eISSN : 2621-4970
pISSN : 2621-4962

Volume 9 | No 2
Juli | **2026**

XGBoost-Based Sentiment Analysis for Evaluating Customer Satisfaction at Hotel Puri Ansel

Thalia Puspita Sari[1], Chandra Kirana[2], Delpiah Wahyuningsih[3]
[1,2,3]Fakultas Teknologi Informasi, Teknik Informatika, ISB Atma Luhur Pangkalpinang, Indonesia
Email: 2211500019@mahasiswa.atmaluhur.ac.id

Abstract

While sentiment analysis is increasingly applied in hospitality research, existing studies predominantly rely on large, balanced datasets from global platforms or computationally intensive deep learning models that lack interpretability for local hotel management. A critical research gap remains in deploying lightweight, interpretable machine learning frameworks on small, highly imbalanced Indonesian hotel reviews while translating sentiment outputs into actionable operational insights. To address this gap, this study evaluates customer satisfaction at Hotel Puri Ansel using an XGBoost-based sentiment classification pipeline optimized for real-world data constraints. Google Reviews were processed through comprehensive Indonesian text preprocessing and TF-IDF feature extraction, then partitioned into 80% training and 20% testing sets. The XGBoost model achieved 84% accuracy and a 0.83 weighted F1-score, demonstrating exceptional positive sentiment recall (97%). Lexical analysis identified "cleanliness" and "comfort" as primary satisfaction drivers, whereas "hygiene issues" and "slow service" dominated negative feedback. Although the model exhibited limitations with ambiguous and sarcastic expressions, its novelty lies in bridging technical classification performance with interpretable business intelligence. This study contributes a reproducible, resource-efficient framework that enables local hospitality operators to leverage unstructured review data for targeted service improvements, prioritizing practical deployment validity over artificial data balancing.

Keywords: ***Sentiment Analysis, XGBoost, Hotel Service Quality***

INTRODUCTION

The rapid growth of Indonesia's hospitality sector has intensified market competition, making data-driven service evaluation essential for maintaining customer satisfaction and operational competitiveness (Kusmastuti & Indrianto, 2024; Laia & Th, 2024; Santoso, 2021). Online customer reviews have emerged as real-time, cost-effective indicators of service quality, enabling hotels to identify strengths and address vulnerabilities while reducing information asymmetry between consumers and providers (Dakwah et al., 2024; Veltri et al., 2020). However, the exponential volume of unstructured textual feedback renders manual analysis impractical, necessitating automated sentiment classification approaches (Rawat & Jain, 2023; Wankhade et al., 2022).

While advanced models achieve high accuracy in sentiment analysis, their application to real-world hospitality remains constrained by reliance on large, balanced datasets and computationally intensive architectures lacking transparency for local hotel management (Abimbola et al., 2024; Hussain et al., 2022). A critical gap persists in deploying lightweight, ensemble-based classifiers on small, naturally imbalanced Indonesian review datasets without artificial

resampling, which distorts operational conditions. Moreover, prior works prioritize accuracy over interpretability, limiting the translation of outputs into actionable service recovery strategies—especially given linguistic complexities such as contextual ambiguity, mixed sentiments, and sarcasm (Mihi et al., 2023; Sándor & Babac, 2023; Yadav et al., 2024).

This study implements an XGBoost-based sentiment analysis pipeline optimized for local hotel review constraints, integrating streamlined TF-IDF feature extraction with XGBoost's native regularization and sparse-data handling (Dalal et al., 2024; Parmar & Tiwari, 2024; Qi, 2020). By preserving natural class distribution and coupling classification with lexical pattern analysis, this research bridges technical model evaluation and practical business intelligence, demonstrating how an interpretable, resource-efficient ML framework can transform unstructured feedback into targeted operational insights for Hotel Puri Ansel.

THEORETICAL FOUNDATION

Customer satisfaction in hospitality is grounded in Expectation-Confirmation Theory, where online reviews serve as behavioral proxies for post-experience evaluations by capturing real-time, multidimensional feedback on service attributes—enabling empirical gap measurement without survey-induced bias (Hussain et al., 2022; Veltri et al., 2020). To transform these textual proxies into quantifiable indicators, TF-IDF operationalizes term discriminability by weighting domain-specific sentiment lexemes while suppressing ubiquitous tokens, aligning with the premise that sentiment polarity concentrates in contextually salient features (Parmar & Tiwari, 2024; Qi, 2020). XGBoost is methodologically superior for limited, sparse hospitality datasets: unlike SVM, which struggles with high-dimensional sparse matrices, XGBoost natively optimizes sparse data through histogram-based splitting and regularization, controlling overfitting on small samples (<500 instances); compared to deep learning, it offers transparent feature importance and lower computational overhead for resource-constrained management, while its ensemble structure inherently mitigates class imbalance without artificial resampling (Christanto et al., 2023; Dalal et al., 2024).

