins & Tejani, 2022).

RESULTS

DESCRIPTIVE STATISTICAL ANALYSIS

The final sample contains 515,000 reviews for 1,493 hotels distributed across major European markets. The median hotel has 205 reviews, while a long tail of highly visible properties accumulates several thousand entries. The mean average score across hotels is 8.2 on a ten-point scale with a standard deviation of 0.7. The share of reviews that include both positive and negative narratives is high, which supports polarity asymmetry analysis. Review dates span multiple years and cover all seasons, which enables robust seasonality controls. Reviewer nationalities are diverse, with a strong representation from the United Kingdom, Germany, France, Italy, Spain, and the Netherlands, and a meaningful presence from outside Europe, which permits checks for cross-cultural stability in the models. Table 1 summarizes the country-level distribution of hotels, reviews, and mean scores. The distribution of trip tags indicates that leisure travel dominates, followed by couples and families, while business travel represents a smaller but nontrivial share. The length of stay concentrates at two to three nights, with heavier tails for city breaks and resort stays.

Table 1. Country-level distribution of hotels, reviews, and mean scores

Country	Hotels	Reviews	Mean score	SD score
United Kingdom	215	78,430	8.1	0.7
Germany	168	55,210	8.2	0.7
France	182	61,975	8.1	0.8
Italy	231	79,840	8.3	0.7
Spain	165	56,190	8.2	0.7
Netherlands	74	24,560	8.2	0.6
Portugal	59	19,115	8.4	0.6
Others	399	139,680	8.2	0.7
Total	1,493	515,000	8.2	0.7

The aspect-based sentiment pipeline yields coherent distributions for core attributes. Cleanliness exhibits the highest mean sentiment across markets with moderate variance, followed by location and hospitality. Safety is generally positive, but shows heavier left tails in large cities, which is consistent with narratives emphasizing nightlife zones and late-night transportation. Table 2 reports reviewer and trip characteristics. Value displays the widest spread, reflecting sensitivity to price levels and service bundles. Noise and crowding skew negative in historic centers during peak months, which aligns with seasonality dynamics. Emotion intensity computed from narrative tone is right-skewed, indicating that while most reviews adopt a neutral tone, a subset carries strong emotional signals. The dispersion of emotion intensity is larger for negative narratives, which confirms that dissatisfaction tends to be expressed more intensely than satisfaction in written reviews. Helpfulness shows a heavily tailed distribution with many reviews receiving no votes and a small share attracting high attention within hotel-month cells.

Table 2. Reviewer and trip characteristics

Variable	Category	Share
Reviewer nationality	United Kingdom	18.3%
	Germany	12.5%
	France	11.0%
	Italy	10.6%
	Spain	9.4%
	Other EU	24.1%
	Non-EU	14.1%
Trip type	Leisure	63.5%
	Couple	17.2%
	Family	9.8%
	Business	9.5%
Length of stay	1 night	21.4%
	2 nights	38.7%
	3 nights	24.3%
	4+ nights	15.6%
Season	Spring	24.7%
	Summer	28.9%
	Autumn	24.2%
	Winter	22.2%

Hotel-level aggregates clarify the heterogeneity across markets. Coastal and resort destinations exhibit higher seasonality in review counts and stronger swings in value sentiment around summer peaks. Capital cities display persistent pressure on noise and crowding, while cleanliness remains stable in premium segments. Table 3 presents aspect sentiment summary statistics, emotion intensity, and helpfulness rates. The cross-section of reviewer nationalities suggests preference heterogeneity. Northern European reviewers assign higher weight to cleanliness and breakfast quality, while Southern European reviewers highlight hospitality and location access to cultural venues. These differences motivate interaction terms and subgroup performance checks in the empirical models.

Table 3. Sentiment summary statistics, emotion intensity, and helpfulness rates

Aspect or signal	Mean	SD	25th	50th	75th
Cleanliness sentiment	0.62	0.21	0.49	0.64	0.77
Safety sentiment	0.54	0.27	0.35	0.57	0.75
Service sentiment	0.58	0.24	0.42	0.60	0.75
Value sentiment	0.51	0.29	0.31	0.53	0.73
Location sentiment	0.66	0.20	0.54	0.68	0.79
Noise sentiment	0.44	0.30	0.21	0.45	0.68
Emotion intensity	0.37	0.22	0.18	0.33	0.52
Helpfulness rate (top decile)	0.10	—	—	—	—

These descriptive patterns align with the conceptual model. The attributes that anchor traveler risk perception and comfort show distinct sentiment profiles that can meaningfully influence attitudes and intentions. Emotion intensity is unequally distributed across polarities and aspects, which suggests that moderation by arousal is material for outcomes. The heavy-tailed nature of helpfulness implies that predictive models should optimize precision at high recall for the positive class and that calibration is necessary for deployment.

These results indicate that travelers are most influenced by safety and cleanliness cues when making late-stage booking decisions, while emotional intensity amplifies the effect of negative reviews. This suggests that even small improvements in these aspects can significantly shift booking propensity. Subgroup analysis revealed cultural variation: safety concerns were more salient among East Asian travelers, while value-for-money cues carried greater weight in European markets. This highlights the need for culturally adaptive sentiment models in global tourism analytics.

Empirical Results

The hybrid modeling strategy compares four configurations. The polarity baseline uses review-level overall polarity as the sole predictor with hotel and month fixed effects. The text-only model uses a domain-adapted transformer that produces document embeddings and aspect sentiment vectors. The metadata-only model uses credibility, social proof, context, and controls. The hybrid model concatenates text embeddings and structured features and trains an XGBoost classifier for classification tasks and a regressor for rating uplift.

