Explainable Multilingual Aspect-Based Sentiment Analysis for Tourism Using SHAP and LIME

Basworo Ardi Pramono
Doctor of Information Systems, Diponegoro University, Semarang, Indonesia | Informatics Engineering Department, Faculty of Information and Communication Technology, Semarang University, Semarang, Indonesia
basworo@usm.ac.id (corresponding author)

Rahmat Gernowo
Doctor of Information Systems, Diponegoro University, Semarang, Indonesia
rahmatgernowo@lecturer.undip.ac.id

Aghus Sofwan
Department of Electrical Engineering, Diponegoro University, Semarang, Indonesia
asofwan@elektro.undip.ac.id

Received: 16 March 2026 | Revised: 23 April 2026 | Accepted: 11 May 2026

 DOI: https://doi.org/10.48084/etasr.18774

ABSTRACT

Online tourism analytics increasingly relies on Aspect-Based Sentiment Analysis (ABSA) to extract fine-grained visitor perceptions; however, prior tourism ABSA studies often emphasize predictive performance while providing limited and rarely quantified evidence on explanation reliability. To address this gap, this study proposes an explainable multilingual ABSA framework for tourism reviews that combines one-vs-rest Logistic Regression (LR) with linear coefficients, SHAP, and LIME, and augments them with a quantitative trustworthiness evaluation. Experiments were conducted on a bilingual corpus of 2,891 Indonesian and English Google Reviews collected from 10 tourist destinations in Central Java and annotated into 9 multi-label classes derived from the dimensions of attractions, amenities, and accessibility, and their sentiment polarities. The selected model achieved a Macro-F1 of 0.4586, a Hamming loss of 0.1505, and an exact match of 0.2512. The global explanation analysis showed substantial agreement between the LR coefficients and SHAP rankings, with overlap@10 generally ranging from 0.70 to 0.80 across most labels. Eraser-based evaluation at $K = 10$ preserved predictions for 0.72–1.00 of cases, indicating strong fidelity of influential features. At the local level, SHAP and LIME consistently highlighted salient tokens associated with both correct and incorrect predictions, while sanity checks showed sharp degradation under model randomization, confirming that the explanations were tied to learned model parameters rather than superficial artifacts. These findings demonstrate that multilingual tourism ABSA can be made both interpretable and quantitatively auditable, thereby providing a transparent analytical basis for tourism service evaluation, destination management, and future decision-support applications.

Keywords-aspect-based sentiment analysis; explainable artificial intelligence; tourism analytics; SHAP; LIME

I. INTRODUCTION

Online reviews have become a critical source of user-generated content for tourism analytics because they capture visitors' detailed experiences with destinations, facilities, and travel access. In this context, Aspect-Based Sentiment Analysis (ABSA) is more informative than document-level sentiment classification because it links opinions to specific service dimensions rather than reducing an entire review to a single polarity label [1-4]. This capability is especially important in tourism, where destination managers require fine-grained evidence about what visitors appreciate or criticize about attractions, amenities, and accessibility.

However, tourism reviews are inherently multilingual and noisy. They often contain informal expressions, lexical variation, code-switching, and cross-lingual cues that complicate reliable aspect and polarity detection [2, 4-6]. Prior studies have shown that ABSA can support tourism

monitoring, service-quality analysis, and recommendation-oriented analytics by identifying visitor perceptions toward destination attributes and facilities [1, 7, 8]. Recent multilingual approaches have also explored hybrid modeling, transformer-based adaptation, and language-specific tuning to address linguistic variability [5, 8, 9]. Even so, most studies still prioritize predictive performance over the interpretability of model decisions.

Recent developments in AI-based review analytics have also introduced LLM-oriented, multimodal, and deployment-aware directions [10, 11]. Transformer and LLM-based approaches, including retrieval-augmented and system-level NLP architectures, offer promising capabilities for richer customer insight extraction, while multimodal sentiment analysis may combine textual reviews with ratings, images, metadata, or social media signals [10, 11]. However, these directions also raise unresolved issues related to cross-domain and cross-country generalization, computational cost, latency, scalability, privacy, fairness, and deployment governance [11-13]. Therefore, while LLM-based, multimodal, and production-level tourism analytics represent important future directions, this study deliberately focuses on text-based explainable multilingual ABSA and quantitative trustworthiness evaluation as an auditable foundation for tourism review analysis.

This limitation is increasingly problematic because tourism analytics is often used to support managerial and policy decisions. High predictive accuracy alone is insufficient if destination managers cannot understand why a model assigns a particular aspect-polarity label. In sentiment analysis and ABSA, Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME are increasingly used to expose influential features and make the model output more transparent [9]. More broadly, recent XAI research emphasizes that local explainers, such as LIME, remain valuable because they provide human-readable instance-level rationales, especially in text classification, but their usefulness depends on a careful evaluation of fidelity, robustness, and consistency rather than visual plausibility alone [14-16]. This concern is also reinforced by recent evidence showing that SHAP can simultaneously deliver meaningful local and global insight, underlining the value of multi-scale explanation for trustworthy model interpretation [17]. However, in tourism-oriented ABSA, explainability is still often limited to illustrative examples rather than systematically evaluated by measurable evidence of explanation quality and reliability [2, 7, 9]. In addition, prior tourism ABSA studies rarely combine multilingual modeling, explicit global-local explanation, and quantitative trustworthiness validation within a single auditable framework.

