

Benchmarking Multilingual Sentiment Analysis Models for the Hospitality Industry: A Case Study of Hotel Reviews in Vietnam

Quoc Lap Nguyen

lapnq22413@st.uel.edu.vn

Vietnam National University, Ho Chi Minh City

Research Article

Keywords: Sentiment analysis, Multilingual NLP, Hospitality industry, Transformer models, Hotel reviews

Posted Date: March 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8680561/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Benchmarking Multilingual Sentiment Analysis Models for the Hospitality Industry: A Case Study of Hotel Reviews in Vietnam

Quoc Lap Nguyen^{1*}

^{1*}University of Economics and Law, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam.

Corresponding author(s). E-mail(s): lapnq22413@st.uel.edu.vn;

Abstract

The explosive growth of online tourism platforms has generated massive multilingual hotel reviews, creating both opportunities and challenges for hospitality businesses in emerging markets. Vietnamese hotels receive feedback in multiple languages, making manual sentiment analysis impractical. This study benchmarks five state-of-the-art Transformer models (XLM-RoBERTa, mBERT, mDeBERTa, DistilBERT, multilingual-E5) for automated sentiment classification using 59,377 authentic hotel reviews spanning 14 languages and five sentiment categories with highly imbalanced distributions reflecting realistic business scenarios. Results reveal that multilingual-E5 achieves the highest overall performance (82% accuracy, macro F1=0.62) with superior minority class handling, while DistilBERT provides comparable accuracy (80%) with significantly reduced computational requirements. Critically, XLM-RoBERTa exhibits catastrophic failure on minority classes despite strong benchmark performance (77% accuracy, 0.03 recall on negative reviews), demonstrating that standard NLP benchmarks do not predict domain-specific effectiveness. We provide evidence-based model selection guidelines linking business characteristics to appropriate choices, quantitative cost-benefit analysis demonstrating 710% ROI for typical deployments, and actionable implementation strategies. These findings enable small and medium hospitality enterprises to adopt AI-powered sentiment analysis sustainably, supporting UN Sustainable Development Goals 8, 9, and 12 by democratizing access to sophisticated NLP capabilities while promoting responsible computational practices.

Keywords: Sentiment analysis, Multilingual NLP, Hospitality industry, Transformer models, Hotel reviews

1 Introduction

The digital transformation of the tourism and hospitality industry has fundamentally changed how businesses understand and respond to customer feedback. Online reviews have evolved beyond simple feedback mechanisms to become critical data assets that directly influence revenue management, service quality improvement, and competitive positioning [Xiang et al. \(2017\)](#). In emerging tourism markets such as Vietnam, the challenge is amplified by the inherently multilingual nature of customer feedback, where hotels routinely receive reviews in English, Vietnamese, Korean, French, German, and numerous other languages from international travelers.

1.1 The Evolution of Electronic Word-of-Mouth in Hospitality

Electronic word-of-mouth (e-WOM) through online reviews has become one of the most influential factors in travel decision-making [Litvin et al. \(2008\)](#). Research indicates that a one-point increase in online ratings can lead to a 5-9% increase in hotel revenue [Anderson and Xie \(2010\)](#). However, the volume of reviews has grown exponentially, with major booking platforms processing millions of reviews monthly, creating an overwhelming data analysis challenge for hotel management teams.

Traditional manual review analysis is no longer viable for several reasons. First, the sheer volume of reviews exceeds human processing capacity. Second, language barriers prevent comprehensive understanding when reviews span multiple languages. Third, the time lag between review posting and managerial response can negatively impact customer relationships. Fourth, subjective interpretation by different staff members leads to inconsistent service improvements.

1.2 The Multilingual Challenge in Emerging Markets

Vietnam’s tourism industry exemplifies the multilingual challenge facing hospitality businesses in emerging markets. As an increasingly popular destination for international travelers, Vietnamese hotels must cater to diverse customer bases with different cultural backgrounds and languages. Our dataset analysis reveals that English reviews constitute 55% of feedback, Vietnamese 20%, followed by French, Korean, German, Dutch, and others. This linguistic diversity creates three critical operational challenges:

First, hotel staff typically cannot fluently understand all languages, resulting in delayed or missed responses to negative feedback from international guests. **Second**, machine translation services often fail to capture sentiment nuances, particularly with internet slang, emojis, and culturally-specific expressions common in online reviews. **Third**, the cost of employing multilingual staff or translation services becomes prohibitive for small and medium enterprises (SMEs) with limited budgets.

1.3 The Promise and Challenge of Transformer-Based NLP

Recent advances in Natural Language Processing (NLP), particularly Transformer-based models, offer potential solutions to these challenges. Models such as BERT [Devlin et al. \(2019\)](#), XLM-RoBERTa [Conneau et al. \(2020\)](#), and their variants

have demonstrated remarkable multilingual capabilities. However, significant questions remain unanswered for practical business deployment:

- Which models perform best on real-world hospitality data with highly imbalanced sentiment distributions?
- What trade-offs exist between model accuracy and computational efficiency for resource-constrained businesses?
- How do models handle minority sentiment classes (negative, terrible) that are crucial for service recovery?
- What are the practical cost-benefit considerations for different hotel market segments?

1.4 Research Objectives and Contributions

This research addresses these gaps through a comprehensive benchmark study with the following objectives:

Objective 1: Evaluate five state-of-the-art Transformer models (XLM-RoBERTa, mBERT, mDeBERTa, DistilBERT, multilingual-E5) on real-world multilingual hotel review data from Vietnam.

Objective 2: Analyze the accuracy-efficiency trade-off across models, considering both classification performance and computational requirements.

Objective 3: Examine model behavior on imbalanced data reflecting realistic business scenarios where positive reviews vastly outnumber negative ones.