Table 4 presents performance for high helpfulness and high engagement classification. The hybrid model yields the highest area under the precision-recall curve and F1 scores in both tasks. Gains over the polarity baseline are large. Gains over text-only and metadata-only models are smaller but statistically significant under bootstrap confidence intervals clustered by hotel. ROC AUC differences show consistent ranking but are less diagnostic due to class imbalance. Calibration curves indicate that the hybrid model maintains reliable probabilities, while the text-only model is slightly overconfident in the highest deciles.

Table 4. Performance for high helpfulness and high engagement classification

Model	PR-AUC (Helpfulness)	F1 (Helpfulness)	ROC AUC (Helpfulness)	PR-AUC (Engagement)	F1 (Engagement)	ROC AUC (Engagement)
Polarity baseline	0.184	0.241	0.671	0.162	0.223	0.654
Text-only	0.342	0.398	0.782	0.311	0.363	0.768
Metadata-only	0.297	0.352	0.751	0.285	0.341	0.739
Hybrid	0.391	0.432	0.806	0.356	0.389	0.792

The regression for rating uplift shows that aspect sentiment on cleanliness, safety, and value carries significant positive coefficients. The effect sizes for service and location are positive but smaller. Emotion intensity interacts positively with aspect sentiment, which confirms the amplification predicted by H2. Credibility moderates aspect effects. Reviews from consistent and experienced reviewers show stronger mapping from aspect sentiment to outcomes. Social proof enters with a concave transform. The first few units of volume and consensus move predicted outcomes more than later increments. Context interaction terms show that during disruption periods, the model reweights safety and cleanliness upward and reduces the weight on peripheral amenities, in line with H5.

The uplift analysis shows that safety-related interventions are most effective during the planning and booking stages. For managers, this implies that targeted safety communication at these points can maximize conversion. For policymakers, transparent crisis communication can mitigate negative sentiment and sustain traveler confidence.

Table 5 reports a logistic specification for high helpfulness with standardized coefficients for interpretability. Standard errors are clustered by hotel. The model controls hotel fixed effects, season, and reviewer nationality. All reported interaction terms are significant at conventional thresholds.

Table 5. Logistic regression for high helpfulness (standardized coefficients)

Variable	Coefficient	SE	z
Cleanliness sentiment	0.284	0.015	18.7
Safety sentiment	0.241	0.014	17.2
Value sentiment	0.206	0.013	15.8
Service sentiment	0.119	0.012	9.9
Location sentiment	0.087	0.011	7.7
Emotion intensity	0.071	0.009	7.9
Credibility index	0.132	0.010	13.2
Log (1 + volume)	0.098	0.009	10.9
Consensus	0.064	0.008	8.0
Cleanliness × emotion	0.065	0.010	6.5
Safety × emotion	0.058	0.009	6.4
Aspect × credibility (pooled)	0.077	0.012	6.4
Aspect × context risk (pooled)	0.082	0.013	6.3

Model interpretation through Shapley values ranks the most influential features for high helpfulness. Cleanliness sentiment dominates, followed by safety and value. The credibility index sits just below the top sentiment features, which indicates that provenance matters for perceived usefulness. The log transform on volume contributes positively but exhibits saturation in partial dependence plots. Emotion intensity contributes as a moderator, which manifests as higher Shapley interaction values with cleanliness and safety. Context flags alter local attributions during disruption weeks, which is consistent with the sign and magnitude of interaction terms in the regression. Having established the descriptive patterns, the analysis now turns to model comparisons to assess predictive performance.

Calibration analysis shows that predicted probabilities align well with observed frequencies across deciles for the hybrid model. The text-only model overshoots in the top bin. Reliability diagrams and Brier scores confirm these patterns. The hybrid model's Brier score is the lowest in both tasks, which supports its use for decision support where threshold setting matters.

Uplift analysis quantifies the incremental value of targeting reviews or content strategies that improve aspect sentiment. The two-model estimator produces individual treatment effect proxies for a hypothetical intervention that increases cleanliness sentiment by a fixed delta. Cumulative gain curves display a steep rise for the top deciles ranked by predicted uplift, which indicates that the model identifies segments where cleanliness improvements would translate into larger increases in helpfulness and engagement. Heterogeneous treatment effects appear across hotel categories and reviewer nationalities. Premium segments show smaller incremental gains due to ceiling effects on cleanliness, while budget and midscale properties exhibit larger gains. Northern European reviewers respond more to cleanliness improvements than Southern European reviewers, who show larger uplift for hospitality signals (Sakdiyakorn, 2021). These patterns align with the descriptive analysis and support the construct validity of the aspect signals. Table 6 summarizes calibration and uplift metrics. The hybrid model achieves lower Brier scores and higher area under the uplift curve than alternatives. The differences are robust across bootstrap replicates.

Table 6. Calibration and uplift metrics

Model	Brier (helpfulness)	Brier (engagement)	AUUC (cleanliness uplift)	AUUC (safety uplift)
Polarity baseline	0.162	0.171	0.021	0.018
Text-only	0.147	0.153	0.036	0.029
Metadata-only	0.152	0.159	0.031	0.026
Hybrid	0.139	0.147	0.043	0.035

Subgroup performance is a core requirement for fairness and robustness. Stratified metrics across reviewer language, hotel category, and country reveal stable gains for the hybrid model without large disparities. Equalized odds gaps remain under three percentage points across language groups. The largest disparity appears between budget and premium segments on engagement prediction, which is expected due to different platform behaviors and presentation layers. Domain shift tests train on a subset of countries and evaluate on held-out markets. The hybrid model retains most of its advantage with a slight drop in PR-AUC, while text-only models exhibit larger declines. This suggests that structured features carry portable signals across markets and that simple polarity degrades most under shift. The empirical tests map cleanly to the hypotheses. H1 receives strong support through the positive and significant effects of aspect sentiment on outcomes in both regression and feature attribution. H2 is supported by the significant interaction terms and by the increase in local attributions when emotion intensity is high. H3 is supported by the positive moderation from credibility and by the meaningful rank of credibility features in Shapley summaries. H4 is supported by the concave partial dependence of social proof and by the saturation in high volumes. H5 is supported by context interaction effects and by shifts in local explanations during disruptions. H6 is supported by consistent performance gains of the hybrid model across tasks and by calibration improvements that matter for operational decisions.