A second gap concerns the explanation of trustworthiness. Existing tourism ABSA studies rarely provide a unified protocol that jointly examines global drivers, local rationales, and quantitative validation of explanatory behavior. In particular, agreement across explanation methods, faithfulness under feature removal, robustness under perturbation, and sanity checks under randomization are not yet standard practice in bilingual tourism ABSA [18, 19]. This gap aligns with the XAI literature, which has repeatedly shown that explanation methods may vary in stability, robustness, and descriptive quality across datasets, model families, and application domains, making benchmarking and structured evaluation essential [14, 15]. As a result, many explanation outputs remain intuitively appealing but empirically under-validated, especially in multilingual and applied NLP settings.

The core problem addressed in this study is not only how to classify multilingual tourism reviews at the aspect level, but also how to make such predictions transparent, auditable, and quantitatively trustworthy for practical tourism analytics. To address these gaps, this study proposes an explainable multilingual ABSA framework for tourism reviews using One-vs-Rest Logistic Regression (LR-OvR) as the model of record, combined with linear coefficients, SHAP and LIME, for global and local interpretation. The study uses a bilingual Indonesian-English corpus of reviews from 10 tourist destinations in Central Java, annotated into nine multi-label classes based on three aspect dimensions and sentiment polarities. The main contributions are twofold. First, the study develops a compact, explainable ABSA pipeline for multilingual tourism reviews. Second, it quantitatively audits explanation trustworthiness through agreement analysis, eraser-based faithfulness evaluation, stability testing, and sanity checks, drawing on established XAI evaluation perspectives [14, 15, 20, 21]. In this way, this study positions multilingual tourism ABSA not merely as a sentiment classification task, but as an auditable, explainable analytics framework. Downstream destination ranking and DSS-oriented integration are reserved for future work.

The proposed method is designed to address three linked gaps: multilingual tourism ABSA under noisy bilingual conditions, the lack of explicit global-local explanation, and the limited use of quantitative trustworthiness evidence in prior tourism-oriented XAI studies. The unique contribution of this study, therefore, lies in integrating a benchmarked and interpretable model of record with complementary SHAP/LIME explanations and a quantitative trustworthiness protocol for multilingual tourism ABSA, rather than presenting explanation as a purely illustrative add-on.

II. METHODS

A. Dataset and Annotation

This study used a multilingual tourism review corpus comprising 2,891 Google Reviews from 10 tourist destinations in Central Java, Indonesia. The corpus contains reviews written in Indonesian and English and reflects the typical characteristics of tourism user-generated content, including short texts, informal expressions, lexical variation, and occasional code-switching. After deduplication and initial cleaning, each review was treated as a unique document-level instance for multi-label classification under the three-aspect (3A×) tourism framework. The corpus consists of bilingual Indonesian-English reviews and is annotated using a 3A× polarity multi-label scheme, where each review may receive one or more labels corresponding to Attractions, Amenities, and Accessibility under positive, neutral, or negative polarity. The reviews were independently annotated by two human annotators using a standardized 3A× guideline. The global

label distribution confirms that tourism reviews are dominated by Attractions-Positive, followed by Attractions-Neutral, Amenities-Positive, Amenities-Negative, and Amenities-Neutral, while the three Accessibility labels remain relatively sparse. Importantly, Amenities-Neutral is present in the corpus and should not be treated as zero, with counts of 213 for annotator A and 204 for annotator B. This point is methodologically important because it confirms that neutral amenity cues remain a valid and learnable class in the corpus. The dataset used in this study is publicly archived in Zenodo and includes bilingual dual-annotated travel reviews from 10 tourist destinations in Central Java [22].

Text preprocessing was intentionally kept lightweight to preserve lexical cues relevant for explainability. The preprocessing stage included lowercasing, whitespace normalization, removal of non-informative punctuation, and TF-IDF-compatible text cleaning while preserving lexical variants that remained informative for multilingual aspect interpretation. No aggressive stemming or semantic rewriting was applied, because the explainability stage required the lexical surface form to remain traceable for coefficient-, SHAP-, and LIME-based analysis.

Table I summarizes the main characteristics of the multilingual 3A tourism review corpus used in this study. The dataset consists of 2,891 Google Reviews collected from 10 tourist destinations in Central Java and formulated as a review-level multi-label classification task over nine labels derived from the 3A framework and sentiment polarity. The corpus was partitioned into 2,023 training, 434 validation, and 434 test instances to maintain comparability with the benchmark configuration established in the benchmark study. The table also reports the overall annotation consistency, with Exact Match = 0.8137 and mean Jaccard = 0.8438, indicating strong agreement in the consolidated nine-label annotation space. Collectively, these statistics show that the corpus is sufficiently structured and reliable for both benchmark modelling and downstream explainability analysis.