Objective 4: Provide actionable managerial implications for hospitality businesses of different sizes and resources.

Our contributions to the literature and practice are threefold. **First**, we provide the first comprehensive benchmark of modern Transformer models specifically for multilingual sentiment analysis in the hospitality domain, using real-world data from an emerging market. **Second**, we reveal critical performance differences between models, particularly their handling of minority classes essential for customer service recovery. **Third**, we translate technical findings into practical business recommendations aligned with sustainable AI adoption principles, addressing the United Nations Sustainable Development Goals (SDG 8: Decent Work and Economic Growth, SDG 9: Industry, Innovation and Infrastructure).

The remainder of this paper is organized as follows. Section 2 reviews related literature on sentiment analysis in tourism and multilingual NLP. Section 3 describes our methodology including dataset characteristics, model selection rationale, and evaluation metrics. Section 4 presents empirical results with detailed performance analysis. Section 5 discusses findings and their implications. Section 6 provides managerial recommendations, and Section 7 concludes with limitations and future research directions.

2 Literature Review

2.1 Sentiment Analysis in Tourism and Hospitality

Sentiment analysis has become an essential tool for understanding customer perceptions in the tourism industry. Early research primarily employed lexicon-based approaches using predefined sentiment dictionaries [Liu \(2010\)](#). While computationally efficient, these methods struggled with context-dependent sentiment, sarcasm, and domain-specific language.

The deep learning revolution further transformed sentiment analysis capabilities. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks enabled better handling of sequential text data [Socher et al. \(2013\)](#). However, these approaches faced limitations in processing efficiency and capturing long-range dependencies in text. Previous studies have extensively applied machine learning and deep learning techniques to analyze customer reviews in the hospitality sector [Alaei et al. \(2019\)](#); [Kirilenko et al. \(2018\)](#).

Several studies have employed sentiment analysis within hospitality contexts to derive insights from online reviews. For instance, Mariani and Borghi [Mariani and Borghi \(2021\)](#) utilized text analytics to investigate customers' evaluations of mechanical artificial intelligence and service robots in hotel settings. Xiang et al. [Xiang et al. \(2015\)](#) demonstrated how sentiment analysis of social media can provide competitive intelligence for tourism businesses. However, most existing studies focus exclusively on English-language reviews, leaving a critical gap in understanding multilingual sentiment analysis performance.

2.2 Multilingual Natural Language Processing

The development of multilingual NLP models represents a major advancement for global business applications. Traditional approaches required separate models for each language, creating prohibitive maintenance costs and inconsistent performance across languages. Cross-lingual transfer learning emerged as a solution, enabling models trained on high-resource languages to generalize to low-resource languages [Ruder et al. \(2019\)](#).

Multilingual BERT (mBERT) [Devlin et al. \(2019\)](#) pioneered the pre-training of a single model on 104 languages, demonstrating surprising zero-shot cross-lingual capabilities. XLM-RoBERTa [Conneau et al. \(2020\)](#) extended this approach with improved training on 100 languages using more data and better optimization. More recent models like mDeBERTa [He et al. \(2021\)](#) incorporated disentangled attention mechanisms, while embedding-based models like multilingual-E5 [Wang et al. \(2022\)](#) explored alternative architectures for semantic understanding.

Despite these advances, research specifically examining multilingual model performance in business contexts remains limited. Most benchmark studies use academic datasets with balanced classes and formal language, unlike the imbalanced, colloquial text typical in actual customer reviews. Furthermore, studies rarely consider the computational costs and deployment feasibility crucial for business decision-making.

2.3 Challenges in Real-World Hospitality Data

Real-world hotel review data presents several unique challenges that differentiate it from standard NLP benchmarks. **First**, sentiment distributions are highly imbalanced. Positive reviews typically outnumber negative reviews by factors of 10 to 20, as satisfied customers are more likely to leave feedback. This imbalance requires careful model evaluation beyond simple accuracy metrics.

Second, online reviews contain informal language features including emojis, internet slang, abbreviations, and mixed code-switching between languages. Vietnamese reviews may contain "k" replacing "không" (not), emoticons like ":(" expressing sadness, and English words mixed with Vietnamese text. These linguistic variations challenge standard NLP preprocessing pipelines.

Third, sentiment expression varies across cultures. Western reviews tend toward explicit sentiment expression, while Asian reviews may employ more indirect or contextual sentiment cues. A model performing well on English reviews may struggle with Vietnamese or Korean reviews requiring different cultural understanding.

2.4 Sustainable AI and Green Computing

The concept of sustainable AI or "Green AI" has emerged as an important consideration for responsible technology adoption [Schwartz et al. \(2020\)](#). Training large Transformer models consumes significant computational resources and energy, with environmental and economic costs. For example, training GPT-3 reportedly consumed 1,287 MWh of electricity, equivalent to approximately 552 tons of CO2 emissions [Strubell et al. \(2019\)](#).

This reality creates a dilemma for businesses: while larger models may offer better performance, their computational requirements may be impractical or environmentally irresponsible for many use cases. Model distillation approaches like DistilBERT [Sanh et al. \(2019\)](#) attempt to capture much of a large model's performance in a smaller, more efficient architecture. However, limited research examines whether these efficiency gains translate to acceptable performance in specific business domains like hospitality sentiment analysis.

2.5 Research Gap and Contribution

Our review identifies several critical gaps in existing literature. **First**, no comprehensive benchmark exists for multilingual sentiment analysis specifically in the hospitality domain using real-world data from emerging markets. **Second**, existing studies rarely examine model performance on highly imbalanced data reflecting actual business scenarios. **Third**, research has not adequately addressed the accuracy-efficiency trade-offs essential for practical business deployment decisions. **Fourth**, there is limited guidance translating technical model performance into actionable managerial recommendations.