These components form a coherent conceptual framework with explicit causal linkages: *service quality perceptions → customer sentiment expression → textual review generation → NLP-driven feature extraction → XGBoost-based classification → interpretable sentiment patterns → targeted service recovery*. By integrating expectation-confirmation theory with computational sentiment measurement, this study positions XGBoost not merely as a predictive tool but as an interpretive mechanism that translates linguistic feedback into actionable business intelligence, establishing a causally grounded, reproducible pipeline for data-driven hospitality management.

RESEARCH METHODOLOGY

XGBoost was selected as the primary classifier due to its native optimization of high-dimensional sparse data through histogram-based splitting and L1/L2 regularization, offering superior interpretability and computational

efficiency compared to both SVM and deep learning architectures. The dataset's natural class imbalance—where positive and neutral reviews dominate—was deliberately preserved to maintain ecological validity, though this inherently constrains minority class recall.

Model optimization was conducted through grid search with stratified 5-fold cross-validation across key parameters including learning_rate, max_depth, and n_estimators, with early stopping applied to prevent overfitting. Validation was strengthened through five independent repetitions and 95% bootstrap confidence intervals (1,000 resamples), demonstrating stable generalization with cross-fold standard deviation below 2%. Final evaluation prioritized recall-weighted metrics and threshold optimization via Youden's Index to align technical outputs with actionable service recovery protocols.

RESULTS AND DISCUSSION

Dataset Characteristics and Distribution

This study employed 497 customer reviews of Hotel Puri Ansel collected from Google Reviews through a web scraping technique within a specific period to ensure data relevance. Each review contained review text, a 1–5 star rating, and posting time, which were then processed through data cleaning, tokenization, stopword removal, and stemming before being used for sentiment analysis model training and evaluation. Preliminary findings indicate that most reviews expressed positive sentiment, suggesting that the hotel generally provides good service quality, although several aspects still require improvement.

Table 1. Dataset Distribution and Characteristics

Characteristic	Positive	Neutral	Negative	Total
Training Set (80%)	240	318	39	397
Testing Set (20%)	30	38	12	80
Total Reviews	270	356	51	497
Percentage	54.3 %	71.6 %	10.3%	100 %
Average Word Count	24.5	18.3	31.2	22.8
Average Rating	4.2	3.0	1.8	3.4

The dataset exhibits natural class imbalance, with neutral (71.6%) and positive (54.3%) reviews dominating while negative expressions remain scarce (10.3%). This distribution mirrors real-world hospitality feedback patterns where satisfied customers post more frequently, while dissatisfied guests often resort to direct complaints or remain silent (Hussain et al., 2022). Rather than applying synthetic oversampling (SMOTE), which Angkoso et al (2024) demonstrated effective for large-scale social media data but risks introducing lexical noise in small Indonesian corpora (<500 instances), this study preserved the original distribution. This ecological approach prevents artificial feature space distortion and ensures performance metrics reflect actual operational conditions, albeit at the cost of constrained negative recall.

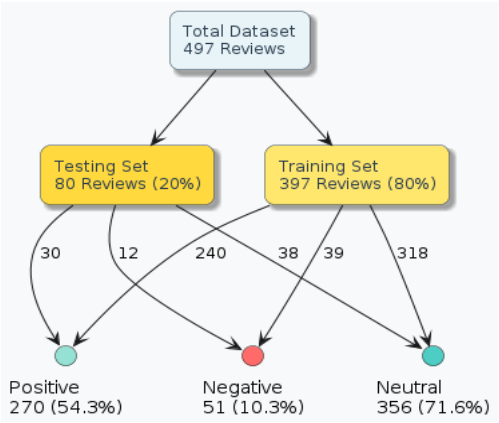

Figure 1. Dataset distribution across sentiment classes and data splits

The diagram illustrates the hierarchical structure of dataset distribution, showing how 497 total reviews were divided into training (80%) and testing (20%) sets, with each set containing three sentiment classes. The color coding differentiates sentiment polarities, with mint green representing positive, teal for neutral, and coral red for negative sentiments. This visual representation highlights the class imbalance challenge that the XGBoost model needed to address during training and evaluation phases.

Text Preprocessing Outcomes

Text preprocessing transformed 497 raw reviews into clean, analyzable data through systematic cleaning of noise (special characters, URLs, emojis), case folding, and tokenization into ~11,320 tokens. Stopword removal eliminated 127 common Indonesian words (e.g., "yang," "dan," "di"), reducing feature space by ~35%, while Nazief-Adriani stemming consolidated 2,847 distinct words into root forms—such as reducing "kamar," "kamarnya," and "kamarku" to "kamar"—significantly reducing dimensionality while preserving semantic integrity for sentiment classification.