TABLE I. CORPUS SUMMARY

Item	Value
Source	Google Reviews
Geographic scope	10 tourist destinations in Central Java
Total reviews	2,891
Task type	Review level multi-label classification
Label space	3A × polarity = 9 labels
Train	2,023
Validation	434
Test	434
Annotators	2
Exact-match (9-label)	0.8137
Jaccard mean	0.8438

Table II reports the global label distribution produced by the two independent annotators. For each label, the table presents two pieces of information: the absolute frequency (count) and the relative proportion (prop.) with respect to the total number of labels assigned by each annotator. Because the task is multi-label, a single review may contribute more than one label; therefore, the reported proportions should be interpreted as shares of the annotator's total assigned labels rather than shares of the total number of reviews. The table shows that Attractions-Positive is the dominant category for both annotators, followed by Attractions-Neutral and Amenities-Positive, whereas most Accessibility-related and negative labels remain relatively sparse. This pattern confirms the presence of class imbalance in the corpus and justifies the use of Macro-F1, per-label threshold tuning, and robust comparative evaluation across models.

TABLE II. GLOBAL LABEL DISTRIBUTION PER ANNOTATOR

Label	Annotator A (count, prop.)	Annotator B (count, prop.)
Attractions-Positive	1762 (0.587)	1617 (0.539)
Attractions-Neutral	671 (0.224)	711 (0.237)
Attractions-Negative	46 (0.015)	40 (0.013)
Amenities-Positive	660 (0.220)	712 (0.237)
Amenities-Neutral	213 (0.071)	204 (0.068)
Amenities-Negative	267 (0.089)	269 (0.090)
Accessibility-Positive	124 (0.041)	124 (0.041)
Accessibility-Neutral	53 (0.018)	46 (0.015)
Accessibility-Negative	57 (0.019)	51 (0.017)

Class imbalance was addressed at the evaluation and decision-threshold level rather than through synthetic resampling. Since the label distribution is highly skewed, especially for accessibility-related and negative labels, Macro-F1 was used as the primary evaluation metric, complemented by Hamming-loss and Exact-match. In addition, per-label threshold tuning was applied on the validation set to reduce the bias of a fixed global threshold toward majority labels. Synthetic oversampling was not applied because the objective was to explain the model behavior on the observed tourism review distribution, and altering the training distribution could change the feature patterns being audited by SHAP and LIME.

TABLE III. INTER-ANNOTATOR AGREEMENT

Label	Cohen's κ
Attractions-Positive	0.7375
Attractions-Neutral	0.6709
Attractions-Negative	0.8121
Amenities-Positive	0.8628
Amenities-Neutral	0.8342
Amenities-Negative	0.8183
Accessibility-Positive	0.9087
Accessibility-Neutral	0.8672
Accessibility-Negative	0.5671

Table III presents the inter-annotator agreement for each of the nine labels using Cohen's κ. Overall, the agreement values range from 0.5671 to 0.9087, indicating substantial to almost perfect agreement across most labels. The strongest agreement is observed for Accessibility-Positive (κ = 0.9087), while Amenities-Positive (κ = 0.8628), Amenities-Neutral (κ = 0.8342), and Amenities-Negative (κ = 0.8183) also show consistently high reliability. The relatively lower agreement for Accessibility-Negative (κ = 0.5671) suggests that this label is more difficult to annotate, likely due to sparse support and the contextual ambiguity of access-related complaints in short tourism reviews. Overall, the table confirms that the annotation protocol was reliable enough to support downstream modeling,

while also indicating which labels are inherently more challenging and may contribute to lower predictive performance in later evaluation stages.

Figure 1 presents the end-to-end workflow of the proposed explainable multilingual ABSA framework. Starting from bilingual tourism reviews, the pipeline proceeds through cleaning, deduplication, dual annotation under the 3A× polarity scheme, and train-validation-test splitting. A benchmark stage was then employed to compare multiple model families, after which LR-OvR was selected due to its competitive performance and transparent feature-level interpretability. The selected model was subsequently analyzed through global explanation methods (LR coefficients and SHAP) and local explanation methods (SHAP and LIME). Finally, the explanation outputs were validated through agreement, eraser-based faithfulness, stability, and sanity tests, while future work may extend these outputs toward downstream destination evaluation and decision-support applications.

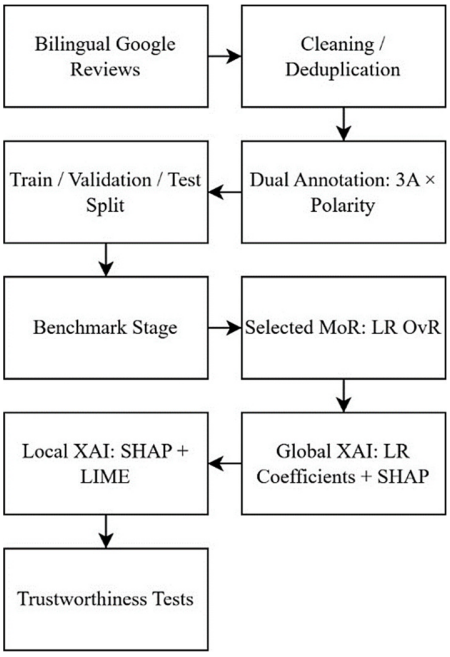

Fig. 1. Proposed explainable multilingual ABSA framework.