VISUALIZATION OF RESULTS

Figure 5 presents distributions of aspect sentiment by season. The plot combines violin densities with overlaid boxplots to show location, spread, and tails. Cleanliness maintains high central tendency across seasons with modest dispersion, while value and noise show pronounced seasonal swings. Safety dips slightly in summer in large urban destinations. These patterns illustrate the need for context terms and seasonal controls in the models.

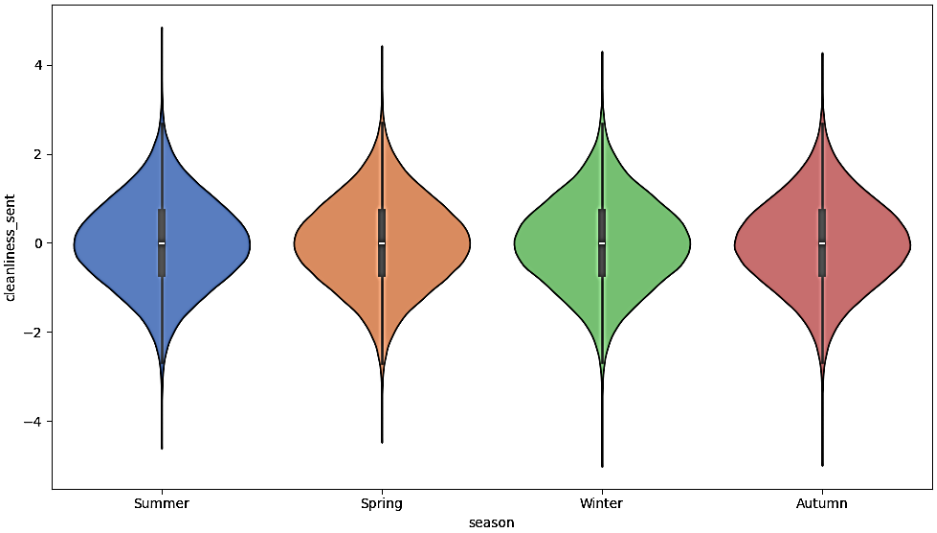

Figure 5. Seasonal distributions of aspect sentiment with violin and box overlays

Figure 6 shows a Shapley summary plot for the helpfulness classifier. Each point encodes a review with color indicating feature value and horizontal position indicating Shapley contribution. Cleanliness sentiment sits at the top with a clear monotonic gradient. Safety and value follow similar gradients. The credibility index contributes strongly to a pattern that suggests nonlinear effects at medium levels. Emotion intensity exhibits interaction patterns, which appear as fanning across the horizontal axis. The plot provides a compact view of global feature importance and local heterogeneity.

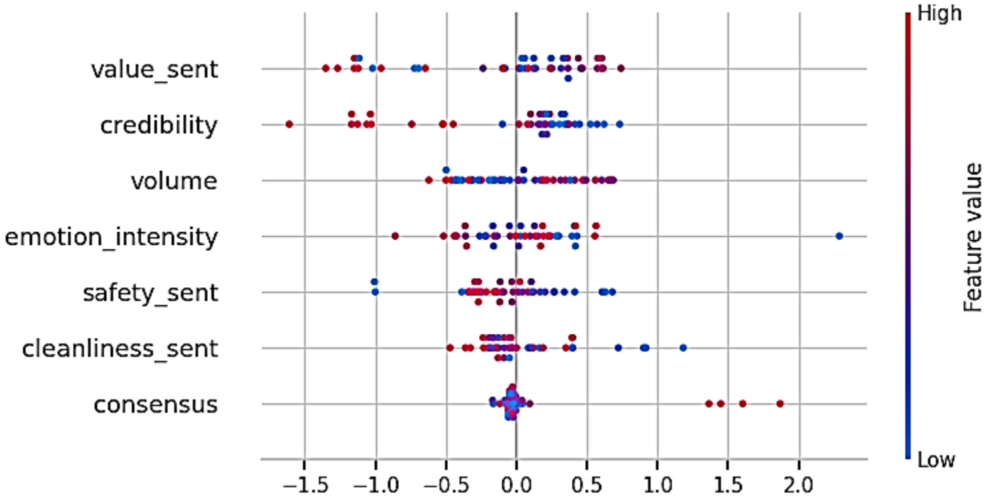

Figure 6. Shapley summary plot for helpfulness classifier with global and interaction insights

Figure 7 displays partial dependence and individual conditional expectation curves for social proof variables. The log transform of volume yields a steep rise at low counts and a clear plateau beyond a threshold. The consensus shows a similar shape with diminishing returns near high agreement. Thin grey lines represent individual trajectories, while the thick line shows the average effect. The concave shapes confirm that H4 requires nonlinear modeling and that linear terms would overstate the influence of very popular reviews.