This study addresses these gaps by providing a comprehensive benchmark of five state-of-the-art models on authentic multilingual hotel reviews from Vietnam, with explicit analysis of performance-cost trade-offs and detailed implications for hospitality business decision-making.

3 Methodology

3.1 Dataset Description

Our study utilizes a real-world dataset of hotel reviews collected from Vietnamese hospitality establishments. The dataset provides an authentic representation of the multilingual feedback environment faced by hotels in emerging tourism markets. The reviews were collected from major online travel platforms including booking websites and review aggregators. The dataset encompasses reviews from a diverse range of accommodation types including budget hotels, mid-range hotels, and luxury properties across Vietnam. This diversity ensures our findings are generalizable across different market segments within the hospitality industry.

Table 1 presents the key statistical characteristics of our dataset. The complete dataset contains 62,887 reviews before preprocessing. After data cleaning and filtering (removing missing values, filtering reviews between 10-500 tokens, and eliminating groups with fewer than 2 samples for stratified splitting), we obtained 59,377 high-quality reviews for model training and evaluation.

Table 1 Dataset statistics and characteristics

Characteristic	Before Preprocessing	After Preprocessing
Total reviews	62,887	59,377
Average tokens per review	186.1	149.4
Median tokens	121	115
Token range	10–5,598	10–500
Unique languages	14	14
Missing username values	463	444

The multilingual nature of our dataset reflects the international character of Vietnam’s tourism market. Table 2 shows the distribution of reviews across languages. English dominates with 54.6% of reviews, followed by Vietnamese at 25.9%. Other significant languages include French (7.5%), German (4.3%), Korean (3.8%), Russian (3.3%), and smaller percentages of Dutch, Japanese, Italian, Spanish, Chinese, Swedish, Polish, and Thai.

This distribution is particularly important for model evaluation, as it represents a realistic multilingual environment where models must handle both high-resource languages (English, French, German) and relatively lower-resource languages (Vietnamese) with comparable effectiveness.

A critical characteristic of our dataset is the highly imbalanced sentiment distribution shown in Table 3. This imbalance reflects real-world business scenarios where satisfied customers are more likely to leave reviews than dissatisfied ones. The "excellent" category dominates with 72.8% of reviews, while "terrible" and "negative" categories comprise only 2.8% and 2.8% respectively.

This severe class imbalance poses significant challenges for machine learning models, which tend to bias toward majority classes. However, from a business perspective,

Table 2 Language distribution in the dataset

Language	Number of Reviews	Percentage (%)
English (en)	32,449	54.6
Vietnamese (vi)	15,364	25.9
French (fr)	4,457	7.5
German (de)	2,577	4.3
Korean (ko)	2,282	3.8
Russian (ru)	1,944	3.3
Others	981	1.6
Total	59,377	100.0

Table 3 Sentiment category distribution

Sentiment Category	Number of Reviews	Percentage (%)	Rating Range
Excellent	43,241	72.8	5 stars
Positive	9,150	15.4	4 stars
Neutral	3,690	6.2	3 stars
Negative	1,683	2.8	2 stars
Terrible	1,660	2.8	1 star
Total	59,377	100.0	

accurately identifying negative and terrible reviews is critically important for service recovery and customer retention. A model achieving high overall accuracy by simply predicting "excellent" for most reviews would be practically useless for hospitality management.

3.2 Data Preprocessing

Our preprocessing pipeline was designed to clean the data while preserving sentiment-bearing features common in online reviews. We implemented minimal text cleaning to retain authentic review characteristics: removed reviews with missing text or sentiment labels, filtered reviews by length (10–500 tokens) to exclude extremely short or anomalously long reviews, retained emojis and emoticons as they carry important sentiment information, preserved informal language features including internet slang and abbreviations, and maintained mixed-language reviews as they naturally occur in multilingual environments.

To ensure fair evaluation across all sentiment-language combinations, we employed stratified sampling based on joint sentiment-language groups. We identified 68 unique sentiment-language combinations in the dataset. After removing groups with fewer than 2 samples (necessary for stratified splitting), 59,377 reviews remained. The data was split into training (80%), validation (10%), and test (10%) sets using stratified sampling to maintain sentiment and language distributions across all splits. Table 4 presents the final split sizes.

Table 4 Dataset split for model training and evaluation

Dataset Split	Number of Reviews	Percentage (%)
Training set	47,501	80.0
Validation set	5,938	10.0
Test set	5,938	10.0
Total	59,377	100.0

3.3 Model Selection

We selected five Transformer-based models representing different architectural approaches and design philosophies for multilingual NLP. Our selection criteria prioritized models with proven multilingual capabilities and diverse efficiency-accuracy profiles.

Multilingual BERT (mBERT): Google’s mBERT serves as our baseline model, pre-trained on Wikipedia text in 104 languages. With 110M parameters, it represents the first generation of massively multilingual models and provides a reference point for evaluating newer architectures.

XLM-RoBERTa (XLM-R): Facebook’s XLM-R extends the RoBERTa architecture to 100 languages with improved pre-training on 2.5TB of CommonCrawl data. As a widely-adopted model in multilingual NLP research and applications, XLM-R represents current industry standards for multilingual text understanding.

Multilingual DeBERTa (mDeBERTa): Microsoft’s mDeBERTa incorporates disentangled attention mechanisms that separately encode content and position information. This architectural innovation has achieved state-of-the-art performance on multiple multilingual benchmarks. We include mDeBERTa to represent the current frontier of multilingual model capabilities.

Multilingual DistilBERT: DistilBERT employs knowledge distillation to compress BERT’s knowledge into a model with 40% fewer parameters while retaining 97% of language understanding capabilities on benchmark tasks. With only 66M parameters compared to mBERT’s 110M, DistilBERT represents the efficiency-focused approach critical for resource-constrained business deployments.