B. Benchmark Summary and Selection of the Model of Record

Before the explainability stage, the same corpus was benchmarked using classical machine-learning models, multilingual transformers, and calibrated soft-voting ensembles. The benchmark results show that the weighted soft-voting ensemble achieved the highest Macro-F1 (0.4596), while tuned LR OvR reached a nearly identical Macro-F1 (0.4586) with a Hamming-loss of 0.1505 and an Exact-match of 0.2512. Since the performance difference was marginal, but LR offered substantially clearer feature-level interpretability, this study selected LR-OvR as the Model of Record (MoR) for the explainability stage. The selection of LR-OvR reflects a deliberate trade-off between predictive performance and interpretability. Although the weighted soft-voting ensemble achieved a slightly higher Macro-F1, its gain over tuned LR-OvR was marginal, while LR-OvR provides direct coefficient-based feature transparency and a reproducible basis for auditable explanations. SHAP and LIME were selected as complementary explanation methods: SHAP supports additive global/local attribution and comparison with LR coefficients, whereas LIME provides sparse local rationales for individual review cases [11, 23, 24]. Alternative XAI methods, including integrated gradients, anchors, counterfactual explanations, concept-based explanations, and rationale-based approaches, are acknowledged as future comparative work [25-27].

The selected MoR uses a TF-IDF word+character n-gram representation with per-label threshold tuning, following the benchmark configuration established in the benchmark study. This configuration is appropriate for short and noisy bilingual tourism reviews because it preserves sparse lexical signals while remaining transparent enough for coefficient-based and perturbation-based explanation methods such as SHAP and LIME. The final configuration uses a TF-IDF union of word- and character-level n-grams within the benchmarked LR-OvR pipeline, with threshold calibration performed independently for each label on the validation split. This configuration was retained unchanged for the explainability stage to ensure that the reported SHAP, LIME, and trustworthiness analyses remain tied to the same empirically benchmarked classifier.

TABLE IV. BENCHMARK SUMMARY

Model	Macro-F1	Hamming-loss	Exact-match
LR OvR (tuned)	0.4586	0.1505	0.2512
mBERT (fine-tuned)	0.4124	0.1516	0.2857
XLM-R (fine-tuned)	0.3844	0.1851	0.2442
Ensemble Soft-Voting (weighted, tuned)	0.4596	0.1365	0.3065

Although the weighted soft-voting ensemble achieved the highest Macro-F1, its gain over tuned LR-OvR was marginal. In contrast, LR-OvR outperformed the single multilingual transformer baselines and provided substantially clearer feature-level interpretability. Therefore, LR-OvR was selected as the MoR for the explainability stage.

C. Explainability Protocol

The explainability protocol was designed to provide both global and local evidence. At the global level, the study uses LR coefficients and SHAP LinearExplainer to identify the dominant lexical drivers for each 3A× polarity label. At the local level, SHAP and LIME are applied to representative true-positive, false-positive, and false-negative cases to analyze why the model succeeds or fails on specific review instances. This dual perspective allows the study to move beyond raw predictive performance and examine how multilingual tourism signals are operationalised in the model. The overall XAI stage is therefore anchored in the benchmarked LR-OvR pipeline.

LIME was included in the local explanation stage because it provides sparse, human-readable rationales around individual predictions, making it particularly suitable for inspecting short tourism reviews and identifying which lexical cues activate a specific aspect-polarity label. This role aligns with the broader XAI literature, where LIME remains a widely used local

explainer due to its model-agnostic design and intuitive surrogate-based approximation of local decision boundaries [14, 20]. However, prior studies also emphasise that LIME explanations can be sensitive to perturbation design and feature representation, meaning that their usefulness depends on careful validation rather than visual plausibility alone [14]. For this reason, this study does not rely on LIME as an isolated interpretability device. Instead, LIME is used alongside SHAP so that local explanations can be cross-read against a second explanation mechanism and later assessed within a broader trustworthiness framework, including agreement, faithfulness, stability, and sanity-oriented evaluation [17].

D. Trustworthiness Evaluation

To avoid treating explainability as a purely illustrative exercise, this study evaluates explanation behavior using four complementary components: agreement, eraser-based faithfulness, stability, and sanity checks. Agreement is measured by comparing ranked features from the LR coefficients and SHAP. Faithfulness is evaluated using top-K eraser tests. Stability is assessed under mild preprocessing perturbations, and sanity checks are conducted under randomization settings to verify that explanation rankings degrade when the learned structure is destroyed. This evaluation layer is built on top of the benchmarked LR-OvR model and uses the same data split as the benchmark study, ensuring that the explainability findings remain tied to the same empirical foundation.

The robustness analysis in this study focuses on mild preprocessing perturbations and sanity checks under randomization. These tests evaluate whether explanation rankings remain stable under realistic text variation and whether they collapse when the learned model structure is destroyed. However, this protocol should not be interpreted as a full adversarial robustness evaluation. Adversarial settings such as intentional spelling attacks, semantic-preserving perturbations, adversarial code-switching, and manipulated review expressions remain beyond the scope of this study and should be examined in future work.

