Ingénierie des Systèmes d'Information
Vol. 31, No. 1, January, 2026, pp. 123-135
Journal homepage: http://iieta.org/journals/isi

# Aspect-Based Sentiment Analysis of Multilingual Hotel Reviews in Jakarta Using Multilingual Transformer Models: A Comparative Study of Multilingual BERT and Cross-Lingual RoBERTa

Arghanta Wijna Suryabrata[1*], Harco Leslie Hendric Spits Warnars[1], Maybin K. Muyeba[2]

[1] Computer Science Department, Bina Nusantara University, Jakarta 11480, Indonesia
[2] School of Science, Engineering & Environment (SEE), University of Salford, Manchester M5 4WT, United Kingdom

Corresponding Author Email: arghanta.suryabrata@binus.ac.id



https://doi.org/10.18280/isi.310112

**Received:** 31 July 2025
**Revised:** 26 November 2025
**Accepted:** 18 January 2026
**Available online:** 31 January 2026

***Keywords:***
*aspect-based sentiment analysis, multilingual transformers, multilingual Bidirectional Encoder Representations from Transformers, Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach, sentence-pair classification, hotel review analytics, multilingual sentiment analysis*

**ABSTRACT**

Online reviews have become a critical source of information for evaluating service quality in the hospitality industry. However, extracting fine-grained insights from multilingual user-generated content remains challenging due to linguistic variability and the presence of code-mixed expressions. Aspect-based sentiment analysis (ABSA) provides an effective framework for identifying customer opinions toward specific service attributes. This study investigates the effectiveness of multilingual transformer models for ABSA on online reviews of five-star hotels in Jakarta. A large-scale dataset comprising more than 96,000 reviews collected from TripAdvisor and Google Reviews was analyzed. The proposed framework adopts a sentence-pair classification strategy that reformulates ABSA as a natural language inference task, enabling transformer models to capture aspect–sentiment relationships more effectively. Two multilingual pretrained language models—multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa)—were fine-tuned and systematically compared. Experimental results show that XLM-RoBERTa achieved the best performance with an accuracy of 97.20% and an F1-score of 0.9729, slightly outperforming mBERT while requiring higher computational resources. In contrast, mBERT demonstrated greater stability across validation folds. Aspect-level sentiment analysis further revealed that cleanliness, facilities, and service are the most positively perceived aspects of five-star hotels in Jakarta, while pricing remains the primary source of negative sentiment. These findings demonstrate the effectiveness of transformer-based ABSA for multilingual hospitality reviews and provide actionable insights for data-driven decision-making in the luxury hotel sector.

## 1. INTRODUCTION

The increasing volume and diversity of online reviews for five-star hotels in Jakarta present significant technical challenges for sentiment analysis, particularly in a multilingual context where reviews are often written in Indonesian and English. Addressing these challenges requires advanced natural language processing (NLP) models capable of extracting nuanced opinions from code-mixed, user-generated content. Recent developments in transformer-based language models, specifically multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa) have shown promise in handling such multilingual and cross-lingual sentiment analysis tasks [1, 2].

A key methodological innovation in aspect-based sentiment analysis (ABSA) is the use of sentence pairing, or sentence-pair classification. Instead of analyzing each review as a single sequence, this approach reformulates ABSA as a natural language inference (NLI) problem by pairing each review sentence with an auxiliary sentence that represents a candidate aspect-sentiment combination. The model is then fine-tuned to predict whether the aspect and sentiment are present in the review. This method leverages the pretraining objectives of transformer models, which are inherently designed to understand relationships between sentence pairs [3]. The sentence pairing approach has been shown to improve the detection of complex or ambiguous aspect expressions and to reduce errors caused by out-of-vocabulary (OOV) words, which are common in informal, user-generated reviews.

Technically, mBERT and XLM-RoBERTa differ in their tokenization and training strategies. mBERT uses WordPiece tokenization, which effectively breaks down rare or misspelled words into subword units, minimizing OOV issues in Indonesian-English reviews. XLM-RoBERTa, on the other hand, employs SentencePiece tokenization with a language-agnostic vocabulary, allowing for more flexible handling of code-switched and morphologically rich text [1]. Empirical studies indicate that XLM-RoBERTa generally outperforms mBERT in multilingual ABSA tasks, especially when dealing with code-mixed or low-resource languages, due to its larger

and more balanced training corpus [2]. However, mBERT can demonstrate greater stability and efficiency in scenarios with limited data or highly contextualized aspect vocabularies [3].

This research focuses on a technical comparison of mBERT and XLM-RoBERTa for ABSA of five-star hotel reviews in Jakarta, utilizing the sentence pairing strategy. By systematically evaluating both models on a bilingual dataset, this study aims to identify their respective strengths and limitations in handling sentence-pair classification, tokenization, and multilingual context adaptation. The findings are expected to inform best practices for deploying ABSA frameworks in real-world, code-mixed hospitality review scenarios.

The main contributions of this paper are as follows:

(1) We conduct a comprehensive technical comparison of the sentence pairing approach for ABSA, utilizing both XLM-RoBERTa and mBERT models, specifically within the context of bilingual, code-mixed hotel reviews from Jakarta.

(2) We empirically evaluate which model and sentence-pairing strategy combination—mBERT with sentence pairing or XLM-RoBERTa with sentence pairing—delivers superior performance in real-world multilingual review scenarios.

(3) Aspect-level results are interpreted to derive actionable insights into guest satisfaction patterns at five-star hotels in Jakarta.

A complete list of abbreviations and acronyms used throughout this paper is provided in the Nomenclature section.

## 2. RELEVANT WORK

The increasing reliance on online reviews in the hospitality sector has spurred significant research into sentiment analysis, particularly within the context of hotel guest experiences. This section reviews key studies and methodologies relevant to ABSA and the application of multilingual language models, with a focus on five-star hotels in Jakarta and the comparative evaluation of mBERT and XLM-RoBERTa using sentence-pairing and fine-tuning approaches.

### 2.1 Sentiment analysis in hospitality and tourism

Online reviews are a critical factor influencing consumer decision-making and hotel performance. Previous research has established the strong impact of positive online reviews on customer trust, booking intentions, and overall hotel reputation [4-6], with studies further demonstrating that review attributes such as overall rating and volume significantly shape consumers' consideration sets during the hotel selection process. In the context of Jakarta's luxury hotel market, the competitive landscape underscores the importance of leveraging sentiment analysis to extract actionable insights from guest feedback, enabling hotels to enhance service quality and maintain a competitive edge [7].

### 2.2 Aspect-based sentiment analysis and methodological advances

Traditional sentiment analysis methods often fail to capture the nuanced opinions expressed in guest reviews. ABSA addresses this limitation by identifying specific aspects of hotel experiences (such as cleanliness, service, or location) and the sentiments associated with each aspect [8]. The ABSA process typically involves aspect extraction, categorization, and sentiment classification, with machine learning and deep learning models-such as Support Vector Machines (SVM), Naive Bayes, and especially transformer-based models-demonstrating strong performance in this domain [9-11].

Recent studies highlight the effectiveness of deep learning models, particularly those based on the BERT architecture, for ABSA tasks. Research [12] demonstrated that multi-label classification using BERT-based models achieved high accuracy in identifying aspect categories and their associated sentiments in German hotel reviews. Similarly, research [13] applied ABSA with zero-shot learning using RoBERTa to extract actionable insights from online reviews in the Indonesian tourism sector, emphasizing the method's adaptability across languages and domains.

### 2.3 Multilingual models: Multilingual Bidirectional Encoder Representations from Transformers and Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach

Multilingual language models have become essential tools for sentiment analysis in diverse linguistic contexts. mBERT, trained on Wikipedia data from 104 languages, and XLM-RoBERTa, trained on a larger and more balanced CommonCrawl corpus covering 100 languages, represent state-of-the-art approaches for cross-lingual understanding [1]. Research [2] compared these models for Indonesian-English bilingual text classification, finding that XLM-RoBERTa consistently outperformed mBERT, particularly in scenarios with limited labeled Indonesian data. The integration of English data was shown to further enhance model performance, a finding directly relevant to the analysis of bilingual hotel reviews in Jakarta.