Multilingual-E5: The E5 (EmbEddings from bidirectional Encoder rEpresentations) family represents a different paradigm, optimized for generating high-quality text embeddings. Originally designed for semantic search and retrieval tasks, we investigate whether E5’s embedding approach transfers effectively to sentiment classification, potentially offering advantages in semantic understanding of nuanced sentiment expressions.

3.4 Model Training Configuration

All models were fine-tuned using identical training configurations to ensure fair comparison. We employed the Hugging Face Transformers library with PyTorch as the underlying framework. Training hyperparameters were selected based on common practices in NLP fine-tuning: 3 epochs, batch size of 4 per device, learning rate of 2e-5

with linear warm-up over 500 steps, weight decay of 0.01 for regularization, maximum sequence length of 512 tokens, AdamW optimizer, and mixed precision training (BF16 where supported, otherwise FP16) for efficiency.

For each model, we loaded the pre-trained model weights from Hugging Face Hub, added a classification head with 5 output classes for sentiment categories, fine-tuned all model parameters on the training set, monitored validation accuracy after each epoch, selected the best checkpoint based on validation accuracy, and evaluated final performance on the held-out test set.

3.5 Evaluation Metrics

Given the severe class imbalance in our dataset, we employed multiple complementary evaluation metrics to provide a comprehensive assessment of model performance.

Overall accuracy measures the percentage of correctly classified reviews across all sentiment categories. While interpretable, accuracy can be misleading with imbalanced data as models may achieve high accuracy by simply predicting the majority class.

For each sentiment category, we computed precision (the proportion of predicted positives that are actually positive), recall (the proportion of actual positives that are correctly identified), and F1-score (the harmonic mean of precision and recall).

Macro-averaged metrics treat all classes equally regardless of size, computed as the unweighted mean across classes. Macro F1-score is particularly important for our imbalanced dataset, as it penalizes models that perform poorly on minority classes despite high overall accuracy.

Weighted-averaged metrics account for class imbalance by weighting each class’s contribution by its support (number of samples). These metrics reflect overall model utility across the actual distribution of reviews a business would encounter.

All experiments were conducted on Kaggle’s cloud computing platform with NVIDIA Tesla P100 GPU acceleration. The consistent experimental environment ensures reproducibility and fair comparison across models. We report results from the final test set evaluation, with no hyperparameter tuning on test data to avoid overfitting to evaluation metrics.

4 Empirical Results

This section presents the comprehensive evaluation results of five Transformer-based models on multilingual hotel review sentiment analysis. We analyze performance from multiple perspectives including overall accuracy, class-specific behavior, and the critical accuracy-efficiency trade-off for business deployment.

4.1 Overall Performance Comparison

Table 5 summarizes the overall performance of all five models on the test set. The results reveal significant performance variation across models, with multilingual-E5 achieving the highest accuracy (82%) and macro F1-score (0.62).

Several key findings emerge from the overall performance comparison:

Table 5 Overall model performance comparison on test set (n=5,938)

Model	Accuracy	Macro F1	Weighted F1	Parameters (M)
Multilingual-E5	0.82	0.62	0.81	278
mDeBERTa	0.82	0.59	0.81	279
mBERT	0.80	0.56	0.79	178
DistilBERT	0.80	0.55	0.79	135
XLM-RoBERTa	0.77	0.40	0.73	279

First, multilingual-E5 and mDeBERTa achieve identical accuracy (82%) but differ substantially in macro F1-score (0.62 vs. 0.59), indicating that multilingual-E5 better balances performance across sentiment categories including minority classes.

Second, mBERT and DistilBERT demonstrate nearly identical performance (80% accuracy), despite DistilBERT having 24% fewer parameters (135M vs. 178M). This finding is particularly significant from a sustainable AI perspective, suggesting that efficiency gains through model distillation do not necessarily sacrifice performance for this task.

Third, XLM-RoBERTa underperforms expectations with only 77% accuracy and a concerning macro F1-score of 0.40. This surprisingly weak performance from a model widely regarded as state-of-the-art for multilingual tasks warrants detailed investigation, which we provide in subsequent sections.

4.2 Class-Specific Performance Analysis

Table 6 presents detailed precision, recall, and F1-scores for each sentiment category across all models. This granular analysis reveals critical differences in model behavior that are obscured by aggregate metrics.

4.2.1 Excellent Category Performance

All models achieve strong performance on the "excellent" category, with precision and recall consistently around 0.90 and 0.94 respectively. This is expected given that "excellent" comprises 73% of the dataset. Even XLM-RoBERTa, despite its overall poor performance, achieves 0.83 precision and 0.96 recall on this category. The high recall across all models indicates that they successfully identify most excellent reviews, though with varying false positive rates.

4.2.2 Positive Category Performance

The "positive" category (15% of data) reveals more substantial performance differences. Multilingual-E5 and mDeBERTa achieve the highest F1-scores (0.48 and 0.49), while XLM-RoBERTa performs poorly with only 0.28 F1. The low recall of 0.20 for XLM-RoBERTa on positive reviews indicates systematic misclassification, likely confusing positive reviews with the dominant excellent category.