Taken together, the methodological contribution of this study lies in combining a benchmarked and interpretable MoR with complementary global-local explanation and a quantitative trustworthiness protocol for multilingual tourism ABSA.

III. RESULTS AND DISCUSSION

A. MoR Performance and Analytical Rationale

The explainability analysis is anchored to the tuned LR-OvR model. On the test set, the model achieved a Macro-F1 of 0.4586, a Hamming-loss of 0.1505, and an exact-match of 0.2512, confirming that a sparse linear baseline remains competitive for multilingual multi-label tourism ABSA. Although the weighted soft-voting ensemble slightly outperformed LR-OvR in overall Macro-F1, the margin was small, whereas LR offered substantially clearer feature-level interpretability. This trade-off justifies the use of LR-OvR as the analytical basis for the XAI stage.

This result is important for two reasons. First, it shows that interpretable linear models remain highly relevant for noisy bilingual tourism reviews, where lexical sparsity, informal expressions, and mixed-language patterns often reduce the practical advantage of more complex architectures. Second, it establishes a stable methodological bridge between predictive benchmarking and explanation analysis: the XAI layer is applied to a benchmarked classifier whose performance is already empirically defensible. In this sense, this study does not position explainability as a cosmetic add-on, but as an auditable analytical extension of a competitive ABSA baseline. The benchmark comparison against multilingual transformers and weighted ensemble modeling shows that the proposed explainable pipeline is grounded in a competitive empirical baseline rather than an arbitrarily selected interpretable model.

B. Global Explanation Results: Dominant Drivers Across 3A Labels

Global explanation analysis reveals that the learned lexical drivers are strongly aligned with intuitive tourism semantics. The LR coefficients extracted from the TF-IDF word+character representation show that Attractions-Positive is dominated by evaluative terms associated with aesthetic appreciation and destination appeal, including great, cool, exciting, beautiful, lively, and their bilingual variants, such as nice and very. For Amenities-Positive, the dominant features are associated with comfort, cleanliness, and service quality, including clean, comfortable, spacious, friendly, with a pool, and affordable. In contrast, Accessibility-Negative is primarily driven by complaint-oriented features related to parking, entry, travel distance, and circulation, such as car park, enter, must, not, far, and road. These patterns indicate that the model captures operationally meaningful distinctions across the 3A× dimensions rather than relying on arbitrary lexical artifacts.

The SHAP global importance analysis supports the same interpretation. In the current draft, the representative SHAP plots confirm that positive attraction labels are associated with affective and evaluative words describing scenery, heritage, or enjoyment. In contrast, negative accessibility labels are driven by mobility-related bottlenecks such as parking access, distance, and movement constraints. This consistency between LR coefficients and SHAP rankings is analytically important because it suggests that the global explanation patterns are not method-specific. Instead, both explanation mechanisms converge on a similar view of the model's dominant lexical signals.

From a tourism-management perspective, these global drivers are already actionable. Positive attraction signals point to destination strengths that can be maintained or amplified in destination branding, while negative accessibility signals identify friction points that may require operational intervention, such as parking organisation, entry flow management, or pedestrian access improvement. This demonstrates that global explainability in multilingual ABSA is not merely descriptive, but can be directly translated into service-quality interpretation under the 3A framework. These findings also have direct practical implications for the tourism industry. Positive attraction and amenity drivers can support destination branding and service-quality strengthening, while

negative accessibility drivers can help managers identify operational bottlenecks requiring intervention. The study also extends recent XAI work by bringing local-global explanation into a multilingual tourism ABSA setting and adding quantitative trustworthiness evaluation [17].

TABLE V. REPRESENTATIVE GLOBAL LEXICAL DRIVERS IDENTIFIED BY LR COEFFICIENTS AND SHAP FOR SELECTED 3A LABELS

Label	Top LR features	Top SHAP features	Interpretation
Attractions-Positive	great, cool, exciting, beautiful, lively, nice, very	great, really, cool, fun, beautiful, lively, nice, very, place, family	Positive attraction sentiment is driven by aesthetic appreciation and enjoyable destination experiences.
Amenities-Positive	clean, comfortable, spacious, friendly, swimming pool, affordable	clean, comfortable, spacious, available, plenty of swimming pools, friendly, affordable, nice, price	Positive amenity sentiment is shaped by cleanliness, comfort, facility completeness, and service quality.
Accessibility -Negative	parking, enter, must, not, far, road	parking, enter, must, not, far, road, access, car park, busy, to	Negative accessibility sentiment is mainly associated with parking problems, difficult entry, and long travel distances.

C. Agreement and Trustworthiness of the Explanations

A key contribution of this study is that the explainability outputs are evaluated quantitatively rather than being presented only as visual illustrations. At the global level, agreement between LR coefficients and SHAP rankings was assessed using Overlap@K and Spearman correlation. The current draft shows that Overlap@10 generally falls in the 0.70–0.80 range for most labels, indicating substantial convergence between the two explanation methods. This means that the most influential lexical signals identified by the linear coefficients are also consistently recovered by SHAP, which strengthens the credibility of the extracted global drivers.