Research [14] also reported that XLM-RoBERTa surpassed monolingual Indonesian BERT models in text classification tasks, underscoring its suitability for multilingual sentiment analysis. The ability of XLM-RoBERTa to handle code-switching and capture nuanced expressions in both Indonesian and English is particularly advantageous in the context of Jakarta's hospitality industry, where guest reviews frequently mix languages [2].

### 2.4 Sentence-pairing and fine-tuning for aspect-based sentiment analysis

A notable methodological advancement in ABSA involves transforming the task into a sentence-pair classification problem. Research [3] adapted the Natural Language Inference-Binary (NLI-B) approach, pairing each review with auxiliary sentences representing aspect-sentiment combinations. Fine-tuning mBERT on this sentence-pair dataset yielded significant improvements in F1-score compared to traditional single-sentence classification methods. This approach is particularly effective in handling domain-specific vocabulary and out-of-vocabulary (OOV) words, which are common in user-generated hotel reviews.

The sentence-pairing method, combined with fine-tuning of multilingual models, enables more accurate detection of aspect-specific sentiments and supports robust analysis across different languages. This methodological framework forms the basis for the current study's comparative evaluation of mBERT and XLM-RoBERTa in the context of five-star hotel reviews in Jakarta.

## 3. METHODOLOGY

This study employs a quantitative approach using ABSA to extract and classify sentiment from online reviews of five-star hotels in Jakarta. The methodology adapts a sentence-paired classification strategy with fine-tuning of multilingual pretrained language models, specifically mBERT and XLM-RoBERTa, to compare their performance in bilingual (Indonesian-English) hotel review sentiment analysis.

### 3.1 Data collection

Over 100,000 raw reviews were collected from TripAdvisor and Google Reviews, focusing on 43 five-star hotels in Jakarta. Reviews included in the dataset were those written in Indonesian, English, or a mix of both, provided they contained non-empty textual content beyond a simple rating. Reviews published from 2022 onwards were included to capture post-pandemic guest experiences. Reviews that were empty or contained only a rating were excluded from the dataset.

### 3.2 Data preprocessing

Data cleaning involved removing duplicate reviews (especially those appearing on both platforms) and eliminating empty entries. Reviews were checked for completeness, but no advanced normalization or stop-word removal was performed beyond basic quality control. The dataset was structured to focus on review content. Language filtering was applied to retain only Indonesian and English language reviews.

### 3.3 Labelling and annotation

Aspect and sentiment labels were assigned automatically using a zero shot classification pipeline applied after preprocessing. Seven aspect categories were defined: cleanliness, facilities, food, location, price, room, and service. A multilingual zero-shot classifier was used to assign one or more of these aspects to each review, with a confidence threshold of 0.5 to allow multi-aspect reviews while filtering out low-confidence assignments. This labeling approach follows the methodology adapted from research [3]. After aspect extraction, the dataset was exploded at the aspect level, so that each (review, aspect) pair became a separate record. This transformation yielded 352,720 aspect-level instances from the 96,746 cleaned reviews. A second zero-shot classification step then assigned a sentiment label (positive, negative, or neutral) to each aspect-level instance. To keep the experiments computationally feasible while preserving the original distribution of aspects and sentiment polarities, a simple random sample of 20,000 aspect-level instances was drawn from the exploded and labelled dataset. This sampled subset forms the common basis for all subsequent model training and evaluation.

### 3.4 Feature engineering and representation

Text data was tokenized using the respective tokenizers for each model: WordPiece for mBERT and SentencePiece for XLM-RoBERTa. This process splits words into subword units, reducing out-of-vocabulary issues and enabling the models to handle code-switching and domain-specific vocabulary. Each pair of review and auxiliary sentences was converted into the model-specific input format, including special tokens and embedding representations.

### 3.5 Data splitting and sentence pair construction

The sampled set of 20,000 aspect-level instances was split once into training (70%), validation (15%), and test (15%) subsets using simple random sampling without stratification. The same split indices were reused for all models to ensure a fair comparison between baseline single-sentence classifiers and sentence-pair models.

For the sentence-pair models, each aspect-level instance in the train, validation, and test subsets was transformed into NLI–style sentence pairs. Following prior work, two auxiliary sentences were constructed for each (review, aspect) record: one stating that the aspect is associated with a positive sentiment and one stating that the aspect is associated with a negative sentiment. Each auxiliary sentence was then paired with the original review text to form a sentence pair.

The sentence-pair labels were defined as follows: a pair receives label 1 if the sentiment in the auxiliary sentence matches the zero-shot sentiment label for that aspect (e.g., positive–positive), and 0 otherwise (e.g., positive–negative). Neutral sentiment instances from the zero-shot step were excluded from training, so each aspect-level instance contributes at most two sentence pairs. As a result, the sentence-pair dataset contains exactly twice as many examples as the underlying aspect-level dataset (e.g., 40,000 sentence pairs derived from 20,000 aspect-level instances in the full split).

The baseline models (without sentence pairing) were trained directly on the same 20,000 aspect-level instances using the identical 70/15/15 split, but without constructing auxiliary sentences. This design guarantees that performance differences between baseline and sentence-pair models are attributable to the input formulation rather than differences in data partitions.

### 3.6 Model development

Both mBERT and XLM-RoBERTa were fine-tuned for the sentence-pair classification task. A classification layer was added on top of the pretrained transformer outputs. Training was conducted with a maximum of 10 epochs, a batch size of 128, and a learning rate of 2e-5 for mBERT and 5e-5 for XLM-RoBERTa. Early stopping was applied based on validation loss, so the actual number of epochs varied by model: training stopped automatically when validation performance ceased to improve, resulting in 6 epochs for mBERT and 4 epochs for XLM-RoBERTa in the best runs. Dropout regularization was also used to mitigate overfitting.

Given the pronounced class imbalance in the data, where positive sentiment constitutes more than 90% of all reviews, no artificial balancing (such as oversampling or undersampling) was applied to the training data. Instead, class imbalance was addressed at the algorithmic level by incorporating weighted loss functions during model training. This ensures that the model does not become biased toward the majority class and maintains robust predictive performance for minority classes, such as negative and neutral sentiments.

To evaluate the impact of the sentence-pairing approach, we developed two sets of models for both mBERT and XLM-RoBERTa. The baseline models were fine-tuned without the use of sentence pairing, where each review was classified for aspect and sentiment without pairing with auxiliary sentences.

In contrast, the experimental models were fine-tuned using the sentence-pairing approach, where each review was paired with an auxiliary sentence representing an aspect-sentiment combination. This setup allowed us to directly compare the effect of sentence-pairing on model performance, as both model types were trained and evaluated on the same datasplits.

### 3.7 Evaluation metrics

Model performance was evaluated using the F1-score as the primary metric, providing a balanced measure of precision and recall. Confusion matrices were generated for both models to analyze performance across aspect and sentiment categories. Additional analyses included error analysis (especially for code-switched reviews).

### 3.8 Data analysis and visualization

The best-performing model was subsequently applied to the entire dataset of 96,746 reviews, utilizing the sentence-pair approach to classify each entry. The classification results were then visualized using heatmaps, graphs, and other data visualization techniques. These visualizations provided valuable insights into guest expectations and preferences, helping to identify key trends and areas for improvement within the hotel experience.

### 3.9 Tools and software

Data collection for this research was conducted using Apify, a robust web scraping and automation platform, to extract hotel review data from both Google Reviews and TripAdvisor. Apify enabled efficient, large-scale, and scheduled extraction, ensuring comprehensive coverage of five-star hotel reviews in Jakarta.