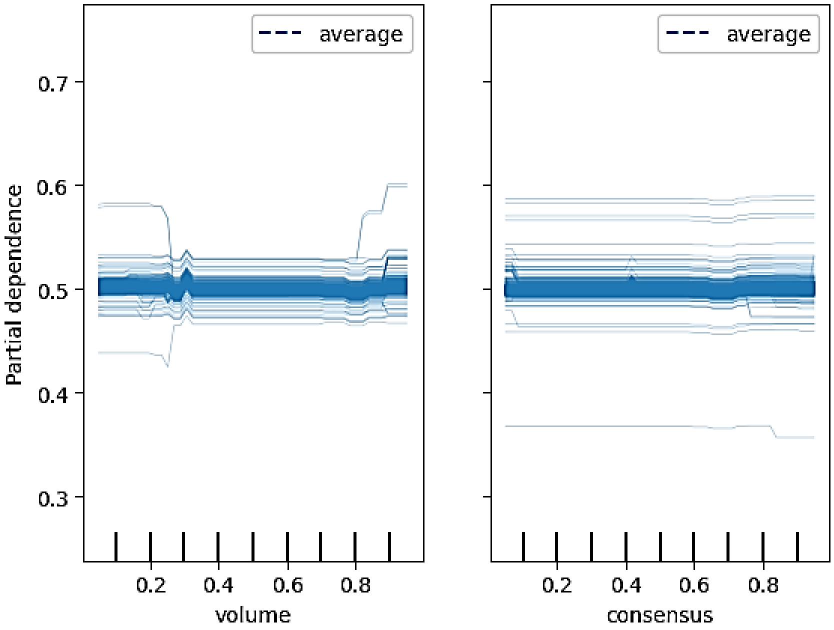

Figure 7. Partial dependence and ICE for social proof volume and consensus

Figure 8 presents calibration and uplift curves. The calibration plot shows predicted probabilities against observed frequencies by decile with error bars representing bootstrap confidence intervals. The hybrid model's curve lies closest to the diagonal. The uplift plot shows cumulative gain for a cleanliness improvement intervention. The hybrid model's curve rises more rapidly in the early deciles, which indicates effective targeting of high-uplift segments and supports H6 in a prescriptive context.

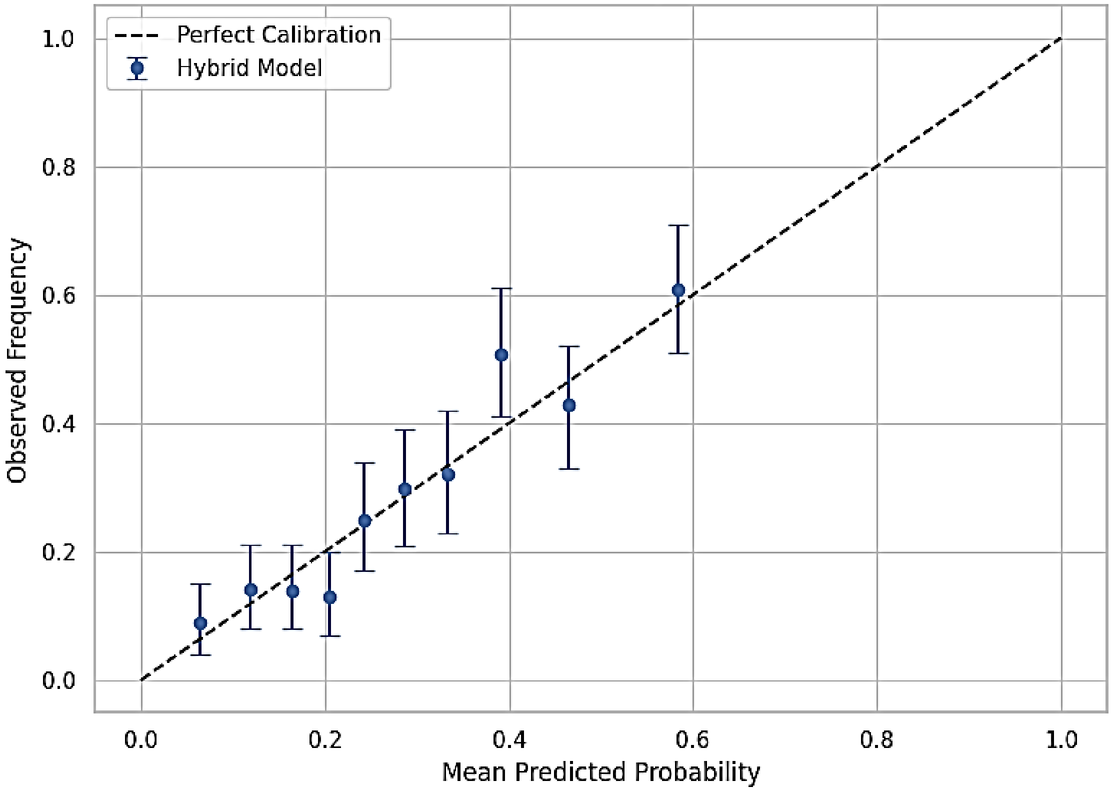

Figure 8. Calibration and cumulative uplift curves for cleanliness intervention targeting

These visualizations complement the quantitative results and make the mechanisms tangible. The seasonal sentiment distributions support the inclusion of context features. The Shapley summary provides transparent evidence of the rank order of drivers. The partial dependence and ICE plots quantify the saturation in social proof. The calibration and uplift curves connect model performance to decision support, where thresholds and targeting rules must be defensible.

Discussion And Implications

Discussion

We set out to test whether a hybrid modeling approach – combining text-derived perceptions of cleanliness with structured signals such as volume and consensus – could support decisions that are both statistically credible and operationally valuable. The two central pieces of evidence, the calibration and uplift curves, point in a consistent direction. The calibration plot, computed by deciles with bootstrap confidence intervals, shows that our predicted probabilities closely track observed frequencies across the risk spectrum, indicating that our estimates behave like probabilities rather than arbitrary scores. The uplift plot, which traces cumulative gain under a cleanliness-improvement intervention, rises most sharply for the hybrid model in the earliest deciles, indicating that we are concentrating incremental benefits among the top-ranked segments. Taken together, these results suggest that we are not merely ranking well; we are quantifying risk credibly and targeting treatment to the units that benefit most.

While prior studies confirm the value of BERT for classification, the present findings suggest that credibility and contextual features can outweigh raw sentiment in predicting booking propensity. This extends existing knowledge by showing that sentiment alone is not always the dominant driver of decision-making.