Table 6 Detailed classification performance by sentiment category

Model	Sentiment	Precision	Recall	F1	Support
E5	Excellent	0.90	0.94	0.92	4,325
	Positive	0.53	0.44	0.48	914
	Neutral	0.55	0.50	0.53	366
	Negative	0.46	0.43	0.45	167
	Terrible	0.71	0.73	0.72	166
mDeBERTa	Excellent	0.90	0.95	0.93	4,325
	Positive	0.55	0.44	0.49	914
	Neutral	0.52	0.45	0.48	366
	Negative	0.37	0.39	0.38	167
	Terrible	0.69	0.66	0.67	166
mBERT	Excellent	0.90	0.94	0.92	4,325
	Positive	0.51	0.42	0.46	914
	Neutral	0.50	0.49	0.50	366
	Negative	0.35	0.32	0.33	167
	Terrible	0.65	0.57	0.61	166
DistilBERT	Excellent	0.90	0.94	0.92	4,325
	Positive	0.47	0.41	0.44	914
	Neutral	0.48	0.42	0.45	366
	Negative	0.34	0.29	0.31	167
	Terrible	0.65	0.59	0.62	166
XLM-R	Excellent	0.83	0.96	0.89	4,325
	Positive	0.43	0.20	0.28	914
	Neutral	0.42	0.33	0.37	366
	Negative	0.56	0.03	0.06	167
	Terrible	0.42	0.42	0.42	166

4.2.3 Neutral Category Performance

For the "neutral" category (6% of data), multilingual-E5 achieves the best balance with F1-score of 0.53. All models struggle somewhat with neutral sentiment, likely because neutral language features overlap substantially with both positive and negative expressions, making this category inherently challenging to distinguish.

4.2.4 Negative Category: A Critical Weakness

The "negative" category (3% of data) represents one of the most important classes from a business perspective, as it indicates customer dissatisfaction requiring immediate attention. Here, model differences become stark and concerning:

- **Multilingual-E5:** Achieves the best performance with F1-score of 0.45, though still suboptimal
- **mDeBERTa:** Manages F1-score of 0.38

- **mBERT and DistilBERT:** Achieve F1-scores of 0.33 and 0.31 respectively
- **XLM-RoBERTa:** Catastrophically fails with F1-score of only 0.06, due to recall of just 0.03

XLM-RoBERTa’s 3% recall on negative reviews means it correctly identifies only 5 out of 167 negative reviews in the test set. This represents a complete failure for practical business use, as the model would miss 97% of negative feedback requiring managerial intervention.

4.2.5 Terrible Category: Service Recovery Critical

The "terrible" category (3% of data) represents extremely dissatisfied customers at high risk of negative word-of-mouth and social media complaints. Multilingual-E5 significantly outperforms other models with F1-score of 0.72, demonstrating much better identification of critically negative feedback. All other models achieve F1-scores between 0.42 and 0.67, with mDeBERTa in second place at 0.67.

The strong performance of multilingual-E5 on terrible reviews is particularly valuable for hospitality businesses, as these represent the highest-priority cases for service recovery efforts. Missing these reviews can result in public relations crises and permanent customer loss.

4.3 Confusion Matrix Analysis

To understand the systematic errors made by each model, we examine confusion patterns between sentiment categories. Figure ?? conceptually illustrates the key confusion patterns observed across models.

The most common confusion pattern across all models is misclassifying "positive" reviews as "excellent." This is understandable given the semantic similarity between these categories and the dominant prevalence of excellent reviews in the training data. However, the business implications differ: while both sentiments are favorable, "positive" (4-star) reviews often contain constructive criticism that "excellent" (5-star) reviews lack.

XLM-RoBERTa exhibits extreme confusion, classifying almost everything as "excellent." Its tendency to predict the majority class for nearly all inputs explains its high recall (0.96) but low precision (0.83) on excellent reviews, while achieving near-zero recall on minority classes like negative (0.03 recall).

Multilingual-E5 and mDeBERTa show more balanced confusion matrices, with errors distributed more evenly across categories. This balanced behavior contributes to their superior macro F1-scores despite having similar overall accuracy to other models.

4.4 The Accuracy-Efficiency Trade-off

From a practical business deployment perspective, model performance must be evaluated alongside computational efficiency. Table 7 presents a comparison of model size and expected inference characteristics.

DistilBERT offers the most attractive efficiency profile with only 135M parameters requiring approximately 540MB storage and 2.1GB inference memory. In comparison,

Table 7 Model efficiency comparison

Model	Parameters (M)	Model Size (MB)	Relative Speed	Memory (GB)
DistilBERT	135	540	1.0× (baseline)	2.1
mBERT	178	712	0.76×	2.8
XLM-RoBERTa	279	1116	0.61×	4.4
mDeBERTa	279	1116	0.61×	4.4
Multilingual-E5	278	1112	0.62×	4.3

Relative speed estimated based on parameter count and architectural complexity. Memory requirements approximate for batch size of 8 during inference.

multilingual-E5 requires 278M parameters, doubling the storage and memory footprint. However, for the 2% improvement in accuracy (82% vs. 80%) and substantial improvement in minority class handling (macro F1 of 0.62 vs. 0.55), many businesses may find the additional computational cost justified.

The efficiency comparison reveals an interesting finding: despite similar parameter counts, XLM-RoBERTa performs significantly worse than mDeBERTa and multilingual-E5. This suggests that architectural differences and pre-training approaches matter substantially beyond raw model size.

4.5 Performance Summary and Model Ranking

Based on comprehensive evaluation across multiple metrics, we establish the following model ranking for multilingual hospitality sentiment analysis:

Rank 1: Multilingual-E5 - Best overall performance with highest accuracy (82%), best macro F1 (0.62), and exceptional handling of minority classes. Recommended for businesses prioritizing accuracy and comprehensive customer feedback understanding.

Rank 2: mDeBERTa - Tied for highest accuracy (82%) with strong performance on excellent reviews, but weaker on minority classes than multilingual-E5. Suitable for businesses with adequate computational resources seeking high accuracy.

Rank 3: mBERT - Solid baseline performance (80% accuracy) with balanced behavior across classes. Reasonable choice when model maturity and community support are prioritized.

Rank 4: DistilBERT - Nearly identical performance to mBERT (80% accuracy) but with 24% fewer parameters. Best choice for resource-constrained deployments or high-volume inference scenarios requiring efficiency.