Faithfulness was further examined using eraser-based evaluation. For $K = 10$, the proportion of unchanged predictions after preserving only the most influential features ranged from 0.72 to 1.00 across labels, while the corresponding logit changes remained consistent with the importance of the removed or retained features. These findings indicate that the highlighted tokens are not merely correlated with the outputs but are meaningfully tied to the model's decision process. In other words, the explanation features show strong fidelity to the classifier's actual behavior.

Taken together, the agreement, eraser, stability, and sanity results support the argument that the proposed framework delivers auditable explanations rather than decorative visualisations. This combination of benchmarked modelling, global-local explanation, and quantitative trustworthiness evidence constitutes the main analytical contribution of the study beyond standard tourism ABSA reporting. This is a substantive distinction from much of the prior tourism ABSA literature, where explanation is often limited to a few example excerpts without systematic quantitative validation.

Table VI provides a compact summary of explanation trustworthiness. The results show substantial agreement between LR coefficients and SHAP, strong eraser-based faithfulness for most labels, robust stability under mild perturbations, and sharp degradation under randomization. Together, these findings indicate that the explanations are both interpretable and quantitatively reliable.

TABLE VI. SUMMARY OF AGREEMENT AND TRUSTWORTHINESS EVIDENCE

Evaluation component	Indicator	Value	Interpretation
Cross-method agreement	Overlap@10 (LR-coef vs SHAP)	0.60–0.80 (median ≈ 0.70)	Substantial alignment on core global drivers
Cross-method agreement	Spearman ρ	0.58–0.72	LR and SHAP rankings are directionally consistent
Eraser faithfulness	Fidelity decision at K=10	0.89–1.00 for most labels; 0.72 for Amenities-Positive	Top-K features well reproduce most labels, while Amenities-Positive remains more diffuse
Eraser faithfulness	Unchanged predictions at K=10	0.72–1.00	Highlighted features preserve decision-relevant evidence
Sufficiency	$\lvert\Delta logit\_keep\rvert$ at K=10	0.06–0.27	Keeping only Top-K features usually retains the decision with moderate confidence loss
Comprehensive-ness	$\Delta logit\_erase$ at K=10	0.07–0.48	Removing Top-K features consistently reduces model confidence
Stability under perturbation	Overlap@10	0.70–0.80	Global explanations remain robust under mild text perturbations
Stability under perturbation	Jaccard-top$K$	0.54–0.67	Top-K feature sets remain largely stable after perturbation
Sanity under randomization	Overlap@10	0.00–0.20	Explanations collapse when the learned structure is randomised
Sanity under randomization	Jaccard-top$K$	0.00–0.13	Supports dependence on learned model parameters rather than artifacts

D. Local Explanations and Error Patterns

At the local level, SHAP and LIME provide consistent token-level rationales for both correct and incorrect predictions. In representative true-positive cases, both methods highlight a compact set of semantically intuitive tokens, such as good, beautiful, nice, or friendly, that directly support positive aspect predictions. This confirms that the local explanations remain readable and semantically aligned with human interpretation, especially for dominant positive labels in the attractions and amenities dimensions.

This error structure is consistent with the data characteristics observed in the corpus benchmark. The label space is clearly imbalanced, and minority labels, especially within the accessibility dimension, have less support and greater contextual ambiguity. As a result, some lower-

frequency labels are inherently more difficult to predict and, correspondingly, more difficult to explain with the same level of confidence as dominant positive labels. The local case analysis, therefore, serves two functions simultaneously: it strengthens the interpretability claim, and it provides diagnostic evidence for future data curation, label refinement, or post-processing rules.

In general, the proposed framework shows that multilingual tourism ABSA can be made both interpretable and quantitatively auditable. The combined evidence from global agreement, local rationale analysis, eraser-based faithfulness, stability testing, and sanity checks indicates that the extracted explanations are meaningfully tied to the learned model behavior. This makes the outputs more transparent and trustworthy for tourism analytics beyond raw classification performance alone. Table VII shows that the local explanations are generally consistent with the predicted accessibility labels. In the true-positive cases, the highlighted tokens directly reflect either access difficulty (parking, difficult) or proximity (nearby, home). In the false-negative case, the relevant access-related cues are identified, but their contribution remains insufficient to activate the final label. In the false-positive case, access vocabulary triggers a negative prediction even though the review lacks a clear complaint. These cases indicate that local errors are mainly associated with threshold sensitivity and ambiguous wording related to access.