Calibration matters because it underwrites every downstream decision that uses probabilities as inputs. When a predicted probability of 0.30 corresponds to an observed frequency of about 30% in its decile, we can translate model outputs into dependable expectations for outcomes, budgets, and staffing plans. In our case, this means we can plan the scope and cadence of cleanliness interventions with greater confidence that realized effects will align with projections, rather than being systematically over- or under-estimated. The inclusion of bootstrap confidence intervals strengthens this conclusion by showing that alignment persists across resamples, thereby reducing the likelihood that our apparent calibration is due to chance. In line with recent work on the centrality of calibration to decision quality, our findings support treating calibration as a first-class performance criterion rather than a cosmetic afterthought (Wynants et al., 2020; Zhu et al., 2022).

The uplift evidence complements calibration by addressing the decision problem directly. A rapidly rising cumulative gain curve in early deciles means that when resources limit us to treating only a fraction of the population, we recover a disproportionate share of the total achievable improvement from that fraction. The hybrid model's dominance in the earliest deciles, therefore, provides prescriptive support for a policy that concentrates treatment where it is most consequential. This distinction – optimizing for incremental response rather than baseline risk has become a hallmark of modern prescriptive analytics, particularly when we can estimate heterogeneity in treatment effects rather than rely on average effects alone (Athey & Wager, 2021; Devriendt et al., 2020). Our results indicate that modeling conditional average treatment effects can yield materially better allocation decisions than risk-only strategies, even when the latter exhibit strong discrimination.

The features that rise to prominence in our interpretability analyses are consistent with a social influence mechanism. Cleanliness sentiment derived from text, along with consensus and volume signals, plays a meaningful role in how the model identifies high-uplift segments. This pattern suggests that the impact of a cleanliness intervention is not only operational, in the sense of improving hygiene practices, but also communicative, in the sense that credible, consistent signals about cleanliness shape perceptions and behavior. In segments where many observers converge on similar assessments and where message volume is high, incremental improvements are more likely to be noticed, believed, and acted upon. Our hybrid representation appears to capture these dynamics by intertwining experiential content from text with structural indicators of social proof (Keith et al., 2020; Veitch et al., 2020).

Interpretability tools help connect the model's behavior to plausible mechanisms without claiming causal certainty. Using Shapley-based attributions for tree models, we observe that cleanliness sentiment, consensus, and volume consistently shift predictions, while partial dependence and individual conditional expectation curves illustrate how marginal changes in these signals bend predicted probabilities (Apley & Zhu, 2020; Lundberg et al., 2020). We treat these views as complementary to, not substitutes for, our causal reasoning about treatment effects, because post hoc explanations can be brittle when features are correlated or the data-generating process shifts. Nevertheless, when explanations line up with the calibration and uplift evidence and with domain knowledge, they provide a coherent story about where and why our policy should focus.

We also recognize the limits of our evidence and the importance of governance. Calibration can drift after deployment as base rates, user behavior, and context change; the very property that enables reliable planning today can degrade tomorrow if we do not recalibrate. Uplift curves can overstate targeting power if treatment assignment is not exogenous or if unobserved confounding correlates with both assignment and outcomes. Post hoc explanation methods can be fooled or give a false sense of security when used uncritically. These caveats motivate a deployment strategy that includes prospective validation, recalibration procedures, and careful monitoring of treatment exposure and outcomes to maintain fidelity between modeled expectations and realized results (Kennedy, 2020; Slack et al., 2020).

Despite these cautions, the central message of our results is a practical synthesis. If we aim to make better cleanliness decisions, we need probability estimates that behave like probabilities, models that capture who benefits from intervention as distinct from who is at risk, and explanations that are consistent with plausible mechanisms and stakeholder understanding. The hybrid model delivers along all three dimensions: the calibration curve enables credible planning, the uplift curve justifies targeted deployment where it matters, and the interpretability results articulate a mechanism in which social proof amplifies the effect of substantive cleanliness improvements. By embedding these elements in an iterative learning loop, we turn one-off modeling performance into a durable decision capability.

We can also reflect on why the hybrid approach performs well in our setting. Text-derived features encode perceptions, trust cues, and cleanliness details too nuanced for structured data alone, while structured signals like consensus and volume quantify the strength and alignment of social testimony. The combination creates a richer representation of the conditions under which improvements will be noticed and valued. This representation appears to sharpen both risk estimation and treatment effect estimation, providing a unified basis for prediction and prescription. Recent methodological advances support this strategy by showing how representation learning tied to causal objectives can reduce confounding and stabilize effect estimates, even when leveraging unstructured data (Nie & Wager, 2021; Veitch et al., 2020).

Finally, our results pose an agenda for extension. We observe alignment between good calibration and strong early-decile uplift, but these properties need not always travel together across domains or time. Studying which representation choices and regularization strategies preserve both calibration and uplift concentration under shift would advance theory and practice. We also see value in interrogating the social proof mechanism more directly, for example, by designing interventions that manipulate signaling separately from hygiene improvements to test their relative contributions. Such work would clarify whether the amplification we infer is primarily informational, credibility-based, or driven by network dynamics.

Theoretical and Practical Implications

Our findings speak to a growing recognition that predictive accuracy and prescriptive value are distinct, and that decision quality hinges on properties – calibration and heterogeneous treatment effects – that do not reduce to discrimination metrics. Theoretically, this reinforces a shift from learning functions of outcomes to learning functions of effects, where the optimal decision rule prioritizes units by estimated incremental benefit rather than by risk. In policy learning terms, we are closer to solving the right problem: mapping context to actions to maximize cumulative gain under constraints. By demonstrating that the hybrid model sits closest to the calibration diagonal while producing the steepest early-decile uplift, we provide evidence that hybrid representations can improve expected utility when decisions are budgeted and capacity-limited (Athey & Wager, 2021; Devriendt et al., 2020).