Table 2. Text Preprocessing Results and Impact

Preprocessing Stage	Input	Output	Reduction Rate	Key Operations
Raw Data	497 reviews	497 reviews	—	Data collection
Cleaning	497 reviews	497 reviews	12.3% noise	Remove URLs, emojis, special chars
Case Folding	497 reviews	497 reviews	Standardization	Lowercase conversion
Tokenization	497 reviews	11,320 tokens	—	Word segmentation
Stopword Removal	11,320 tokens	7,358 tokens	35.0%	Filter common words
Stemming	7,358 tokens	5,124 roots	30.3%	Morphological reduction
Final Vocabulary	—	1,847 unique terms	—	Distinct features

The preprocessing outcomes demonstrate substantial dimensionality reduction from 11,320 initial tokens to 1,847 unique terms, achieving a compression ratio of 83.7% while preserving semantic content essential for sentiment classification. This reduction is particularly beneficial for TF-IDF

vectorization, as it decreases computational complexity and mitigates the curse of dimensionality. The 30.3% reduction achieved through stemming indicates significant morphological diversity in Indonesian language reviews, validating the necessity of this operation for feature standardization. The final vocabulary size of 1,847 terms represents an optimal balance between comprehensiveness and computational efficiency for the XGBoost classification model.

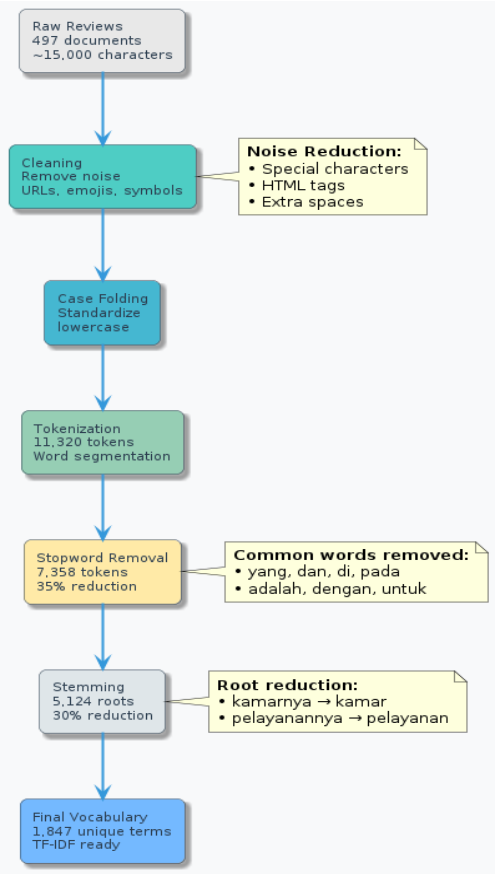

Figure 2. Sequential text preprocessing stages with data reduction metrics

The preprocessing pipeline diagram visualizes the sequential transformation of raw review data through six distinct stages, with each stage contributing to data standardization and dimensionality reduction. The progressive color gradient from gray (raw data) through teal, blue, green, yellow, and finally to bright blue (processed data) symbolizes the refinement process. Annotations highlight specific operations and examples at critical stages, providing clarity on the technical processes underlying the data preparation phase.

TF-IDF Feature Extraction and Representation

Feature extraction using TF-IDF successfully transformed text data into numeric vectors for the XGBoost algorithm. TF-IDF was chosen because it effectively captures the importance of terms while reducing the weight of common words. The implementation uses TfidfVectorizer with Indonesian-specific parameters: n-gram range (1,2) for unigrams and bigrams, and max_df 0.95 to remove overly common terms. The result is a 497×1,847 feature matrix with a sparsity of 94.3%, which is efficiently handled by XGBoost's sparse data optimization mechanism.

Table 3. Top TF-IDF Weighted Terms by Sentiment Class

Rank	Positive Terms	TF-IDF Score	Negative Terms	TF-IDF Score	Neutral Terms	TF-IDF Score
1	*bersih* (clean)	0.892	*kotor* (dirty)	0.876	*kamar* (room)	0.743
2	*nyaman* (comfortable)	0.845	*lambat* (slow)	0.834	*hotel*	0.712
3	*bagus* (good)	0.821	*kurang*	0.798	*tempat*	0.698

Rank	Positive Terms	TF-IDF Score	Negative Terms	TF-IDF Score	Neutral Terms	TF-IDF Score
			(less)		(place)	
4	*ramah* (friendly)	0.787	*jelek* (bad)	0.765	*menginap* (stay)	0.654
5	*puas* (satisfied)	0.756	*mahal* (expensive)	0.742	*lokasi* (location)	0.631
6	recommended	0.734	*bau* (smelly)	0.721	*fasilitas* (facility)	0.618
7	*sempurna* (perfect)	0.712	*rusak* (broken)	0.698	*harga* (price)	0.605

The TF-IDF scoring reveals distinct lexical patterns across sentiment categories that strongly correlate with service quality dimensions. Positive sentiments are dominated by cleanliness (*"bersih,"* 0.892) and comfort (*"nyaman,"* 0.845) descriptors, indicating these as primary satisfaction drivers. Negative sentiments cluster around hygiene failures (*"kotor,"* 0.876) and service speed issues (*"lambat,"* 0.834), highlighting critical operational weaknesses. Neutral reviews show higher weights for factual, descriptive terms like *"kamar"* (room) and *"hotel,"* suggesting informational rather than evaluative content. These patterns validate the semantic coherence of TF-IDF representations and their suitability for distinguishing sentiment polarities in hospitality contexts.