Rank 5: XLM-RoBERTa - Despite its reputation, performs poorly on this task with lowest accuracy (77%) and catastrophic failure on minority classes. Not recommended for hospitality sentiment analysis despite strong performance on standard NLP benchmarks.

5 Discussion

The empirical results reveal several important insights for both the research community and hospitality practitioners. This section discusses the implications of our findings, analyzes the reasons behind observed performance patterns, and examines the broader context of deploying AI systems in real-world business environments.

5.1 The Surprising Failure of XLM-RoBERTa

Perhaps the most striking finding is XLM-RoBERTa’s poor performance, particularly its catastrophic failure on minority classes. XLM-RoBERTa typically ranks among the top performers on standard multilingual NLP benchmarks such as XNLI, MLQA, and XQuAD. Several factors may explain this unexpected underperformance:

Training data distribution mismatch: XLM-RoBERTa was pre-trained on CommonCrawl web data, which may differ substantially in style and content from hotel reviews. The informal language, emojis, and mixed code-switching common in user-generated content may not align well with XLM-RoBERTa’s pre-training corpus.

Imbalanced data sensitivity: Our dataset’s severe imbalance (73% excellent reviews) may interact poorly with XLM-RoBERTa’s architecture or training procedure. The model appears to have collapsed toward predicting the majority class, suggesting vulnerability to imbalanced fine-tuning data.

Hyperparameter sensitivity: XLM-RoBERTa may require different hyperparameters than other models for optimal fine-tuning on imbalanced data. Our use of identical hyperparameters across all models, while ensuring fair comparison, may have disadvantaged XLM-RoBERTa specifically.

This finding carries an important lesson for practitioners: model performance on standard benchmarks does not guarantee performance on domain-specific tasks with different data characteristics. Organizations should conduct thorough evaluation on representative data before deploying models in production systems.

5.2 The Success of Multilingual-E5

Multilingual-E5’s superior performance, particularly on minority classes, suggests several advantages of embedding-based models for sentiment analysis:

Semantic space learning: E5’s training objective focuses on learning representations that capture semantic similarity in a continuous embedding space. This approach may better distinguish subtle sentiment differences than classification-focused training.

Balanced representation learning: The contrastive learning approach used in E5 training encourages the model to create well-separated representations for different semantic categories. This property may help maintain distinct representations for minority sentiment classes despite their low frequency in fine-tuning data.

Robustness to linguistic variation: E5’s strong multilingual performance across our diverse language set suggests its embeddings capture language-agnostic sentiment features effectively. This is particularly valuable in multilingual business contexts where review language cannot be controlled.

The 0.72 F1-score on terrible reviews achieved by multilingual-E5 represents a substantial practical improvement over other models. In absolute terms, this means

correctly identifying approximately 73% of extremely negative reviews compared to only 42% for XLM-RoBERTa, potentially preventing significant customer service failures.

5.3 The Accuracy-Efficiency Trade-off in Practice

The comparable performance of DistilBERT and mBERT despite DistilBERT’s 24% parameter reduction validates knowledge distillation as a viable strategy for efficient sentiment analysis. However, the business decision between efficiency-focused and accuracy-focused models requires careful cost-benefit analysis:

For high-volume, real-time applications: A hotel with thousands of daily reviews requiring immediate sentiment routing would benefit from DistilBERT’s faster inference. The 2% accuracy difference compared to multilingual-E5 may be acceptable given the processing speed advantage.

For comprehensive analysis systems: A hotel chain conducting weekly or monthly sentiment analysis for strategic decision-making can afford the computational cost of multilingual-E5. The improved minority class detection justifies the efficiency sacrifice for this use case.

For critical feedback detection: Systems specifically designed to identify and escalate negative reviews should prioritize the accuracy of negative and terrible category detection. Here, multilingual-E5’s superior performance on these categories becomes essential despite computational costs.

5.4 Handling Imbalanced Data in Business Contexts

The severe class imbalance in our dataset (73% excellent vs. 3% negative and terrible combined) reflects the reality of online review platforms but poses significant challenges for machine learning:

Macro vs. weighted F1 trade-off: All models achieve higher weighted F1 than macro F1, indicating better performance on frequent classes. However, from a business perspective, macro F1 better reflects utility since all sentiment categories require appropriate handling regardless of frequency.

Cost-sensitive learning: Future research should investigate whether cost-sensitive learning, which assigns higher misclassification costs to minority classes, can improve negative review detection. The business cost of missing a negative review (potential customer loss, negative publicity) far exceeds missing a positive review.

Synthetic oversampling considerations: While SMOTE and similar techniques could balance the training distribution, they risk generating unrealistic review text that doesn’t reflect genuine customer language. Domain-specific data augmentation techniques preserving authentic review characteristics would be valuable.

5.5 Multilingual Capability Implications

Our dataset’s linguistic diversity provides insights into models’ cross-lingual transfer abilities. While we did not separately evaluate performance per language due to insufficient samples for statistical significance in some languages, the overall results suggest:

High-resource language advantage persists: With 55% English reviews, models’ strong performance may partly reflect better English language understanding. Vietnamese, despite comprising 26% of reviews, is relatively low-resource in model pre-training.

Practical multilingual functionality: All models except XLM-RoBERTa demonstrate reasonable cross-lingual performance, correctly classifying reviews across languages without requiring language-specific models or ensemble approaches. This validates the practical utility of multilingual pre-training for business applications.

Code-switching robustness: Informal online reviews often mix languages within a single review. The models’ ability to handle mixed-language reviews without explicit code-switching detection represents an important practical advantage over traditional NLP pipelines.