TABLE VII. REPRESENTATIVE LOCAL EXPLANATION CASES

Case	Review Excerpt	Pred./True	Key Tokens
TP	"The only problem is parking."	Acc-Neg / Acc-Neg	parking, difficult
TP	"Tourist attractions near home…"	Acc-Pos / Acc-Pos	nearby, home
FN	"it's a pain to get there from the parking area."	Acc-Neg / Acc-Neg	pain, parking
FP	"plenty of parking spaces…"	Acc-Neg / Neutral	car park, area

IV. CONCLUSIONS

This study presented an explainable multilingual aspect-based sentiment analysis framework for tourism reviews using LR OvR and combining linear coefficients, SHAP, and LIME for global and local explanation. Using a bilingual Indonesian-English corpus of tourism reviews under the 3A framework, this study showed that an interpretable linear model can remain competitive while also supporting transparent inspection of label-specific lexical drivers. The findings demonstrate that the proposed explainability layer is not limited to visual interpretation alone. Global analysis showed semantically coherent drivers across representative 3A labels, local explanation cases revealed interpretable token-level rationales for both correct and incorrect predictions, and the trustworthiness evaluation provided quantitative support through cross-method agreement, eraser-based faithfulness, robustness under perturbation, and sanity degradation under randomization. Together, these results indicate that the explanation outputs are meaningfully related to the learned decision behavior of the model.

The main contribution of this work lies in presenting multilingual tourism ABSA not only as a predictive classification task, but as an auditable, explainable analytics framework whose explanations can be examined systematically and quantitatively. This is particularly important in tourism settings, where aspect-level insights are more useful when the reasoning behind model outputs can be inspected and justified. This transparency can support practical decisions related to service improvement, destination positioning, and operational issue prioritization in tourism management.

This study has several limitations related to generalization and deployment. First, the corpus is limited to bilingual Google Reviews from 10 tourist destinations in Central Java; therefore, the findings should not yet be generalized to other countries, tourism cultures, or review platforms without external validation. Cross-dataset, cross-region, and cross-country evaluation is needed to assess whether the proposed framework remains reliable under different linguistic, cultural, and platform-specific conditions. Second, although the LR OvR model with TF-IDF is computationally lighter than transformer-based architectures and is potentially suitable for dashboard-style tourism analytics, this study did not evaluate real-time latency, throughput, production scalability, or stakeholder-facing deployment yet. These aspects should be examined in future work before the framework is implemented as an operational tourism intelligence system.

In addition, the explanation evaluation in this study is computational rather than user-centered. Although SHAP and LIME explanations are assessed through agreement, eraser-based faithfulness, stability, and sanity checks, their usefulness has not yet been formally evaluated with tourism managers or destination stakeholders. Future work should include stakeholder-based evaluation to examine whether explanations improve interpretability, trust, and decision-making in practical destination management contexts.

DECLARATIONS OF COMPETING INTERESTS

The authors declare no conflicting interests that could have influenced the results of this study.

ACKNOWLEDGEMENT

Not applicable to this study.

DATA AVAILABILITY

The dataset used in this study is available at [22].

REFERENCES

[1] M. Agua, N. Antonio, M. P. Carrasco, and C. Rassal, "Large Language Models Powered Aspect-Based Sentiment Analysis for Enhanced Customer Insights," *Tourism & Management Studies*, vol. 21, no. 1, pp. 1–19, Jan. 2025, https://doi.org/10.18089/tms.20250101.

[2] A. Jain, A. Bansal, and S. Tomar, "Aspect-Based Sentiment Analysis of Online Reviews for Business Intelligence:," *International Journal of Information Technologies and Systems Approach*, vol. 15, no. 3, pp. 1–21, Aug. 2022, https://doi.org/10.4018/IJITSA.307029.

[3] S. O. E. Putri, A. A. Arifiyanti, and A. R. E. Najaf, "Convolutional Neural Network Approach for Aspect-Based Sentiment Analysis of Tourism Reviews," *bit-Tech*, vol. 8, no. 1, pp. 448–458, Aug. 2025, https://doi.org/10.32877/bt.v8i1.2582.

[4] A. Murzakhmetov, M. Satymbekov, A. Bapanov, and N. Beisov, "Sentiment Analysis of Tourist Reviews About Kazakhstan Using a Hybrid Stacking Ensemble Approach," *Computation*, vol. 13, no. 10, Oct. 2025, Art. no. 240, https://doi.org/10.3390/computation13100240.

[5] A. Alsehaimi, A. Babour, and D. Alahmadi, "Toward Transparent Modeling: A Scoping Review of Explainability for Arabic Sentiment Analysis," *Applied Sciences*, vol. 15, no. 19, Oct. 2025, Art. no. 10659, https://doi.org/10.3390/app151910659.

[6] A. A. Maruf, F. Khanam, Md. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung, "Challenges and Opportunities of Text-Based Emotion Detection: A Survey," *IEEE Access*, vol. 12, pp. 18416–18450, 2024, https://doi.org/10.1109/ACCESS.2024.3356357.

[7] M. R. A. Yudianto, P. Sukmasetya, R. A. Hasani, and Maimunah, "Aspect-Based Sentiment Analysis of Borobudur Temple Reviews Use Support Vector Machine Algorithm," *E3S Web of Conferences*, vol. 500, 2024, Art. no. 01005, https://doi.org/10.1051/e3sconf/202450001005.

[8] G. I. Bhaskara, I. G. A. Sastrawan, and I. G. B. A. Yudiastina, "Sentiment and Sunsets: Analysing Online Reviews of Kuta Beach in Bali," *E-Journal of Tourism*, Mar. 2024, Art. no. 76, https://doi.org/10.24922/eot.v11i1.114486.