All data preprocessing, modelling, and evaluation steps were conducted using Python 3.11, leveraging a suite of modern machine learning libraries. The primary deep learning frameworks include HuggingFace Transformers (version 4.51.3) and PyTorch (version 2.7.0 with CUDA 12.4 support. Experiments were executed in a Google Colab cloud environment, utilizing NVIDIA A100 GPUs and 64 GB RAM to accelerate model training and inference. For data manipulation and analysis, Pandas (version 2.2.3) and NumPy (version 1.26.4) were used, while Matplotlib (3.8.4) and Seaborn (0.13.2) facilitated visualization. Additional libraries such as tqdm were employed for progress tracking, and the datasets' library was used for streamlined data handling. All package versions were explicitly managed and installed to maintain a consistent and reproducible environment across all stages of the workflow.

## 4. THEORETICAL FRAMEWORK AND MODEL RATIONALE

The theoretical justification for using ABSA with multilingual transformer models and sentence-pair classification, as well as the rationale for model selection, is provided in this section.

### 4.1 Theoretical framework

This research is anchored in several foundational theories that explain the influence of online reviews in the hospitality industry and provide the conceptual basis for the analytical approach:

•Social Proof Theory posits that individuals look to the behavior and opinions of others-such as online reviewers-especially in uncertain situations, to guide their own decisions. In the context of hotel selection, positive reviews serve as social validation, shaping consumer trust and booking intentions [15, 16].

•The Information Adoption Model explains how consumers process and utilize information from online reviews. This model suggests that the credibility, relevance, and perceived expertise of the reviewer influence how potential guests interpret and act upon review content [16].

•Expectation Confirmation Theory provides a framework for understanding customer satisfaction, emphasizing the gap between pre-consumption expectations and post-consumption experiences as articulated in reviews [16].

These theories collectively justify the focus on sentiment analysis of online hotel reviews as a method for uncovering actionable insights into guest satisfaction, preferences, and service quality. They also inform the selection of analytical techniques that can capture nuanced, aspect-specific sentiment from diverse and multilingual user-generated content.

### 4.2 Aspect-based sentiment analysis and model rationale

ABSA extends traditional sentiment analysis by identifying not only whether a review is positive or negative overall, but also which specific aspects of the service (such as cleanliness, service, or price) are associated with those sentiments [8, 12]. This level of granularity is particularly important in the hospitality sector, where guest satisfaction depends on multiple dimensions of the hotel experience and management needs to know precisely which aspects drive positive or negative perceptions [12, 13]. ABSA therefore enables hotels to move beyond aggregate ratings and extract targeted, actionable insights from large volumes of unstructured review text [13].

Recent advances in transformer-based language models make ABSA especially effective in multilingual and code-mixed settings [17]. Multilingual models such as mBERT and XLM-RoBERTa are pretrained on large corpora covering many languages, allowing them to capture shared semantic structures across Indonesian and English reviews [14]. Their deep contextual representations help resolve challenges such as informal language, spelling variations, and domain-specific expressions that are common in hotel reviews, providing a robust foundation for aspect-level sentiment classification on bilingual guest feedback [12].

The methodological innovation in this study lies in adapting a sentence-pair classification strategy for ABSA. Instead of treating each review as a single sequence, the task is reformulated so that each review is paired with an auxiliary sentence representing a candidate aspect–sentiment statement, and the model predicts whether the sentiment expressed in the review supports that statement [17]. This formulation is inspired by prior work that recasts ABSA and targeted ABSA as Natural Language Inference–like problems and demonstrates strong performance gains for BERT-based models compared with single-sentence classification [17]. By leveraging transformer models' ability to reason over sentence pairs, this approach improves the detection of aspect-specific sentiment, particularly in reviews that mention multiple

aspects or use implicit, context-dependent expressions [12, 17].

### 4.3 Integration of theory and analytical methods

The integration of these theoretical perspectives with advanced analytical methods ensures that the research not only addresses the technical challenge of extracting nuanced sentiment from multilingual reviews but also aligns with established frameworks on consumer behavior and satisfaction in hospitality. The use of ABSA, supported by transformer-based models, operationalizes the theoretical constructs, such as Social Proof and Expectation Confirmation, by systematically mapping review content to specific aspects and sentiments. This alignment allows for the extraction of actionable insights that are directly relevant to hotel management and industry stakeholders, transforming qualitative guest feedback into quantitative performance metrics.

### 4.4 Methodological considerations and novelty

•Model Selection Justification: The choice of mBERT and XLM-RoBERTa is grounded in their demonstrated effectiveness for cross-lingual and domain-adaptive sentiment analysis tasks. While mBERT provides a stable baseline for multilingual understanding, XLM-RoBERTa offers potential performance advantages due to its larger training corpus and capacity to handle code-switching, a critical feature for Jakarta's bilingual review landscape.

•Sentence-Pairing Approach: By leveraging auxiliary sentences and sentence-pair classification, the study addresses the limitations of traditional single-sentence ABSA approaches, such as difficulty in handling code-switching and domain-specific vocabulary [3].

•Novelty in Application: This research represents one of the first comparative applications of the sentence-pair ABSA strategy specifically for the Indonesian hospitality sector using these two transformer architectures. By benchmarking their performance on a real-world, code-mixed dataset of 5-star hotel reviews, the study provides empirical evidence on the efficacy of transfer learning for operational reputation management.

### 4.5 Limitations

Potential limitations include reliance on automated sentiment labeling, which may introduce labeling noise; the absence of manual validation for all aspect-sentiment pairs; and possible dataset bias toward guests more likely to leave online reviews.

## 5. PROPOSED METHODS

This section builds on the data collection, preprocessing, and aspect labeling pipeline described in Section 3 and focuses on the construction of sentence-pair inputs and the training configuration for mBERT and XLM-RoBERTa.

### 5.1 Sentence-pair input construction

The proposed method reformulates aspect-based sentiment analysis as a sentence-pair classification problem. For each aspect-level instance produced in Section 3, a pair is constructed where the first segment contains the original review text and the second segment contains an auxiliary sentence that encodes the target aspect and sentiment polarity.

This transformation allows the models to jointly attend to both the review content and an explicit description of the aspectsentiment hypothesis, improving discrimination for minority sentiment classes.

### 5.2 Model-specific tokenization and encoding

The constructed sentence pairs are tokenized using the native tokenizers of each model. For mBERT, WordPiece tokenization is applied, and the pair is encoded in the form "[CLS] review text [SEP] auxiliary sentence [SEP]" with segment embeddings distinguishing the review and auxiliary parts. For XLM-RoBERTa, SentencePiece tokenization is used with an analogous sequence layout, relying on special tokens to separate the two segments.

Both models receive identical sentence-pair inputs at the semantic level, which ensures that performance differences can be attributed to architectural and pretraining differences rather than to disparities in input representation.

### 5.3 Model training and fine-tuning

•Both mBERT and XLM-RoBERTa were fine-tuned using the sentence-pairing approach, with a classification head added atop the final transformer layer. For comparison, we also trained baseline versions of each model using single-sentence classification, omitting the auxiliary sentence pairing step. This allowed us to isolate and quantify the contribution of the sentence-pairing methodology to overall model performance.

•Training was performed for up to 10 epochs, using a batch size of 128 and a learning rate of 2e-5 for mBERT and 5e-5 for XLM-RoBERTa. Early stopping was implemented to prevent overfitting: training halted automatically when validation loss stopped improving, resulting in the best models being saved after 6 epochs for mBERT and 4 epochs for XLM-RoBERTa. This approach ensured optimal model performance without unnecessary training cycles.

### 5.4 Rationale for method selection

The sentence-pairing approach is adopted based on the findings of research [3], who demonstrated that this method yields superior performance over single-sentence classification for ABSA tasks with mBERT. This study extends their methodology by including XLM-RoBERTa for direct comparison, leveraging its advantages in multilingual and low-resource contexts [1, 2].

### 5.5 Evaluation metrics and analysis

•Model performance is primarily evaluated using the F1-score, providing a balanced measure of precision and recall for aspect-sentiment classification.

•Confusion matrices are generated to analyze misclassifications across aspect and sentiment categories.

•Additional metrics such as accuracy, precision, and recall are reported as relevant [2].

•Error analysis focuses on challenging cases, including code-mixed and ambiguous reviews, to identify strengths and limitations of each model.