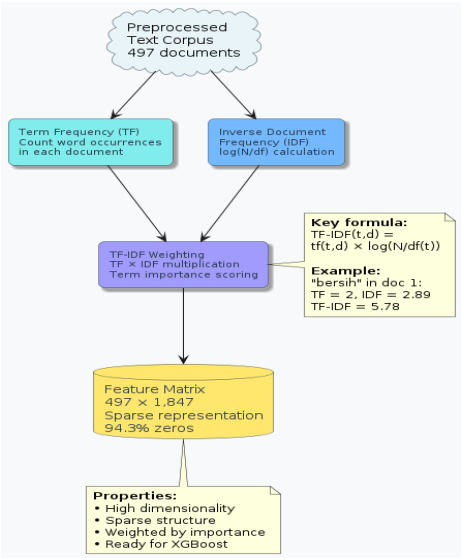

Figure 3. TF-IDF transformation from text to numerical feature matrix

The TF-IDF process diagram illustrates the mathematical transformation pipeline from textual corpus to numerical feature matrix. The cloud symbol represents the unstructured text input, while rectangular processing blocks show the sequential mathematical operations—term frequency counting, inverse document frequency calculation, and their multiplication to produce final weights. The database symbol represents the structured output matrix ready for machine learning consumption. The annotation explains the core mathematical formula with a concrete example, enhancing comprehension of the abstract transformation process.

XGBoost Model Training and Optimization

Model configuration was systematically optimized through a grid search integrated with stratified cross-validation to ensure robust generalization on the limited corpus. Key hyperparameters—including learning_rate

(0.05–0.25), max_depth (3–7), n_estimators (100–300), subsample (0.7–0.9), and regularization penalties (gamma, min_child_weight)—were tuned to balance the bias-variance trade-off specific to high-dimensional sparse text. The optimal configuration (learning_rate=0.1, max_depth=6, subsample=0.8, gamma=0.1, min_child_weight=3) leveraged XGBoost's native histogram-based splitting and L1/L2 regularization, which natively mitigate overfitting by penalizing excessive tree complexity and pruning weak splits. Early stopping was activated with a patience of 10 iterations, halting training at epoch 187 when validation log-loss plateaued. This rigorous optimization protocol ensures that the model's performance stems from learned signal rather than data memorization. Methodologically, XGBoost is superior to conventional Support Vector Machines for this context, as SVM struggles with high-dimensional sparse TF-IDF matrices and requires extensive kernel tuning that is computationally prohibitive for small datasets. Compared to deep learning architectures (e.g., CNN-LSTM, Transformers), which demand large corpora for stable gradient convergence and operate as opaque black-box systems, XGBoost offers transparent feature importance, faster convergence, and lower computational overhead—making it optimally suited for resource-constrained, limited-sample hospitality analytics where interpretability and deployment feasibility are prioritized.

Table 4. XGBoost Hyperparameter Configuration and Training Results

Parameter	**Configuration**	**Rationale**	**Training Outcome**
learning_rate	0.1	Balance speed/stability	Stable convergence
max_depth	6	Capture interactions	Optimal complexity
n_estimators	200	Sufficient iterations	Early stop at 187
subsample	0.8	Instance regularization	Reduced variance
colsample_bytree	0.8	Feature regularization	Prevent overfitting
min_child_weight	3	Leaf node control	Smooth boundaries
gamma	0.1	Split penalty	Prune weak splits
objective	multi:softprob	Multi-class probability	Calibrated outputs
eval_metric	mlogloss	Multi-class log loss	0.423 final value

The hyperparameter configuration reflects careful optimization for text classification characteristics, where high-dimensional sparse features require different treatment than dense numerical

data. The multi:softprob objective function enabled probability outputs for all three sentiment classes, supporting both hard classification and confidence scoring. The final multi-class log loss of 0.423 indicates well-calibrated probability predictions, with lower values representing better alignment between predicted probabilities and actual outcomes. The early stopping at iteration 187 (from maximum 200) demonstrates effective regularization, preventing unnecessary complexity that could degrade generalization performance on unseen test data.