[9] S. Rajarajeswari, D. Ashwin, N. S. Kumar, R. Vishnal, and N. Vishwa, "Online Review Sentimental Analysis," *International Journal of Innovative Science and Research Technology*, pp. 2152–2155, May 2025, https://doi.org/10.38124/ijisrt/25apr1434.

[10] X. Chen, H. Xie, S. J. Qin, Y. Chai, X. Tao, and F. L. Wang, "Cognitive-Inspired Deep Learning Models for Aspect-Based Sentiment Analysis: A Retrospective Overview and Bibliometric Analysis," *Cognitive Computation*, vol. 16, no. 6, pp. 3518–3556, Nov. 2024, https://doi.org/10.1007/s12559-024-10331-y.

[11] Y. Zhao, J. Zhang, Y. Tong, Z. Li, X. Yu, and S. Tsai, "Design of an Enterprise Public Opinion Monitoring System Based on Natural Language Processing: Sentiment Analysis and Management of Public Opinion Data," *Journal of Global Information Management*, vol. 33, no. 1, pp. 1–35, June 2025, https://doi.org/10.4018/JGIM.381306.

[12] A. Orive, A. Agirre, H. L. Truong, I. Sarachaga, and M. Marcos, "Quality of Service Aware Orchestration for Cloud–Edge Continuum Applications," *Sensors*, vol. 22, no. 5, Feb. 2022, Art. no. 1755, https://doi.org/10.3390/s22051755.

[13] F. Tanveer *et al.*, "Balancing privacy and performance in healthcare: A federated learning framework for sensitive data," *Digital Health*, vol. 11, May 2025, Art. no. 20552076251381769, https://doi.org/10.1177/20552076251381769.

[14] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, Sept. 2023, https://doi.org/10.1007/s10618-023-00933-9.

[15] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022, https://doi.org/10.1109/ACCESS.2022.3204051.

[16] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," *Frontiers in Artificial Intelligence*, vol. 8, Sept. 2025, Art. no. 1627078, https://doi.org/10.3389/frai.2025.1627078.

[17] S. Khanapur, J. S. Nayak, B. S. Rajeshwari, M. Namratha, C. B. Bharadwaj, and R. Bhardwaj, "SHAP-Based Explainability for Local and Global Insights in Alzheimer's Detection," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, pp. 30940–30947, Feb. 2026, https://doi.org/10.48084/etasr.13932.

[18] A. Adak, B. Pradhan, and N. Shukla, "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review," *Foods*, vol. 11, no. 10, May 2022, Art. no. 1500, https://doi.org/10.3390/foods11101500.

[19] E. I. Setiawan, F. Ferry, J. Santoso, S. Sumpeno, K. Fujisawa, and M. H. Purnomo, "Bidirectional GRU for Targeted Aspect-Based Sentiment Analysis Based on Character-Enhanced Token-Embedding and Multi-Level Attention," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 392–407, Oct. 2020, https://doi.org/10.22266/ijies2020.1031.35.

[20] S. Mirzaei, H. Mao, R. R. O. Al-Nima, and W. L. Woo, "Explainable AI Evaluation: A Top-Down Approach for Selecting Optimal Explanations for Black Box Models," *Information*, vol. 15, no. 1, Dec. 2023, Art. no. 4, https://doi.org/10.3390/info15010004.

[21] E. Albini, A. Rago, P. Baroni, and F. Toni, "Achieving descriptive accuracy in explanations via argumentation: The case of probabilistic classifiers," *Frontiers in Artificial Intelligence*, vol. 6, Apr. 2023, Art. no. 1099407, https://doi.org/10.3389/frai.2023.1099407.

[22] B. A. Pramono, "Dual-Annotated Central Java Tourism Review Dataset: 10 Tourist Destinations with Annotator A and Annotator B." Zenodo, Mar. 10, 2026, https://doi.org/10.5281/ZENODO.18937916.

[23] M. Danilevsky, S. Dhanorkar, Y. Li, L. Popa, K. Qian, and A. Xu, "Explainability for Natural Language Processing," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug. 2021, pp. 4033–4034, https://doi.org/10.1145/3447548.3470808.

[24] F. Stoehr *et al.*, "Natural language processing for automatic evaluation of free-text answers — a feasibility study based on the European Diploma in Radiology examination," *Insights into Imaging*, vol. 14, no. 1, Sept. 2023, Art. no. 150, https://doi.org/10.1186/s13244-023-01507-5.

[25] E. Balkir, S. Kiritchenko, I. Nejadgholi, and K. Fraser, "Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 2022, pp. 80–92, https://doi.org/10.18653/v1/2022.trustnlp-1.8.

[26] J. Yuan, J. Vig, and N. Rajani, "iSEA: An Interactive Pipeline for Semantic Error Analysis of NLP Models," in *27th International Conference on Intelligent User Interfaces*, Mar. 2022, pp. 878–888, https://doi.org/10.1145/3490099.3511146.

[27] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh, "Rationalization for explainable NLP: a survey," *Frontiers in Artificial Intelligence*, vol. 6, Sept. 2023, Art. no. 1225093, https://doi.org/10.3389/frai.2023.1225093.