•Confidence analysis compares the probability scores assigned to predictions by both models

### 5.6 Visualization and interpretation

•Visualizations include:
-Bar charts for aspect frequency across hotels.
-Stacked bar or pie charts for sentiment distribution per aspect.
•Heatmaps to illustrate correlations between different aspects.
•Visualizations are designed to accommodate bilingual data and highlight actionable insights for hotel management.

### 5.7 Reproducibility

All code and processed datasets are publicly available at the repositories listed in the Appendix to support reproducibility and further research.

## 6. RESULTS AND DISCUSSION

### 6.1 Results

The experimental results demonstrated the effectiveness of ABSA using multilingual transformer models on Jakarta hotel reviews. Initial preprocessing reduced the raw dataset from 129,519 entries to 96,746 usable reviews after deduplication, language filtering, and text normalization. Zero-shot aspect extraction was then applied, and the data were exploded at the aspect level, yielding 352,720 (review, aspect) instances. A second zero-shot classifier assigned a sentiment label to each instance. From this exploded and labeled dataset, a random sample of 20,000 aspect-level records was drawn to balance computational cost and representativeness. This 20,000-instance subset was split once into training (70%), validation (15%), and test (15%) sets, and the same split was reused for both baseline and sentence-pair models. For the sentence-pair experiments, these aspect-level instances were further transformed into auxiliary sentence–review pairs, as summarized in Table 1.

Table 1 illustrates the transformation of raw review data into aspect-labeled sentence pairs, forming the core input for the sentence-pair modeling approach. Each review is duplicated for each relevant aspect, some entries spanning multiple aspects (e.g., "kamar nya sangat bersih..." is associated with both cleanliness and facilities), while others may remain unassigned (NaN), reflecting the aspect annotation process and the inherent diversity of the dataset.

Both mBERT and XLM-RoBERTa demonstrated robust classification capabilities, with XLM-RoBERTa achieving marginally superior accuracy. On the test set, XLM-RoBERTa attained a classification accuracy of 97.20% (F1-score: 97.29%), outperforming mBERT's 96.63% (F1-score: 96.66%). Cross-validation revealed divergent stability profiles: mBERT exhibited lower variability (±0.35% accuracy) compared to XLM-RoBERTa (±1.84%), suggesting greater consistency in handling linguistic nuances. The Matthews Correlation Coefficient (MCC) further validated XLM-RoBERTa's balanced classification capability (MCC: 0.944 vs. 0.933 for mBERT), particularly in managing class imbalances.

The experimental findings demonstrate that integrating sentence-pair techniques with XLM-RoBERTa yields superior performance in multilingual sentiment analysis for Jakarta's luxury hotel reviews. While both models benefited from sentence pairing, XLM-RoBERTa showed enhanced contextual understanding through its deeper transformer architecture, enabling better resolution of complex multilingual expressions like Indonesian-English code-mixed phrases.

Tables 2 and 3 collectively summarize the performance trajectory of the mBERT model across training and validation phases. Table 2 details the steady improvement in training metrics, with loss decreasing from 0.3 to 0.06 and accuracy rising from 85% to 98% over six epochs, indicating effective learning and model convergence. Complementing this, Table 3 evaluates the model's performance on the validation set, tracking not only accuracy but also precision, recall, and F1-score across epochs. The alignment between rising training and validation scores—culminating in consistent metrics near 97%—demonstrates robust generalization and a low risk of overfitting. This combined analysis reaffirms the efficacy of the training regimen and underscores the model's capacity to maintain high performance on unseen data.

**Table 1.** Sample of exploded review-aspect pairs by aspect

| Text | Aspect |
|---|---|
| kamar nya sangat bersih saat saya berjunjung s... | cleanliness |
| kamar nya sangat bersih saat saya berjunjung s... | facilities |
| Old but legend | NaN |
| Wow amazing cuma sayang lagi pandemi covid ... | location |
| Wow amazing cuma sayang lagi pandemi covid ... | price |
| Wow amazing cuma sayang lagi pandemi covid ... | facilities |

**Table 2.** Training metrics of the mBERT model

| Epoch | Train Loss | Train Accuracy |
|---|---|---|
| 1 | 0.3 | 0.85 |
| 2 | 0.12 | 0.96 |
| 3 | 0.1 | 0.96 |
| 4 | 0.08 | 0.97 |
| 5 | 0.07 | 0.97 |
| 6 | 0.06 | 0.98 |

Note: mBERT = multilingual Bidirectional Encoder Representations from Transformers.

**Table 3.** Validation metrics of the mBERT model

| Epoch | Validation Loss | Validation Accuracy | Validation Precision | Validation Recall | Validation F1 |
|---|---|---|---|---|---|
| 1 | 0.11 | 0.95 | 0.94 | 0.96 | 0.95 |
| 2 | 0.1 | 0.96 | 0.97 | 0.95 | 0.96 |
| 3 | 0.11 | 0.96 | 0.97 | 0.95 | 0.96 |
| 4 | 0.09 | 0.97 | 0.96 | 0.97 | 0.97 |
| 5 | 0.09 | 0.97 | 0.96 | 0.98 | 0.97 |
| 6 | 0.09 | 0.97 | 0.96 | 0.97 | 0.97 |

Note: mBERT = multilingual Bidirectional Encoder Representations from Transformers.

**Table 1.** Training metrics of the XLM-RoBERTa model

| Epoch | Train Loss | Train Accuracy |
|---|---|---|
| 1 | 0.29 | 0.89 |
| 2 | 0.14 | 0.95 |
| 3 | 0.11 | 0.96 |
| 4 | 0.1 | 0.97 |
| 5 | 0.1 | 0.97 |
| 6 | 0.09 | 0.97 |
| 7 | 0.08 | 0.97 |
| 8 | 0.1 | 0.97 |
| 9 | 0.08 | 0.97 |
| 10 | 0.07 | 0.97 |

Note: XLM-RoBERTa = Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach.

**Table 2.** Validation metrics of the XLM-RoBERTa model

| Epoch | Validation Loss | Validation Accuracy | Validation Precision | Validation Recall | Validation F1 |
|---|---|---|---|---|---|
| 1 | 0.15 | 0.95 | 0.96 | 0.95 | 0.95 |
| 2 | 0.12 | 0.96 | 0.96 | 0.96 | 0.96 |
| 3 | 0.09 | 0.97 | 0.95 | 0.99 | 0.97 |
| 4 | 0.09 | 0.97 | 0.96 | 0.98 | 0.97 |
| 5 | 0.09 | 0.97 | 0.96 | 0.98 | 0.97 |
| 6 | 0.09 | 0.97 | 0.97 | 0.98 | 0.97 |
| 7 | 0.10 | 0.97 | 0.96 | 0.98 | 0.97 |
| 8 | 0.10 | 0.97 | 0.96 | 0.97 | 0.97 |
| 9 | 0.08 | 0.97 | 0.96 | 0.99 | 0.97 |
| 10 | 0.08 | 0.98 | 0.97 | 0.98 | 0.98 |

Note: XLM-RoBERTa = Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach.

Tables 4 and 5 collectively illustrate the performance dynamics of the XLM-RoBERTa model over an extended ten-epoch training cycle. Table 4 shows a rapid reduction in training loss from 0.29 to 0.07, alongside an increase in training accuracy from 89% to 97%, indicating effective optimization and early convergence. Complementary to this, Table 5 captures the model's validation performance, with accuracy and F1-scores consistently reaching 97–98% and minimal validation loss throughout the epochs. The close alignment between training and validation metrics suggests minimal overfitting, and the incremental gains realized in extended epochs justify the additional computational investment. Together, these tables demonstrate that prolonged training yields marginal yet meaningful improvements in generalization while maintaining robust classification across unseen data.

These results indicate that XLM-RoBERTa's advantage arises primarily from its broader multilingual pretraining corpus and its tokenizer design, which together improve robustness to noisy, code-mixed hotel reviews. Because XLM-RoBERTa is exposed to far more Indonesian and mixed-language content during pretraining, it can represent infrequent or non-standard expressions in guest reviews more effectively than mBERT, reducing misclassification for minority sentiment classes.