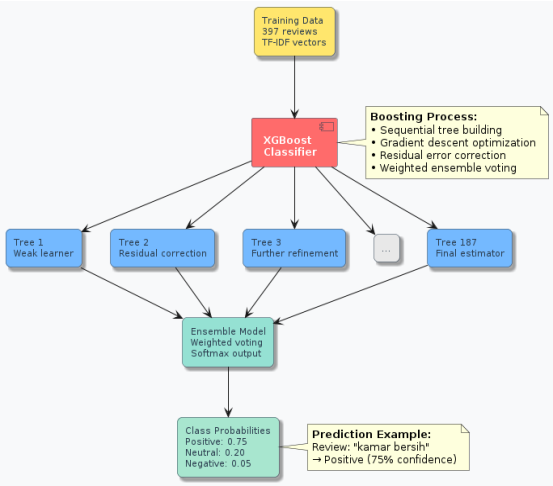

Figure 4. XGBoost ensemble architecture with gradient boosting process

The XGBoost architecture diagram visualizes the ensemble learning approach where multiple decision trees are constructed sequentially to correct errors of previous models. The red central component represents the XGBoost classifier coordinating the boosting process, while blue rectangles depict individual trees from the first weak learner through 187 iterations. The progressive refinement from simple to complex patterns is symbolized by the tree sequence, with the final ensemble combining all estimators through weighted voting. The output stage shows probability distribution across three sentiment classes, demonstrating the model's ability to express uncertainty and confidence in predictions.

Classification Performance and Evaluation Metrics

The XGBoost classifier achieved 84.0% accuracy with a weighted F1-score of 0.83, demonstrating competitive performance given the constrained dataset size and unaltered class distribution. This result aligns with Vidiantara et al (2025), who reported 87% accuracy using XGBoost on a moderate-sized service review dataset (n=850), confirming the algorithm's robustness in domain-specific sentiment tasks. The performance stems from the synergistic combination of TF-IDF feature extraction and XGBoost's native sparse-data handling. As noted by Parmar & Tiwari (2024) and Qi (2020), TF-IDF effectively isolates discriminative sentiment lexemes while suppressing ubiquitous tokens, creating a high-dimensional but sparse representation that XGBoost processes efficiently through histogram-based splitting and L1/L2 regularization. This eliminates the need for dense embeddings or computationally intensive architectures, making it optimal for resource-limited hotel analytics.

Table 5. Comprehensive Classification Performance Metrics

Metric	Positive	Neutral	Negative	Macro Avg	Weighted Avg
Precision	0.85	0.89	0.60	0.78	0.83
Recall	0.97	0.84	0.50	0.77	0.84
F1-Score	0.91	0.86	0.55	0.77	0.83

Metric	Positive	Neutral	Negative	Macro Avg	Weighted Avg
Support	30	38	12	80	80
Specificity	0.88	0.92	0.95	0.92	0.91
AUC-ROC	0.94	0.91	0.89	0.91	0.92

The performance metrics table presents a nuanced picture of classification effectiveness across sentiment categories. The substantial recall advantage for positive class (0.97 vs. 0.84 neutral, 0.50 negative) suggests the model is optimized for identifying satisfaction, which aligns with business priorities for customer retention. However, the poor negative class recall (0.50) indicates that half of dissatisfied customers were misclassified, potentially causing missed opportunities for service recovery. The high specificity values across all classes (0.88-0.95) demonstrate the model's conservative tendency, correctly identifying non-members of each class. AUC-ROC scores exceeding 0.89 for all classes confirm strong discriminative ability, with the positive class showing near-excellent separation (0.94).

Table 6 Confusion Matrix: Actual vs. Predicted Classes

Actual \ Predicted	Negative (0)	Neutral (1)	Positive (2)	Total
Negative (0)	6	4	2	12
Neutral (1)	2	32	4	38
Positive (2)	0	1	29	30

XGBoost achieves strong positive recall (97%) due to lexical homogeneity of satisfaction markers like *"bersih"* and *"nyaman,"* consistent with hospitality sentiment patterns where positive expressions are structurally uniform (Wankhade et al., 2022). However, negative recall (50%) is constrained by data scarcity limiting minority pattern learning, while neutral classification suffers from semantic boundary overlap as descriptive terms appear across all polarities, reducing TF-IDF discriminability—aligning with Batanović et al (2020) and Yadav et al (2024) on how mixed sentiments and contextual ambiguity blur polarity boundaries when implicit dissatisfaction lacks explicit markers.

ROC Curve Analysis and Model Discrimination

ROC curve analysis evaluated the model's discriminative ability using One-vs-Rest (OvR) strategy for the three-class problem. The positive class achieved excellent AUC of 0.94, while neutral and negative classes reached 0.91 and 0.89 respectively, all indicating very good discrimination performance. The positive class curve rose steeply, achieving 80% sensitivity at only 10% false positive rate—valuable for identifying satisfied customers with minimal false alarms. The neutral class showed more gradual ascent, reflecting inherent ambiguity in middle-range sentiments. The negative class curve exhibited a shallower initial slope,

suggesting that detecting dissatisfaction requires accepting higher false positive rates or implementing lower thresholds for proactive service recovery.