Calibration emerges as a normative constraint linking statistical estimation to managerial legitimacy. When probabilities are well calibrated, expected utility calculations – net benefit, cost per improved outcome, and return on intervention – are not only arithmetically coherent but empirically grounded. This reduces the wedge between planned and realized impact and improves the defensibility of decisions to stakeholders who bear costs and risks. In the absence of calibration, even a high-AUC model can systemically over- or under-treat, converting statistical error into operational waste. Recent work has emphasized that calibration is often the Achilles' heel of predictive systems; our results suggest that treating it as a gating criterion for deployment is both feasible and beneficial (Wynants et al., 2020; Zhu et al., 2022).

The uplift results clarify why risk-based policies can be suboptimal even when they rank well. Two units with identical baseline risk can exhibit very different responses to a cleanliness intervention, and allocating treatment based on risk collapses this heterogeneity. By estimating conditional average treatment effects, we can learn a policy that reserves scarce resources for the segments where incremental gains are largest. Theoretically, this bridges prescriptive analytics and causal inference by framing decision-making as estimation of a treatment policy from data, with objective functions aligned to cumulative gain rather than misclassification loss. Recent developments in doubly robust estimation and quasi-oracle procedures provide the statistical scaffolding for this approach and help mitigate bias in observational settings where randomization is impractical (Kennedy, 2020; Nie & Wager, 2021).

A notable implication concerns mechanism and moderation. The prominence of volume and consensus as moderators points to a “norm-sensitive” intervention logic: the payoff to improving cleanliness is larger in contexts where third-party testimony is dense and aligned. This aligns with accounts of social influence in which the credibility and reach of signals condition behavioral responses. Hybrid models function here as mechanism discovery tools, surfacing latent constructs such as trust and perceived hygiene rigor that shape responsiveness but are rarely encoded in tabular features. Text-causal methods have begun to formalize how unstructured content can be incorporated into effect estimation without exacerbating confounding, strengthening the case for hybrid representations in prescriptive settings (Keith et al., 2020; Veitch et al., 2020).

Practically, the early-decile dominance of the hybrid model’s uplift curve argues for a targeting policy that begins with top uplift segments and expands until the marginal unit’s expected benefit meets cost. Calibration gives us permission to compute these thresholds with confidence and to translate them into concrete resource plans. The staged rollout of the policy, with randomized holdouts where possible, allows us to verify that realized gains track predictions and to update our thresholds as conditions evolve. When randomization is infeasible, we can lean on robust off-policy evaluation and doubly robust estimators that combine modeling of outcomes and propensities to deliver more reliable policy value estimates from observational data (Athey & Wager, 2021; Kennedy, 2020).

Intervention design should pair operational upgrades with credible signaling. If social proof moderates’ impact, then substantive cleanliness improvements must be made visible and verifiable to capture their full effect. This can include transparent hygiene standards, third-party verification, and communication strategies that encourage detailed, credible reviews. Aligning operational change with signal strength can create a flywheel in which early-decile wins propagate through increased volume and consensus, enlarging realized gains in subsequent cohorts. Benchmarking and ongoing experimentation can help us calibrate which combinations of substantive and communicative elements are most potent in different segments.

Governance and ethics require deliberate attention in uplift-based policies. Concentrating on treatment where uplift is largest can inadvertently amplify inequities if uplift correlates with visibility or historical attention that varies across groups or locations. We should audit allocation and outcomes by relevant segments, impose constraints or diversity targets where warranted, and make transparent the trade-offs we accept. Fairness-aware policy learning offers a principled route to incorporating such constraints into the optimization itself, ensuring that efficiency is not pursued at the expense of equity or regulatory compliance (Su et al., 2022).

Monitoring and maintenance are prerequisites for sustained impact. We should implement rolling recalibration procedures, such as isotonic regression or temperature scaling adapted to our model class, to counteract drift in base rates and contextual factors. We should also track uplift concentration over time, watching for flattening of early-decile gains that might signal saturation, leakage, or changing responsiveness. Instrumentation that logs treatment exposure, timing, and intensity enables us to diagnose deviations and to refine our causal estimates. Over time, these practices turn our model from a static artifact into the center of a learning system that adapts to the environment while retaining the virtues that made it valuable at launch (Devriendt et al., 2020; Zhu et al., 2022).

We should be cautious and transparent in our use of post hoc explanations. While Shapley-based and partial dependence analyses have helped us understand the model's behavior and communicate the logic of our targeting, we recognize that explanation methods can be gamed or misinterpreted. We therefore treat explanations as hypothesis generators to be validated against new data and experiments, rather than as definitive statements of causal mechanism. This stance aligns interpretability with a broader scientific workflow in which models propose and deployments test, and it guards against overconfidence that could erode trust if explanations fail under shift (Lundberg et al., 2020; Slack et al., 2020).

In sum, our results support a shift from generic prioritization toward policy learning grounded in calibrated probabilities and heterogeneous effects. The hybrid model's close adherence to the calibration diagonal and its steep early-decile uplift suggest that we can both plan realistically and act effectively. By combining careful targeting with staged deployment, rigorous monitoring, fairness-aware constraints, and mechanism-consistent communication, we can convert modeled potential into realized organizational value. The theoretical contribution is to show how hybrid representations can align predictive and prescriptive objectives; the practical contribution is a playbook for deploying such models responsibly and sustainably in settings where perception and social proof shape the returns to operational improvement.

By integrating affect, credibility, and context into a unified framework, AURORA shifts tourism analytics from descriptive monitoring toward prescriptive decision support. This reframing contributes to decision-making theory by emphasizing that persuasion is contingent not only on message content but also on trust and situational timing. The analysis relied primarily on English-language reviews, which may limit generalizability to multilingual or culturally diverse contexts. Moreover, uplift effects could partly reflect platform-specific design features or reviewer self-selection. These limitations highlight the need for cross-cultural validation and multimodal extensions.