The trajectories in Tables 4 and 5 show that training and validation accuracy remain closely aligned across epochs, with validation F1-scores stabilizing at 0.97–0.98 while validation loss does not increase, which indicates that the model continues to generalize well rather than overfitting to the training data. The absence of a widening gap between training and validation metrics suggests that XLM-RoBERTa is able to exploit the larger sentence-pair dataset without memorizing noise, which is critical for robust deployment on unseen multilingual reviews.

The integration of sentence pairing enhanced classification metrics for both architectures, albeit with computational trade-offs. mBERT showed a 1.12% improvement in F1-score (95.56% → 96.66%) with sentence pairing, while XLM-RoBERTa improved by 0.71% (96.58% → 97.29%). This technique extended training durations — mBERT required 50% more epochs (6 vs. 4), and XLM-RoBERTa's training time increased by 150% (10 vs. 4 epochs). The models' divergent resource requirements highlighted task-specific optimization needs, particularly for real-time hospitality analytics.

For mBERT, the training and validation curves indicate that sentence pairing increases F1-score without introducing divergence between training and validation performance, which implies that the additional contextual signal from auxiliary sentences improves discrimination of minority classes without overfitting. The relatively smooth evolution of validation loss and F1 across epochs further suggests that mBERT benefits from sentence pairing in a stable way, making it suitable for scenarios where predictable behaviour across retraining cycles is important.

In practice, this means that sentence pairing acts as a stronger performance multiplier for mBERT because it aligns closely with its next-sentence pretraining objective and compensates for its smaller, Wikipedia-based corpus, whereas for XLM-RoBERTa the same reformulation mainly refines an already strong decision boundary. Taken together, these findings explain both why XLM-RoBERTa achieves higher peak performance on this ABSA task and why the sentence-pair strategy yields a larger relative improvement for mBERT than for XLM-RoBERTa.

Tables 6 and 7 present the training and validation metrics for the baseline mBERT model, trained without sentence pairing. Table 6 shows that the model achieves rapid convergence, with training accuracy increasing from 94% to 98% and loss decreasing from 0.21 to 0.06 over four epochs. Table 7 complements these findings with validation results: accuracy remains consistently high (96–97%) across all

epochs, and the model maintains stable precision, recall, and F1-scores (95–96%). The close alignment between training and validation scores suggests minimal overfitting and robust generalization, even in the absence of explicit aspect-sentiment relations that sentence pairing provides. This baseline performance offers a critical benchmark against which the benefits of sentence-pair input.

Tables 8 and 9 provide a detailed overview of the baseline XLM-RoBERTa model's training and validation performance, where sentence pairing was not applied. Table 8 shows a steady improvement in training accuracy, increasing from 94% to 97%, while the training loss decreases from 0.21 to 0.08 across four epochs. This indicates efficient model optimization in a relatively short training cycle. Table 9 extends the analysis to the validation set, where accuracy consistently rises from 96% to 97%, with precision, recall, and F1-score remaining stable in the 95–96% range throughout all epochs. The close correspondence between training and validation metrics suggests that the model generalizes well and avoids overfitting, even without explicit aspect pairing. These baseline results serve as a crucial reference point for assessing the added value of sentence-pair input, particularly in terms of nuanced aspect-sentiment detection and the overall robustness observed in models that incorporate this technique.

**Table 6.** Training metrics of the baseline mBERT model

| Epoch | Train Loss | Train Accuracy |
|---|---|---|
| 1 | 0.21 | 0.94 |
| 2 | 0.13 | 0.96 |
| 3 | 0.09 | 0.97 |
| 4 | 0.06 | 0.98 |

Note: mBERT = multilingual Bidirectional Encoder Representations from Transformers.

**Table 7.** Validation metrics of the baseline mBERT model

| Epoch | Validation Loss | Validation Accuracy | Validation Precision | Validation Recall | Validation F1 |
|---|---|---|---|---|---|
| 1 | 0.14 | 0.96 | 0.94 | 0.96 | 0.95 |
| 2 | 0.13 | 0.96 | 0.95 | 0.96 | 0.96 |
| 3 | 0.13 | 0.96 | 0.95 | 0.96 | 0.96 |
| 4 | 0.14 | 0.96 | 0.96 | 0.96 | 0.96 |

Note: mBERT = multilingual Bidirectional Encoder Representations from Transformers.

**Table 8.** Training metrics of the baseline XLM-RoBERTa model

| Epoch | Train Loss | Train Accuracy |
|---|---|---|
| 1 | 0.2 | 0.94 |
| 2 | 0.12 | 0.96 |
| 3 | 0.1 | 0.97 |
| 4 | 0.08 | 0.97 |

Note: XLM-RoBERTa = Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach.

**Table 3.** Validation metrics of the baseline XLM-RoBERTa model

| Epoch | Val idation Loss | Validation Accuracy | Validation Precision | Validation Recall | Validation F1 |
|---|---|---|---|---|---|
| 1 | 0.13 | 0.96 | 0.94 | 0.96 | 0.95 |
| 2 | 0.11 | 0.96 | 0.96 | 0.96 | 0.96 |
| 3 | 0.12 | 0.97 | 0.97 | 0.97 | 0.96 |
| 4 | 0.13 | 0.97 | 0.96 | 0.97 | 0.96 |

Note: XLM-RoBERTa = Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach.

Comparing the baseline and sentence-pair configurations across Tables 6–9 shows that sentence pairing yields consistent gains in F1-score for both architectures without destabilizing validation metrics, indicating that the reformulation to a sentence-pair NLI-style task adds discriminative information rather than simply increasing model capacity. This pattern supports the interpretation that explicit aspect–sentiment pairing helps both mBERT and XLM-RoBERTa focus on fine-grained opinion cues in code-mixed reviews, rather than memorizing surface-level correlations.

Heatmap analysis generated after running all the data through the fine-tuned XLM-RoBERTa model, identified cleanliness as the most positively perceived aspect (99.4% positive sentiment), while pricing generated the highest negativity (11.2% negative sentiment). Facilities and service quality dominated review frequency, accounting for 79,713 and 78,773 mentions, respectively. Notably, 79% of negative pricing reviews coexisted with positive remarks on other aspects, reflecting guests' willingness to critique specific elements while maintaining overall satisfaction.

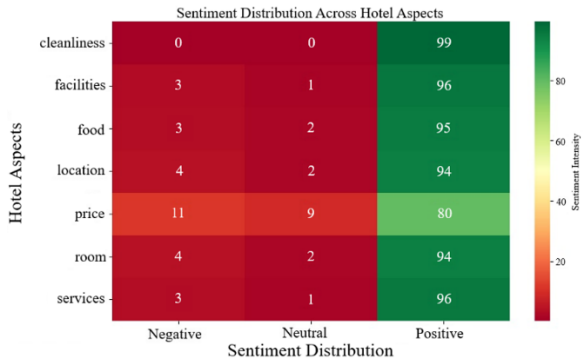


**Figure 1.** Heatmap distribution of sentiment

As visualized in Figure 1, positive sentiment dominates almost all aspects, with cleanliness and facilities showing the highest concentration of strong positive polarity, while price exhibits a visibly thicker band of negative values than other aspects. This pattern confirms that Jakarta's five-star hotels are

consistently praised for tangible quality attributes such as cleanliness, facilities, and service, whereas perceived value for money remains a relative weakness. The heatmap also reveals that negative opinions on price often co-occur with positive sentiment on other aspects, indicating that guests are generally satisfied with the overall experience but remain sensitive to pricing and cost-related issues.