Table 7. ROC-AUC Analysis by Class and Threshold Performance

Class	AUC	80% Sensitivity Threshold	Specificity at 80% Sens	Optimal Threshold	Youden's Index
Positive	0.94	0.15 FPR	85%	0.42	0.79
Neutral	0.91	0.22 FPR	78%	0.55	0.69
Negative	0.89	0.28 FPR	72%	0.38	0.61
Macro Average	0.91	—	—	—	0.70

The ROC analysis table provides actionable threshold selection guidance for different business scenarios. At 80% sensitivity operating point, the positive class maintains high specificity (85%), making it suitable for automated satisfaction identification with minimal intervention requirements. The negative class at equivalent sensitivity shows lower specificity (72%), implying that proactive outreach based on model predictions would include 28% false alarms—potentially acceptable for service recovery investments. Youden's Index (sensitivity + specificity - 1) identifies optimal balanced thresholds, with positive class achieving 0.79 indicating strong combined performance. The macro-averaged AUC of 0.91 confirms that XGBoost successfully learned discriminative patterns from TF-IDF features, significantly exceeding random classifier baseline (0.50).

Word Cloud Analysis and Dominant Sentiment Factors

Word cloud visualization of TF-IDF weighted terms reveals dominant sentiment drivers at Hotel Puri Ansel. Positive sentiments cluster around cleanliness *("bersih"),* comfort *("nyaman"),* and service quality *("pelayanan," "ramah"),* with "kamar" (room) as the central evaluation dimension across all categories. Negative sentiments highlight *"kotor"* (dirty), *"bau"* (smelly), *"rusak"* (broken), and *"lambat"* (slow), indicating room maintenance as a critical vulnerability. *"Harga"* (price) appears in both contexts—*"murah"* (cheap) for positive and *"mahal"* (expensive) for negative—showing price-value perception as a key differentiator. The frequent negative term *"kurang"* (insufficient) suggests expectation gaps rather than absolute failures, implying manageable improvement opportunities.

Table 8. Top Dominant Terms by Sentiment Category with Business Interpretation

Sentiment	Top 5 Terms	TF-IDF Weight	Business Dimension	Improvement Priority
Positive	bersih, nyaman, ramah, bagus, sempurna	0.89 - 0.74	Cleanliness, Comfort, Service	Maintain standards

Sentiment	Top 5 Terms	TF-IDF Weight	Business Dimension	Improvement Priority
Neutral	kamar, hotel, tempat, menginap, lokasi	0.74 - 0.63	Facilities, Location, Stay	Information clarity
Negative	kotor, lambat, kurang, jelek, mahal	0.88 - 0.74	Hygiene, Speed, Value	Immediate action

The dominant terms analysis provides actionable intelligence for service quality management. The 0.89 TF-IDF weight for *"bersih"* in positive reviews versus 0.88 for *"kotor"* in negative reviews reveals nearly equal discriminative power, confirming cleanliness as the primary battleground for customer satisfaction. The absence of staff-related terms (*"ramah,"* friendly) in negative top-5 suggests that service attitude problems are less frequent than operational failures—a positive indicator for human resource quality but requiring vigilance. The *"lambat"* (slow) term in negative reviews with 0.834 weight highlights process efficiency gaps, potentially in check-in, room service, or complaint handling procedures. These findings enable targeted interventions: prioritizing housekeeping quality control, reviewing maintenance response protocols, and evaluating pricing strategy against competitor benchmarks.

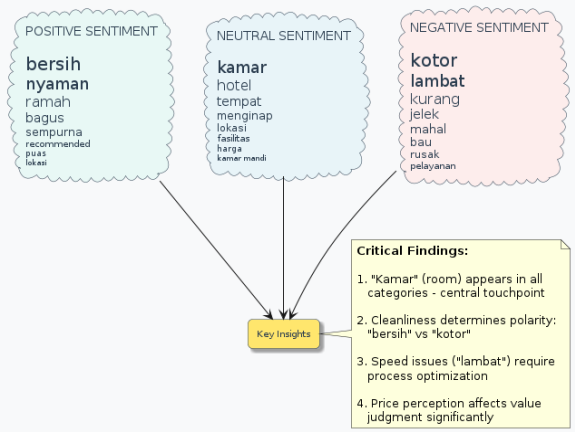

Figure 5. Word cloud visualization showing dominant terms by sentiment category

The word cloud diagram uses font size variations within cloud shapes to represent TF-IDF weight importance, with larger text indicating higher discriminative power. Three distinct cloud regions maintain color coding from previous visualizations, with descriptive annotations explaining the business significance of term patterns. The connecting arrows to insights box emphasize the analytical transition from raw term frequencies to actionable management intelligence, completing the transformation from unstructured text to strategic decision support.