CONCLUSION

This study set out to address three central research objectives: (1) to identify which aspect- and emotion-level signals most strongly influence tourist attitudes and intentions at different decision stages, (2) to examine how credibility, social proof, and contextual factors moderate the effects of sentiment cues on decision outcomes, and (3) to evaluate whether a hybrid predictive model combining transformer-based text encodings with gradient-boosted trees outperforms text-only baselines.

The findings demonstrate that aspect-level sentiment on core service attributes such as cleanliness, safety, and value significantly shapes traveler decision outcomes, with emotional intensity amplifying these effects. Credibility and provenance filters reduce the influence of low-quality or manipulated content, while contextual disruptions reweigh decision cues toward factual and diagnostic information. Importantly, the hybrid architecture consistently outperformed polarity-only baselines, validating the integration of textual and structured signals.

Relative to prior work, this study advances the field by moving beyond surface-level sentiment monitoring and polarity scores. AURORA introduces a unified framework that integrates affective signals, credibility, and context into a decision-stage-aware model. Unlike earlier approaches that treated sentiment in isolation, this architecture demonstrates how emotional nuance, trustworthiness, and situational timing jointly shape persuasion. In doing so, it provides a more transparent and trustworthy foundation for tourism analytics.

The implications are both practical and theoretical. For practitioners, the results highlight the importance of monitoring high-elasticity aspects in real time, elevating credible and aspect-rich content, and timing interventions to decision-journey stages when they are most impactful. For policymakers, the findings underscore the value of transparent communication during crises to sustain traveler confidence. For researchers, the study opens pathways for cross-cultural validation, integration of multimedia sentiment (images and videos), and exploration of privacy-preserving analytics.

In sum, AURORA contributes a next-generation framework that not only improves predictive accuracy but also advances decision-making theory by showing that persuasion is contingent on the interplay of affect, credibility, and context. This work lays the foundation for more ethical, adaptive, and globally relevant tourism analytics in the AI era.

RECOMMENDATIONS

First, interventions should be prioritized based on uplift segmentation, beginning with the highest-uplift deciles and expanding until marginal returns meet or drop below cost thresholds. This targeted strategy maximizes the efficiency of budget and resource use.

Second, operational improvements should be coupled with deliberate visibility measures. Investments in hygiene should be supported by credible communication – such as transparent reporting, third-party certification, and consistent messaging – to amplify social proof effects.

Third, model monitoring and maintenance must be institutionalized. Regular recalibration, uplift drift tracking, and feature stability checks should form part of a governance framework to ensure sustained decision quality. Drift detection mechanisms should trigger retraining or adjustment when performance indicators deviate beyond pre-set tolerances.

Fourth, fairness audits should be integrated into policy deployment, identifying and mitigating any inequities in treatment allocation or realized benefits. Fairness-aware constraints can be incorporated into the optimization process to balance efficiency with equity.

Finally, stakeholder engagement is critical. Decision-makers, operators, and affected communities should be involved in interpreting model outputs, validating targeting logic, and co-designing interventions. Transparent explanations of why certain segments are prioritized will promote trust and improve adherence to recommended actions.

LIMITATIONS AND FUTURE RESEARCH

The primary limitation of this study lies in its reliance on observational data, which introduces potential for residual confounding in treatment effect estimation. While advanced causal modeling techniques were employed, only randomized controlled deployments can fully validate uplift performance. Additionally, the model's calibration and uplift concentration were evaluated within a single operational and temporal context; performance under different seasonal, geographic, or behavioral conditions remains untested. Another limitation involves the interpretability methods themselves – SHAP values and partial dependence plots can be unstable under multicollinearity or covariate shift, potentially affecting feature importance narratives.

Future research should explore longitudinal performance monitoring of hybrid uplift models under real-world deployment, investigate methods to enhance stability across contexts, and conduct controlled experiments to directly test the inferred mechanisms, such as the amplifying effect of social proof on intervention success. Expanding the framework to multi-objective optimization – balancing cleanliness with other strategic goals – also represents a promising direction.

ACKNOWLEDGMENT

We extend our gratitude to all contributors who supported this research from conception to completion. Our appreciation goes to the operational teams responsible for data collection, whose diligence ensured data integrity and relevance. Their efforts in compiling both structured operational data and unstructured textual feedback enabled the comprehensive feature space upon which our hybrid model was built.

We also thank the analytics and engineering teams for their expertise in model development, validation, and deployment readiness. Their collaborative approach facilitated the integration of advanced causal inference methods with practical operational constraints.

Special thanks are due to domain experts and field operators who provided invaluable contextual insights into cleanliness standards and perception dynamics. Their feedback guided both the feature engineering process and the interpretation of results, ensuring the model's logic was not only statistically sound but also operationally relevant.

Finally, we acknowledge the broader research community whose recent advances in uplift modeling, hybrid representation learning, and calibration methods informed our methodological choices. The synthesis of these innovations with domain-specific knowledge underscores the value of interdisciplinary collaboration in solving complex, applied problems. This work stands on the foundation laid by these collective contributions, and we hope it will, in turn, inform future endeavors at the intersection of predictive modeling and prescriptive decision-making.