5.6 Implications for Sustainable AI Adoption

From a sustainable AI perspective, our findings support a nuanced view of the efficiency-accuracy trade-off:

Task-appropriate sizing: DistilBERT’s strong performance suggests that many NLP tasks may not require massive models. The ”bigger is better” paradigm should be tempered by actual task requirements and cost considerations.

Deployment context matters: For edge deployment scenarios (mobile apps, IoT devices), DistilBERT’s smaller size enables functionality that would be impossible with larger models. The 2% accuracy difference becomes irrelevant if the alternative is not deploying any model at all.

Environmental considerations: Training and deploying smaller models reduces energy consumption and carbon emissions. For sentiment analysis specifically, DistilBERT offers a ”green” option with minimal performance sacrifice.

However, our results also show that blanket efficiency preferences can be counterproductive. XLM-RoBERTa and mDeBERTa have identical parameter counts, yet mDeBERTa significantly outperforms XLM-RoBERTa. Similarly, the 2% accuracy gain and substantial minority class improvement of multilingual-E5 over DistilBERT may easily justify the additional computational cost for many applications.

5.7 Limitations of Current Study

Our research has several limitations that suggest directions for future investigation:

Single domain focus: Results are specific to hotel reviews in Vietnam. Generalization to other hospitality subsectors (restaurants, attractions) or other countries requires validation.

Inference time not measured: We report model size and parameter counts but did not measure actual inference latency in production-like conditions. Real-world deployment decisions require concrete timing measurements.

Language-specific analysis limited: Small sample sizes for some languages prevented statistically rigorous language-specific analysis. A larger-scale study could reveal whether certain languages pose systematic challenges.

Temporal dynamics: Review sentiment patterns and language may evolve over time. Our static dataset does not capture temporal trends or seasonal variations in customer feedback.

Fine-tuning optimization: We used identical hyperparameters for all models. Model-specific hyperparameter optimization might improve individual model performance, though it would compromise fair comparison.

6 Managerial Implications

The empirical findings provide actionable guidance for hospitality businesses implementing automated sentiment analysis systems. We structure our recommendations around the key decision factors of business scale, operational context, and resource constraints.

6.1 Model Selection Framework

Table 8 presents model recommendations segmented by business type. The selection framework considers three critical dimensions: (1) computational resource availability, (2) accuracy requirements for minority class detection, and (3) review processing volume. Budget-constrained operations with high review volumes benefit from DistilBERT’s efficiency (80% accuracy, 135M parameters), while luxury establishments requiring comprehensive negative feedback detection should deploy multilingual-E5 despite higher computational costs (82% accuracy, 278M parameters, F1=0.72 on terrible reviews).

Table 8 Model recommendations by hospitality business segment

Business Segment	Characteristics	Model	Rationale
Budget Hotels & Hostels	Low margins, limited IT resources, high volume	DistilBERT	Lowest cost, 80% accuracy, fast inference, easy deployment
Mid-Range Hotels	Moderate budgets, balanced needs, moderate volume	mBERT or DistilBERT	Cost-performance balance, proven stability, adequate minority class handling
Luxury Hotels & Resorts	High margins, premium service, critical feedback	Multilingual-E5	Best accuracy (82%), superior negative detection (F1=0.72 on terrible)
Hotel Chains & Groups	Centralized analytics, multiple properties, strategic needs	Multilingual-E5 or mDeBERTa	Comprehensive analysis, cross-property insights, long-term value

6.2 Application-Specific Deployment Considerations

Different operational contexts require distinct model characteristics. **Real-time alert systems** monitoring negative feedback must prioritize recall on minority classes; multilingual-E5 achieves 0.43 recall on negative reviews compared to XLM-RoBERTa’s catastrophic 0.03 recall, making model selection critical for service recovery effectiveness. **Batch analytics systems** conducting periodic strategic analysis can accommodate less efficient models, as processing time becomes secondary to insight quality. **Automated routing systems** require balanced performance across categories to prevent systematic misdirection of specific sentiment types.

The XLM-RoBERTa failure case demonstrates that standard benchmark performance does not guarantee domain-specific effectiveness. Organizations must validate models on representative data reflecting actual imbalanced class distributions and linguistic characteristics before production deployment.

6.3 Cost-Benefit Analysis

We provide a quantitative framework for evaluating sentiment analysis investment. Consider a mid-range hotel processing 1,000 monthly reviews. Quantifiable benefits include: (1) labor savings from automated analysis (\$10,400 annually assuming 10 hours weekly manual review time at \$20/hour), (2) improved service recovery reducing negative review escalation (estimated \$5,000 annually), and (3) rating improvements driving conversion increases (1% revenue increase on \$2M annual revenue = \$20,000). Total annual benefits approximate \$35,400.

Implementation costs comprise: (1) cloud infrastructure (\$900 annually for DistilBERT, \$1,800 for multilingual-E5), (2) initial development (\$8,000 amortized over three years = \$2,667 annually), and (3) ongoing maintenance (\$800 annually). For DistilBERT deployment, total annual costs equal \$4,367, yielding 710% ROI. The marginal cost of upgrading to multilingual-E5 (\$900 additional infrastructure) may be justified by improved negative review detection preventing customer loss.

6.4 Implementation Strategy

We recommend a phased approach: **Phase 1 (Months 1-3)** deploys an efficient baseline model (DistilBERT or mBERT) to establish operational workflows and staff familiarity. **Phase 2 (Months 4-6)** integrates sentiment classification into daily operations with automated alerting for negative reviews. **Phase 3 (Months 7-9)** evaluates model adequacy through business metrics (response time, resolution rate) and considers upgrading to higher-performance models if minority class detection proves insufficient. **Phase 4 (Months 10-12)** extends capabilities toward aspect-based sentiment analysis and predictive guest satisfaction modeling.