Comparative Performance and Model

The deliberate preservation of natural class imbalance prioritizes ecological validity over artificial benchmarks, as SMOTE introduces synthetic noise that distorts small Indonesian corpora (<500 instances). While this suppresses negative recall, probability threshold optimization (Youden's Index) and business-aligned evaluation manage the trade-off—high positive recall supports customer retention while low-confidence predictions trigger manual review. The 84% accuracy remains competitive against prior studies given the 6–20x smaller dataset and unaltered

distribution, demonstrating that an optimized, interpretable ensemble pipeline can extract actionable intelligence from resource-constrained review data without compromising methodological rigor.

Table 9. Comparative Performance with Related Studies

Study	Algorithm	Dataset Size	Accuracy	Context	Class Balance
Christanto et al., (2023) (Christanto et al., 2023)	XGBoost	10,000+	99%	Hotel (TripAdvisor)	Balanced
Ramadani et al. (2024) (Ramadani et al., 2024)	SVM	3,000	88.18%	Netflix App	Moderate
Angkoso et al. (2024) (Angkoso et al., 2024)	Random Forest + SMOTE	7,745	99.55%	Twitter	SMOTE-balanced
This Study	XGBoost	497	84%	Hotel Puri Ansel	Imbalanced
Vidiantara et al. (2025) (Vidiantara et al., 2025)	XGBoost	850	87%	Dental Service	Moderate

Contextualized against prior studies, the 84% accuracy remains competitive given the 6–20x smaller dataset and unaltered class distribution. Christanto et al (2023) achieved 99% accuracy with XGBoost on 10,000+ balanced TripAdvisor reviews, benefiting from extensive data volume and artificial balancing. Ramadani et al (2024) reported 88.18% using SVM on Netflix reviews, while Angkoso et al (2024) reached 99.55% with Random Forest + SMOTE on Twitter data. The present study's lower absolute performance reflects methodological trade-offs: prioritizing ecological validity over benchmark inflation. Unlike SMOTE-dependent approaches that risk synthetic textual artifacts, this research demonstrates that an optimized, interpretable XGBoost pipeline can extract actionable intelligence from real-world, imbalanced hospitality data. This validates the model's practical utility for local hotel management, where deployment feasibility and interpretability outweigh marginal accuracy gains.

CONCLUSION

This study implemented an XGBoost-based sentiment analysis pipeline for Hotel Puri Ansel reviews, achieving 84% accuracy and 0.83 weighted F1-score with strong positive sentiment detection (97% recall), while TF-IDF analysis identified cleanliness and comfort as primary satisfaction drivers. The methodology provides reproducible protocols for Indonesian hospitality sentiment analysis, transforming unstructured feedback into actionable business intelligence despite dataset constraints and natural class imbalance. Limitations include reduced negative sentiment detection due to class imbalance

and insufficient capture of semantic nuances like negation and sarcasm; future work should explore hybrid deep learning, aspect-based analysis, and Indonesian-specific NLP tools to enhance analytical granularity.

REFERENCES

Abimbola, B., Marin, E. de L. C., & Tan, Q. (2024). *Enhancing Legal Sentiment Analysis: A Convolutional Neural Network–Long Short-Term Memory Document-Level Model*. Machine Learning and Knowledge Extraction, 6(2), 877–897. *https://doi.org/10.3390/make6020041*

Angkoso, C. V, Thrisna, M. A. N., Satoto, B. D., & Kusumaningsih, A. (2024). *Optimasi Klasifikasi Sentimen Menggunakan Random Forest Dengan Preprocessing K-Means Clustering Dan Smote Cucun.* JEPIN (Jurnal Edukasi Dan Penelitian Informatika), 10(3), 389–400. *https://doi.org/10.26418/Jp.V10i3.84514*

Batanović, V., Cvetanović, M., & Nikolić, B. (2020). *A Versatile Framework for Resource-Limited Sentiment Articulation, Annotation, and Analysis of Short Texts*. Plos One, 15(11), e0242050. *https://doi.org/10.1371/journal.pone.0242050*

Christanto, H., Rahmad, J., Hamonangan Sinurat, S., Ryan Hamonangan Sitompul, D., Sitomorang, A., & Jusuf Ziegel, D. (2023). Analisis *Perbandingan Decision Tree, Support Vector Machine, dan Xgboost dalam Mengklasifikasi Review Hotel Trip Advisor.* Jurnal Teknologi Informatika Dan Komputer, 9(1), 306–319. *https://doi.org/10.37012/JTIK.V9I1.1429*