Based on the results of running the entire dataset through the fine-tuned XLM-RoBERTa model, the aspect-based sentiment classification produced clear trends across all major hotel service dimensions. As summarized in the results table, the model identified overwhelmingly positive sentiment for most aspects: facilities (79,713 positive), services (78,773 positive), cleanliness (33,324 positive), rooms (41,359 positive), location (64,574 positive), and food (29,608 positive). Negative sentiment was most frequently associated with location (2,712) and services (2,690), while the price aspect showed the highest proportion of negative (791) and neutral (624) classifications relative to its total mentions. Cleanliness stood out with the lowest number of negative (100) and neutral (114) cases, and no errors, indicating exceptionally consistent guest satisfaction in this area. Across all aspects, the number of ambiguous or error cases detected by the model remained minimal (e.g., 24 for location, 20 for facilities, and 101 errors in total), demonstrating both the robustness of the classification and the clarity of guest feedback in the dataset. These results highlight that, for five-star hotels in Jakarta, facilities, services, and cleanliness are the most consistently praised aspects, while price remains the most sensitive and divisive factor among guests.

Table 10 summarizes the distribution of sentiment predictions across the seven key aspects in the test set, including counts for positive, negative, and neutral classifications, as well as errors. The results highlight exceptionally high positive sentiment for cleanliness, facilities, and service, reinforcing these as core strengths for Jakarta's five-star hotels. In contrast, the aspect of price records notably higher negative sentiment, indicating ongoing guest dissatisfaction in this area. The relatively low error counts across all aspects further illustrate the reliability and robustness of the classification pipeline. This table provides crucial insights for hotel management, offering a data-driven basis for targeting improvements and sustaining high service standards.

**Table 10.** Distribution of sentiment classification results per aspect

| Aspect | Positive | Negative | Neutral | Error |
|---|---|---|---|---|
| Cleanliness | 33.324 | 100 | 114 | 0 |
| Facilities | 79.713 | 2.311 | 972 | 20 |
| Food | 29.608 | 982 | 467 | 9 |
| Location | 64.574 | 2.712 | 1.08 | 24 |
| Price | 5.642 | 791 | 624 | 8 |
| Room | 41.359 | 1.698 | 726 | 16 |
| Service | 78.773 | 2.69 | 989 | 24 |

From a managerial perspective, Table 10 highlights that interventions should prioritize pricing policies and communication, since price is the only aspect where negative and neutral sentiments form a substantial share of all mentions, while operational efforts in cleanliness, facilities, and service can focus on maintaining already strong performance. The very low error counts across aspects also indicate that these patterns are not artifacts of model uncertainty, but reflect consistent signals in guest feedback that can guide targeted improvements.

### 6.2 Discussion

6.2.1 Addressing the research questions

The results of this study reveal important insights about the application of multilingual language models for sentiment analysis of hotel reviews. The findings demonstrate that both mBERT and XLM-RoBERTa models with sentence pairing techniques can effectively classify sentiment in multilingual hotel reviews, with XLM-RoBERTa showing slightly higher performance metrics but requiring more computational resources. The analysis also provides valuable insights into guest experiences at 5-star hotels in Jakarta, identifying cleanliness, facilities, and service as highly appreciated aspects, while price emerges as a significant challenge.

6.2.2 Success of the methodology

The sentence pairing technique proved highly effective for sentiment analysis in this domain, significantly improving performance for both models compared to their respective baselines. For mBERT, the implementation of sentence pairing increased the F1-score by 1.12%, from 95.56% to 96.66% on test data, while precision improved by 1.41%, from 95.41% to 96.82%. Similarly, XLM-RoBERTa showed improvements with sentence pairing, though the magnitude was somewhat smaller with a 0.71% increase in F1-score and a 0.22% increase in precision. These findings suggest that the sentence-pairing approach may be particularly beneficial for mBERT, potentially due to its architecture's greater responsiveness to contextual information provided through paired sentences.

The zero-shot classification approach for aspect extraction also proved successful, allowing for efficient identification of multiple aspects within single reviews without requiring extensive labeled data. This addresses a common challenge in ABSA, particularly for languages with limited resources, such as Indonesian. The comprehensive evaluation metrics, including Matthews Correlation Coefficient (MCC) and Area Under the ROC Curve (AUC), further validate the robustness of the models, with both achieving AUC values exceeding 0.99.

6.2.3 Performance comparison with previous studies

The findings align with and extend previous research in several key ways. The superior performance of XLM-RoBERTa compared to mBERT is consistent with observations done by research [2], who found that XLM-R outperformed mBERT in text classification tasks for the Indonesian language. Their study attributed this advantage to XLM-R's larger architecture and more extensive, balanced training dataset. However, our research adds nuance to this finding by demonstrating that while XLM-RoBERTa achieves higher peak performance, mBERT offers significantly better stability across cross-validation folds, with 5.3 times lower standard deviation in accuracy.

The effectiveness of sentence pairing for ABSA confirms and builds upon the findings from research [3], who introduced this approach for Indonesian hotel reviews using mBERT. Our results extend their work by demonstrating the applicability of this technique to XLM-RoBERTa and providing a direct performance comparison between the two models. The observed improvements in F1-scores are

comparable to their reported enhancements, though our study reveals interesting differences in how each model responds to the sentence pairing intervention.

Our findings regarding the importance of specific hotel aspects align with research [18], who also found that aspects related to service and facilities significantly influenced guest satisfaction in luxury hotels. However, our results add the insight that cleanliness emerged as the most consistently positively rated aspect (99.4% positive sentiment), while price represented the greatest challenge (11.2% negative sentiment), which may reflect the specific context of Jakarta's luxury hotel market.

6.2.4 Understanding guest experiences

The comprehensive sentiment analysis across seven key aspects provides valuable insights for hotel management. The extremely high positive sentiment for cleanliness (99.4%) suggests that 5-star hotels in Jakarta are exceeding guest expectations in this domain. Similarly, the strong positive sentiment towards facilities (95.8%) and service (96.0%) indicates these are core strengths that should be maintained. These findings align with the regulatory requirements outlined in Peraturan Kementerian Pariwisata dan Ekonomi Kreatif Nomor 4 Tahun 2021, which emphasizes standards for facilities, personnel, and service management.

Conversely, the relatively high negative sentiment regarding price (11.2%) suggests a potential misalignment between guest expectations and perceived value. This presents an opportunity for hotel management to either adjust pricing strategies or enhance the communication of value propositions to guests. The keyword analysis, which highlighted terms like "staff," "kamar," and "breakfast" across sentiment categories, provides further granularity in understanding specific elements that contribute to guest satisfaction or dissatisfaction.

6.2.5 Methodological implications

From a methodological perspective, the findings of this study provide several important insights for the design and deployment of multilingual ABSA models in the hospitality domain. The first concerns the trade-off between peak performance and stability when comparing XLM-RoBERTa and mBERT. XLM-RoBERTa consistently achieved higher accuracy, F1-score, and lower loss on the test set than mBERT, particularly for minority sentiment classes, demonstrating a clear advantage when maximum predictive performance is the primary objective. At the same time, cross-validation results showed that XLM-RoBERTa exhibited greater variability across folds, whereas mBERT produced more stable and predictable scores. This pattern suggests that XLM-RoBERTa is better suited for one-off, high-precision analytical tasks, while mBERT may be preferable in operational settings that require regular retraining and highly consistent behaviour over time.

The architectural differences between the two models help explain this behaviour. XLM-RoBERTa is pretrained on the CommonCrawl-based CC100 corpus, which is substantially larger and more balanced across 100 languages than the Wikipedia-based corpus used for multilingual BERT (mBERT) [1, 18]. This broader and more heterogeneous pretraining data makes XLM-RoBERTa more robust to noisy, user-generated content and better able to model low-resource languages such as Indonesian, especially when mixed with English in code-switched reviews. By contrast, mBERT's reliance on Wikipedia text makes it more sensitive to domain shift when exposed to informal hotel reviews containing slang, spelling variations, and non-standard expressions [19, 20]. These pretraining differences are reflected in the empirical results of this study, where XLM-RoBERTa more effectively captures nuanced aspect–sentiment relations, particularly for negative and neutral classes that are under-represented in the dataset, in line with previous comparative evaluations of the two models on multilingual benchmarks [1].