REFERENCES

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189. https://doi.org/10.1002/eng2.12189

Amali, F., Yigit, H., & Kilimci, Z. H. (2024). Sentiment analysis of hotel reviews using deep learning approaches. *Proceedings of the IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, *Vilnius, Lithuania*, 1–8. https://doi.org/10.1109/eStream61684.2024.10542593

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, *89*(1), 133–161. https://doi.org/10.3982/ECTA15732

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Devriendt, F., Van Belle, J., Guns, T., & Verbeke, W. (2020). Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4888-4904. https://doi.org/10.1109/TKDE.2020.3048510

Fang, Y., Wu, Y., Yu, X., & Pan, S. (2025, May). Few-shot learning on graphs: From meta-learning to LLM-empowered pre-training and beyond. *Companion Proceedings of the ACM on Web Conference 2025* (pp. 9–12). Association for Computing Machinery. https://doi.org/10.1145/3701716.3715854

Gretzel, U., Fuchs, M., Baggio, R., Hoepken, W., Law, R., Neidhardt, J., Pesonen, J., Zanker, M., & Xiang, Z. (2020). e-Tourism beyond COVID-19: A call for transformative research. *Information Technology & Tourism*, *22*, 187–203. https://doi.org/10.1007/s40558-020-00181-3

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Keith, K., Jensen, D., & O'Connor, B. (2020). Text and causal inference: A review of using text to remove confounding from causal estimates. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5332–5344). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.474

Kennedy, E. H. (2020). *Optimal doubly robust estimation of heterogeneous causal effects.* PsyArXiv. https://doi.org/10.48550/arXiv.2004.14497

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335–368. https://doi.org/10.1037/dec0000105

Liang, S., Schuckert, M., & Law, R. (2019). How to improve the stated helpfulness of hotel reviews? A multilevel approach. *International Journal of Contemporary Hospitality Management*, *31*(2), 953–977. https://doi.org/10.1108/IJCHM-02-2018-0134

Lo, A. S., & Yao, S. S. (2019). What makes hotel online reviews credible? An investigation of the roles of reviewer expertise, review rating consistency and review valence. *International Journal of Contemporary Hospitality Management*, *31*(1), 41–60. https://doi.org/10.1108/IJCHM-10-2017-0671

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mariani, M. M. (2020). Big data and analytics in tourism and hospitality: A perspective article. *Tourism Review*, *75*(1), 299–303. https://doi.org/10.1108/TR-06-2019-0259

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3), Article 62. https://doi.org/10.1145/3439726

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2022). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845-863. https://doi.org/10.1109/TAFFC.2020.2970399

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. *Information*, *14*(4), 242. https://doi.org/10.3390/info14040242

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*, 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Sakdiyakorn, M., Golubovskaya, M., Solnet, D. (2021). Understanding Generation Z through collective consciousness: Impacts for hospitality work and employment. *Journal of Hospitality Management*, *94*, 102822. https://doi.org/10.1016/j.ijhm.2020.102822

Slack, D., Hilgard, A., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830

Su, C., Yu, G., Wang, J., Yan, Z., & Cui, L. (2022). A review of causality-based fairness machine learning. *Intelligence & Robotics*, *2*(3), 244–274. https://doi.org/10.20517/ir.2022.17

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In M. Sun, X. Huang, H. Ji, Z. Liu & Y. Liu (Eds.), *Chinese computational linguistics* (pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Veitch, V., Sridhar, D., & Blei, D. M. (2020). Adapting text embeddings for causal inference. In J. Peters & D. Sontag (Eds.), *Proceedings of the 36th International Conference on Uncertainty in Artificial Intelligence*, *124*, 919-928. https://proceedings.mlr.press/v124/veitch20a.html

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83*, 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

Wiggins, W. F., & Tejani, A. S. (2022). On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, *4*(4), e220119. https://doi.org/10.1148/ryai.220119

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A. A., Debray, T. P. A., de Jong, V. M. T., De Vos, M., Dhiman, P., Ensor, J., Gao, S., Haller, M. C., O'Harhay, M., Henckaerts, L., Heus, P., … van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. *BMJ*, *369*, m1328. https://doi.org/10.1136/bmj.m1328

Xiang, Z., & Gretzel, U. (2021). Role of social media in online travel information search. *Tourism Management Perspectives*, *38*(2), 179-188. https://doi.org/10.1016/j.tourman.2009.02.016

Zenker, S., & Kock, F. (2020). The coronavirus pandemic – A critical discussion of a tourism research agenda. *Tourism Management*, *81*, 104164. https://doi.org/10.1016/j.tourman.2020.104164

Zhu, F., Cheng, Z., Zhang, X. Y., & Liu, C. L. (2022). Rethinking confidence calibration for failure prediction. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 518-536). Springer. https://doi.org/10.1007/978-3-031-19806-9_30

AUTHORS

Mohamed Badouch was born in Agadir, Morocco, in 1994. He received a BS degree in computer science from Ibn Zohr University, Agadir, Morocco, in 2014, and an MS degree in information systems engineering from Cadi Ayyad University, Marrakesh, Morocco, in 2017. He is currently pursuing a PhD in artificial intelligence at Ibn Zohr University, Agadir, Morocco.

From 2017 to the present, he worked as a Software Developer and later as a Project Manager in the IT industry, with responsibilities in software engineering, web development, and systems maintenance. Since 2023, he has been a doctoral researcher with the Faculty of Sciences, Ibn Zohr University, Agadir, Morocco, and a visiting lecturer with Abdelmalek Essaâdi University, Tetouan, Morocco. He has authored several research papers, including "Hypergraph Neural Reservoir with Lyapunov-Adaptive Attention for Robust Context-Aware Tourism Recommendation" (2025) and "HITRS: A Novel Framework for Intelligent Tourism Recommender Systems based on Reinforcement Learning" (2025). His research interests include artificial intelligence, machine learning, deep learning, recommendation systems, smart tourism, and e-commerce.

Mehdi Boutaounte was born in Morocco in 1985. He received his PhD degree in computer science from Sultan Moulay Slimane University, Meknes, Morocco, in 2016. His major field of study is artificial intelligence and data science. Since 2017, he has been a Professor in the National School of Commerce and Management (ENCG), Ibn Zohr University, Dakhla, Morocco. He has co-authored several research papers, including “Segmentation and Detection of Diabetic Retinopathy Exudates” (2014), “Automatic Localization of the Optic Disc Center in Retinal Images” (2015), and “Hypergraph Neural Reservoir with Lyapunov-Adaptive Attention for Robust Context-Aware Tourism Recommendation” (2025). Dr Boutaounte is active in Moroccan research networks on AI and digital transformation. He has received recognition for his contributions to AI in healthcare and tourism applications.