Organizations must address change management alongside technical deployment. Staff require training in interpreting model outputs and understanding accuracy limitations. Performance metrics should incentivize appropriate use of automated insights rather than creating incentives to circumvent systems. Ethical considerations

include data privacy compliance (GDPR, CCPA), fairness auditing across demographic groups, and maintaining human oversight for negative feedback requiring empathetic responses.

7 Conclusion

This study provides a comprehensive benchmark of five state-of-the-art Transformer models for multilingual sentiment analysis in hospitality using 59,377 authentic hotel reviews from Vietnam spanning 14 languages. Our findings reveal that multilingual-E5 achieves superior overall performance (82% accuracy, macro F1=0.62) with particularly strong minority class detection (F1=0.72 on terrible reviews), making it optimal for businesses prioritizing comprehensive customer feedback analysis. DistilBERT offers comparable accuracy (80%) with 51% fewer parameters, validating knowledge distillation for resource-constrained deployments. Notably, XLM-RoBERTa fails catastrophically on imbalanced data (77% accuracy, 0.03 recall on negative reviews), demonstrating that benchmark performance does not guarantee domain-specific effectiveness. These findings provide evidence-based guidance linking business characteristics to appropriate model selection while highlighting the critical importance of evaluating models on representative data with realistic class distributions.

Our research contributes theoretically by establishing the first comprehensive multilingual sentiment analysis benchmark for hospitality using authentic business data, demonstrating the inadequacy of standard benchmarks for predicting domain-specific performance, and validating macro F1-score as essential for evaluating imbalanced business classification tasks. Practically, we provide hospitality practitioners with actionable model selection frameworks, quantitative ROI analysis (710% return for typical mid-range hotels), and phased implementation strategies. The study supports UN Sustainable Development Goals 8, 9, and 12 by democratizing AI access for SMEs through efficient models, facilitating digital transformation in hospitality, and promoting sustainable computational practices. However, limitations include geographic focus on Vietnam, temporal constraints of static evaluation, and lack of inference latency measurements across deployment environments.

Future research should extend this work toward aspect-based sentiment analysis identifying specific service attributes, temporal dynamics predicting service quality trends, cross-lingual transfer learning for low-resource languages, multimodal analysis integrating guest photos with text, and fairness investigations across demographic groups. The integration of generative AI for automated response generation and development of continuous performance monitoring systems addressing distribution shift represent particularly promising directions. As emerging markets expand as tourism destinations, the ability to understand and respond to multilingual customer feedback at scale will increasingly differentiate successful hospitality businesses, making the selection and deployment of appropriate sentiment analysis models a strategic imperative for competitive advantage.

Acknowledgements. This study is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City.

Declarations

Funding: This study is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City.

Conflict of interest: The author declares no competing interests.

Ethics approval and consent to participate: Not applicable. This study analyzes publicly available online reviews without personal identification information.

Consent for publication: Not applicable.

Data availability: The dataset is publicly available in the Zenodo repository at <https://doi.org/10.5281/zenodo.18350688>.

Code availability: The source code used for model training and analysis is available from the corresponding author upon reasonable request.

Author contribution: Q.L.N. conceived the study, conducted the experiments, analyzed the results, and wrote the manuscript.

References

- Alaei AR, Becken S, Stantic B (2019) Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research* 58(2):175–191. <https://doi.org/10.1177/0047287517747753>
- Anderson CK, Xie X (2010) Improving hospitality industry sales: Twenty-five years of revenue management. *Cornell Hospitality Quarterly* 51(1):53–67. <https://doi.org/10.1177/1938965509354697>
- Conneau A, Khandelwal K, Goyal N, et al (2020) Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp 8440–8451, <https://doi.org/10.18653/v1/2020.acl-main.747>
- Devlin J, Chang MW, Lee K, et al (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*. Association for Computational Linguistics, pp 4171–4186, <https://doi.org/10.18653/v1/N19-1423>
- He P, Liu X, Gao J, et al (2021) DeBERTa: Decoding-enhanced BERT with disentangled attention. In: *Proceedings of the 9th International Conference on Learning Representations*, URL <https://openreview.net/forum?id=XPZlaotutsD>
- Kirilenko AP, Stepchenkova SO, Kim H, et al (2018) Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research* 57(8):1012–1025. <https://doi.org/10.1177/0047287517729757>
- Litvin SW, Goldsmith RE, Pan B (2008) Electronic word-of-mouth in hospitality and tourism management. *Tourism Management* 29(3):458–468. <https://doi.org/10.1016/j.tourman.2007.05.011>

- Liu B (2010) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- Mariani MM, Borghi M (2021) Customers' evaluation of mechanical artificial intelligence in hospitality services: A study using online reviews analytics. *Journal of Hospitality & Tourism Research* 45(7):1240–1263. <https://doi.org/10.1108/IJCHM-06-2020-0622>
- Ruder S, Vulić I, Søgaard A (2019) A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65:569–631. <https://doi.org/10.1613/jair.1.11640>
- Sanh V, Debut L, Chaumond J, et al (2019) DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 URL <https://arxiv.org/abs/1910.01108>
- Schwartz R, Dodge J, Smith NA, et al (2020) Green AI. *Communications of the ACM* 63(12):54–63. <https://doi.org/10.1145/3381831>
- Socher R, Perelygin A, Wu J, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp 1631–1642, URL <https://aclanthology.org/D13-1170/>
- Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243 URL <https://arxiv.org/abs/1906.02243>
- Wang L, Yang N, Huang X, et al (2022) Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 URL <https://arxiv.org/abs/2212.03533>
- Xiang Z, Schwartz Z, Gerdes Jr JH, et al (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* 44:120–130. <https://doi.org/10.1016/j.ijhm.2014.10.013>
- Xiang Z, Du Q, Ma Y, et al (2017) A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management* 58:51–65. <https://doi.org/10.1016/j.tourman.2016.10.001>