Tokenizer design further contributes to XLM-RoBERTa's advantage. XLM-RoBERTa employs a SentencePiece tokenizer with a language-agnostic subword vocabulary, which is well suited to handling morphologically rich words and mixed-language tokens that frequently appear in Indonesian–English hotel reviews [1, 20]. This configuration allows rare, misspelled, or code-mixed terms to be represented with relatively few, semantically meaningful subwords. In contrast, mBERT's WordPiece tokenizer, as introduced in the original BERT architecture, tends to segment such tokens into longer subword sequences [19]. This increases sequence length and can dilute important contextual cues. In practical terms, XLM-RoBERTa can form more coherent representations for domain-specific expressions such as "upgrade kamar", "late checkout", or mixed Indonesian–English phrases, which supports its higher peak performance on the ABSA task observed in this study.

The results also clarify why the sentence-pairing strategy benefits the two models to different degrees. mBERT was pretrained not only with masked language modeling but also with a next-sentence prediction (NSP) objective [19]. Reformulating ABSA as a sentence-pair classification problem—where each review is paired with an auxiliary aspect–sentiment sentence—closely aligns with this NSP-style pretraining and directly leverages mBERT's capacity to model relationships between two textual segments. Prior work has shown that such sentence-pair formulations can substantially improve ABSA performance for mBERT in Indonesian settings [3, 21, 22], and the present study confirms this pattern: introducing sentence pairing increased mBERT's F1-score by 1.12 points and precision by 1.41 points on the test set compared to its single-sentence baseline. In effect, sentence pairing provides mBERT with additional task structure that sharpens aspect–sentiment boundaries, particularly in reviews containing multiple aspects or implicit sentiment.

XLM-RoBERTa, in contrast, is pretrained solely with a masked language modeling objective and does not include an explicit NSP component [1]. Its strong single-sequence contextual representations, learned from a much larger multilingual corpus, already capture many aspect–sentiment relationships without requiring an auxiliary sentence. As a result, reframing ABSA as a binary sentence-pair classification task yields a smaller but still positive improvement: in this study, sentence pairing increased XLM-RoBERTa's F1-score by 0.71 points and precision by 0.22 points relative to its baseline. Intuitively, sentence pairing refines an already strong decision boundary for XLM-RoBERTa, while it acts as a stronger performance multiplier for mBERT by aligning more directly with its pretraining objectives and compensating for its more limited pretraining corpus.

These observations have direct methodological implications for model and task design. For applications where computational resources are constrained or where stability and reproducibility across training runs are critical, mBERT with

sentence pairing offers an attractive balance between accuracy and consistency. For use cases that prioritize maximum discriminatory power on noisy, code-mixed data and can tolerate somewhat greater variance across runs, XLM-RoBERTa with sentence pairing is a preferable choice. More broadly, the results suggest that sentence-pair formulations are particularly valuable for models whose pretraining includes sentence-level relationship objectives, while models trained solely with masked language modeling may derive more limited incremental benefit from the same task reformulation.

Finally, the study demonstrates that combining zero-shot classification for aspect and sentiment labeling with sentence-pair fine-tuning is a practical and effective strategy in low-resource settings. Zero-shot labeling made it possible to construct a large, aspect-specific training set without manual annotation, while the subsequent sentence-pair models achieved high AUC values (above 0.99) and strong Matthews Correlation Coefficients despite pronounced class imbalance. This pipeline shows that domain practitioners can obtain robust, aspect-level sentiment models for multilingual hotel reviews even when labeled data are scarce, provided that careful attention is paid to model choice, tokenizer behaviour, and the alignment between pretraining objectives and the downstream task formulation.

## 7. CONCLUSION

This study demonstrates a comprehensive approach to ABSA on multilingual online reviews of five-star hotels in Jakarta, leveraging advanced transformer models—mBERT and XLM-RoBERTa—with the sentence pairing technique. By systematically evaluating both models across multiple metrics and validation strategies, the research provides a nuanced understanding of model effectiveness in a real-world, multilingual hospitality context.

The findings position XLM-RoBERTa with sentence pairing as the optimal model for this task, achieving the highest performance across all evaluated metrics, including accuracy, F1-score, MCC, and AUC, while also demonstrating superior capability in handling nuanced sentiment detection across Indonesian and English reviews. mBERT, while slightly less accurate in aggregate (accuracy 96.63%, F1-score 96.66%, MCC 0.9326, AUC 0.9944), exhibited greater consistency and stability across validation folds (standard deviation of accuracy ± 0.35% vs. XLM-RoBERTa's ± 1.84%), highlighting its suitability for applications where reproducibility and reliability are paramount.

From a domain perspective, the analysis of over 350,000 aspect-level review entries reveals that cleanliness, facilities, and service are consistently perceived as the strongest attributes of Jakarta's five-star hotels, each receiving over 94% positive sentiment—cleanliness in particular stands out with a 99.4% positive rating. In contrast, price emerges as the most critical area for improvement, drawing the highest proportion of negative sentiment (11.2%) and the lowest sentiment intensity among all aspects. Key drivers of guest satisfaction, as reflected in the most frequent positive keywords, include staff friendliness, room quality, and breakfast experience, while recurring negative terms point to issues with administrative processes and perceived value for money.

This research advances the field by empirically validating the effectiveness of sentence pairing for ABSA in a multilingual, domain-specific setting and by directly comparing two multilingual transformer models. The methodology—combining zero-shot aspect extraction, fine-tuned sentence pairing, and robust cross-validation—offers a scalable framework that can be adapted to other languages, domains, or service industries facing similar challenges in multilingual user-generated content.

Future applications of this work may include real-time sentiment monitoring dashboards for hotel management, integration with customer relationship management (CRM) systems to trigger targeted service improvements, and extension to other segments of the hospitality or tourism industry. Further research is encouraged to address the remaining challenges of handling ambiguous or mixed-sentiment reviews, optimizing computational efficiency for large-scale deployment, and exploring the impact of additional contextual features such as temporal trends or reviewer demographics.

By providing actionable, aspect-level insights into guest experiences, this study supports data-driven decision-making for hotel operators in Jakarta's competitive luxury market and contributes a replicable approach for sentiment analysis in multilingual, service-oriented domains.

## 8. FUTURE WORK

### 8.1 Current limitations

This research, while providing valuable insights into ABSA for five-star hotel reviews in Jakarta, is subject to several limitations:

•Assumptions: The study assumes that online reviews from TripAdvisor and Google Reviews accurately represent guest experiences at five-star hotels in Jakarta. It also presumes that the multilingual models (mBERT and XLM-RoBERTa) can handle mixed Indonesian-English reviews without significant bias toward either language.

•Number of variables: The analysis was limited to seven main aspects (room, location, services, price, food, cleanliness, and facilities) extracted using zero-shot classification. Other potentially relevant aspects, such as technology, security, or unique guest experiences, were not included in this study.

•Constraints:

-Data limitations: Only 129,519 reviews in Indonesian and English from 2022 onwards were analyzed. Reviews in other languages or from earlier periods were excluded, which may limit the generalizability of findings.

-Sampling: Of the total aspect-level data generated (over 350,000 entries), only 20,000 samples were used for model training and evaluation to ensure computational efficiency. This means a broader range of data variability was not fully captured.

-Model and methodology: The study focused solely on comparing two multilingual transformer models (mBERT and XLM-RoBERTa) using the sentence-pairing approach. Other models or ensemble techniques that could potentially enhance performance were not explored.

•Aspect Detection Pipeline: The current models are unable to autonomously identify specific aspects from raw reviews, necessitating a separate zero-shot learning step for aspect extraction prior to sentiment classification. This two-stage approach may limit end-to-end efficiency and integration.

### 8.2 Recommendations for future research

To address these limitations and advance the field, several directions are recommended for future studies:

•Evaluate sentence-pair ABSA across architectures: Future research should systematically test the sentence-pair classification strategy on additional multilingual encoders such as DeBERTa-v3, mT5, or LLaMA-based models, to determine whether the performance gains observed for mBERT and XLM-RoBERTa in this study generalize across different pretraining objectives and architectures.