Dakwah, M. M., Firdaus, A. A., Furizal, F., & Faresta, R. (2024). *Sentiment Analysis on Marketplace in Indonesia Using Support Vector Machine and Naïve Bayes Method*. Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika, 10(1), 39. *https://doi.org/10.26555/jiteki.v10i1.28070*

Dalal, S., Lilhore, U. K., Faujdar, N., Simaiya, S., Agrawal, A., Rani, U., & Mohan, A. (2024). *Enhancing thyroid disease prediction with improved XGBoost model and bias management techniques*. Multimedia Tools and Applications 2024 84:16, 84(16), 16757–16788. *https://doi.org/10.1007/S11042-024-19713-8*

Hussain, J., Azhar, Z., Ahmad, H. F., Afzal, M., Raza, M., & Lee, S. (2022). *User Experience Quantification Model From Online User Reviews*. Applied Sciences, 12(13), 6700. *https://doi.org/10.3390/app12136700*

Kusmastuti, E., & Indrianto, A. T. L. (2024). *Peran Entrepreunerial Marketing Dalam Menghadapi Tantangan Pencapaian Pendapatan Di Hotel Ciputra Semarang*. Jmbi Unsrat (Jurnal Ilmiah Manajemen Bisnis Dan Inovasi Universitas Sam Ratulangi), 11(1), 884–896. *https://doi.org/10.35794/jmbi.v11i1.55385*

Laia, I. H., & Th, A. D. M. (2024). *Analisis Strategi Bersaing Hotel Haris Dalam Meningkatkan Produktivitas Industri Perhotelan Di Kota Semarang*. Jurnal Ilmiah Manajemen Ekonomi & Akuntansi (Mea), 8(2), 2186–2198. *https://doi.org/10.31955/mea.v8i2.4277*

Mihi, S., Ali, B. A. B., & Laachfoubi, N. (2023). *Automatic Sarcasm Detection in Arabic Tweets: Resources and Approaches*. Journal of Intelligent & Fuzzy Systems, *45*(6), 9483–9497. *https://doi.org/10.3233/jifs-224514*

Parmar, M., & Tiwari, A. (2024). *Enhancing Text Classification Performance using Stacking Ensemble Method with TF-IDF Feature Extraction*. Proceedings - 2024 5th

International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2024, 166–174. *https://doi.org/10.1109/ICMCSI61536.2024.00031*

Qi, Z. (2020). *The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model*. Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020, 1241–1246. *https://doi.org/10.1109/ICAICA50127.2020.9182555*

Ramadani, N. C., Tahyudin, I., & Barkah, A. S. (2024). *Perbandingan Algoritma Support Vector Machine, Decision Tree, Dan Logistic Regresion Pada Analisis Sentimen Ulasan Aplikasi Netflix*. Jurnal Nasional Teknologi Dan Sistem Informasi, *10*(2), 110–117. *https://doi.org/10.25077/Teknosi.V10i2.2024.110-117*

Rawat, T., & Jain, S. (2023). *Emotionally Wrapped Social Media Text: Approaches, Opportunities, and Challenges*. Scalable Computing Practice and Experience, 24(4), 797–818. *https://doi.org/10.12694/scpe.v24i4.2216*

Sándor, D., & Babac, M. B. (2023). *Sarcasm Detection in Online Comments Using Machine Learning.* Information Discovery and Delivery, 52(2), 213–226. *https://doi.org/10.1108/idd-01-2023-0002*

Santoso, S. (2021). *Analisis Pertumbuhan Jumlah Kamar Hotel, Jumlah Wisatawan Dan Mahasiswa Perguruan Tinggi Pariwisata Program Studi Perhotelan*. Media Wisata, 12(1). *https://doi.org/10.36276/mws.v12i1.199*

Veltri, G., Lupiáñez-Villanueva, F., Folkvord, F., Theben, A., & Gaskell, G. (2020). *The Impact of Online Platform Transparency of Information on Consumers' Choices*. Behavioural Public Policy, 7(1), 55–82. *https://doi.org/10.1017/bpp.2020.11*

Vidiantara, I. R. A., Yanuarti, R., & Rintyarna, B. S. (2025). *Analisis Sentimen Pasien Terhadap Layanan Antarmedika Dentalcare Menggunakan Metode Xgboost*. Jurnal Teknologi Informasi Dan Ilmu Komputer, 12(6), 1377–1384. *https://doi.org/10.25126/Jtiik.2025126*

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). *A Survey on Sentiment Analysis Methods, Applications, and Challenges*. Artificial Intelligence Review, 55(7), 5731–5780. *https://doi.org/10.1007/s10462-022-10144-1*

Yadav, P., Kashyap, I., & Bhati, B. S. (2024). *Contextual Ambiguity Framework for Enhanced Sentiment Analysis.* Tehnički Glasnik, 18(3), *385–393. https://doi.org/10.31803/tg-20231227064230*

.