•Develop joint aspect sentiment models: Instead of separating zero-shot aspect extraction and supervised sentiment classification, future studies could design multi-task or end-to-end architectures that simultaneously detect aspects and predict their sentiment, thereby reducing error propagation between stages and simplifying deployment in production hotel review analytics pipelines.

•Broaden variables and aspects: Future work may include additional aspects such as technology features, security, or sustainability-related attributes to capture a more comprehensive picture of guest expectations and to test whether the current seven-aspect schema remains sufficient for luxury hotel analysis in evolving market conditions.

•Widen data sources, timeframes, and regions: Collecting data from more review platforms, extending the temporal range beyond the current post-pandemic window, and incorporating reviews from different hotel categories or cities (for example, other major Indonesian or Southeast Asian destinations) would enable cross-city or cross-country comparisons and help assess the robustness of the proposed approach under domain shift.

•Improve handling of class imbalance and rare patterns: Subsequent studies should explore more advanced techniques for dealing with highly skewed sentiment distributions, such as focal loss, curriculum sampling, or contrastive learning, to further enhance performance on negative and neutral classes and on aspects with fewer mentions, such as price-related complaints identified in this work.

By addressing these limitations and pursuing the suggested directions, future research can provide deeper and more comprehensive insights into ABSA in the hospitality industry, particularly in multilingual and dynamic urban contexts like Jakarta.

## ACKNOWLEDGMENT

The data collection process for this study was greatly facilitated by Apify application, who provided a generous discount on their data scraping services.This manuscript was supported by Perplexity AI, whose tools and platform assisted in the editing process and the development of research notebooks. Their technological support contributed to the clarity and rigor of the analysis.

The author contributed to the writing as follows, A.W. Suryabrata conducted the initial investigation, collected research materials as directed by course assignments, and designed the proposed model under the supervision of their academic advisor. H.L.H.S. Warnars contributed to the conceptualization of the study, provided critical review and editing of the manuscript, and facilitated its submission. M.K. Muyeba offered supervisory guidance on the research topic and reviewed the manuscript.

## REFERENCES

[1] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

[2] Wiciaputra, Y.K., Young, J.C., Rusli, A. (2021). Bilingual text classification in English and Indonesian via transfer learning using XLM-RoBERTa. International Journal of Advances in Soft Computing & Its Applications, 13(3): 73-87. https://doi.org/10.15849/ijasca.211128.06

[3] Azhar, A.N., Khodra, M.L. (2020). Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis. In 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, pp. 1-6. https://doi.org/10.1109/icaicta49861.2020.9428882

[4] Sparks, B.A., Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. Tourism Management, 32(6): 1310-1323. https://doi.org/10.1016/j.tourman.2010.12.011

[5] Hu, X.B., Yang, Y. (2020). Determinants of consumers' choices in hotel online searches: A comparison of consideration and booking stages. International Journal of Hospitality Management, 86: 102370. https://doi.org/10.1016/j.ijhm.2019.102370

[6] Vo, N.T., Hung, V.V., Tuckova, Z., Pham, N.T., Nguyen, L.H. (2022). Guest online review: An extraordinary focus on hotel users' satisfaction, engagement, and loyalty. Journal of Quality Assurance in Hospitality & Tourism, 23(4): 913-944. https://doi.org/10.1080/1528008X.2021.1920550

[7] Hananto, A. (2015). Application of text mining to extract hotel attributes and construct perceptual map of five star hotels from online review: Study of Jakarta and Singapore five-star hotels. ASEAN Marketing Journal, 7(2): 58-80. https://doi.org/10.21002/amj.v7i2.5262

[8] Liu, B. (2012). Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). University of Illinois: Chicago, IL, USA.

[9] Said, F., Manik, L.P. (2022). Aspect-based sentiment analysis on Indonesian presidential election using deep learning. Paradigma-Jurnal Komputer dan Informatika, 24(2): 160-167. https://doi.org/10.31294/paradigma.v24i2.1415

[10] Dewi, M.T., Herdiani, A., Kusumo, D.S. (2018). Multi-aspect sentiment analysis komentar wisata tripadvisor dengan rule-based classifier (Studi Kasus: Bandung Raya). eProceedings of Engineering, 5(1): 1589-1596.

[11] Perdana, S.A.P., Aji, T.B., Ferdiana, R. (2021). Aspect category classification Dengan Pendekatan Machine Learning Menggunakan Dataset Bahasa Indonesia. Jurnal Nasional Teknik Elektro Dan Teknologi Informasi, 10(3): 229-235. https://doi.org/10.22146/jnteti.v10i3.1819

[12] Fehle, J., Münster, L., Schmidt, T., Wolff, C. (2023). Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews. In Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pp.

202-218.
[13] Nawawi, I., Ilmawan, K.F., Maarif, M.R., Syafrudin, M. (2024). Exploring tourist experience through online reviews using aspect-based sentiment analysis with zero-shot learning for hospitality service enhancement. Information, 15(8): 499. https://doi.org/10.3390/info15080499
[14] Putra, I.F., Purwarianti, A. (2020). Improving Indonesian text classification using multilingual language model. In 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, pp. 1-5. https://doi.org/10.1109/ICAICTA49861.2020.9429038
[15] Thumvichit, A., Gampper, C. (2019). Composing responses to negative hotel reviews: A genre analysis. Cogent Arts & Humanities, 6(1): 1629154. https://doi.org/10.1080/23311983.2019.1629154
[16] Hlee, S., Lee, H., Koo, C. (2018). Hospitality and tourism online review research: A systematic analysis and heuristic-systematic model. Sustainability, 10(4): 1141. https://doi.org/10.3390/su10041141
[17] Sun, C., Huang, L., Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588. https://doi.org/10.18653/v1/n19-1035
[18] Özen, İ.A., Özgül Katlav, E. (2023). Aspect-based sentiment analysis on online customer reviews: A case study of technology-supported hotels. Journal of Hospitality and Tourism Technology, 14(2): 102-120. https://doi.org/10.1108/jhtt-12-2020-0319
[19] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, pp. 4171-4186. https://doi.org/10.18653/v1/n19-1423
[20] Pires, T., Schlinger, E., Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4996-5001. https://doi.org/10.18653/v1/P19-1493
[21] Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, pp. 66-71. https://doi.org/10.18653/v1/D18-2012
[22] Yu, T., Joty, S. (2021). Effective fine-tuning methods for cross-lingual adaptation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, pp. 8492-8501. https://doi.org/10.18653/v1/2021.emnlp-main.668

## NOMENCLATURE

| | |
|---|---|
| F1 | F1-score, dimensionless |

**Subscripts**

| | |
|---|---|
| Acc | Accuracy |
| Test | Test dataset |
| Train | Training dataset |
| Val | Validation dataset |

**Abbreviations**

| | |
|---|---|
| ABSA | Aspect-Based Sentiment Analysis |
| AUC | Area Under the Curve |
| BERT | Bidirectional Encoder Representations from Transformers |
| CRM | Customer Relationship Management |
| CUDA | Compute Unified Device Architecture |
| GPU | Graphics Processing Unit |
| MCC | Matthews Correlation Coefficient |
| mBERT | multilingual BERT |
| NLI | Natural Language Inference |
| NLP | Natural Language Processing |
| OOV | Out-of-Vocabulary |
| ROC | Receiver Operating Characteristic (curve) |
| RoBERTa | Robustly Optimized BERT Approach |
| SVM | Support Vector Machine |
| XLM-RoBERTa | Cross-lingual Language Model - Robustly Optimized BERT Approach |

## APPENDIX

The complete Python pipeline is available in the following GitHub-hosted notebook:

**Data cleaning**

https://github.com/Orneus/Thesis/blob/7771c3a72d22f52604e2f9306581e324e4c71397/ABSA_Cleaning.ipynb

**Model training**

https://github.com/Orneus/Thesis/blob/db3f2dedda4bdea20f1391b244bbf9d5c0b76ca5/ABSA_Processing__Final.ipynb

**Data processing**

https://github.com/Orneus/Thesis/blob/db3f2dedda4bdea20f1391b244bbf9d5c0b76ca5/ABSA_deployment.ipynb