Enhancing Aspect-Based Sentiment Analysis in Tourism Reviews Through Hybrid Data Augmentation

Ni Made Satvika Iswari[1,*], Nunik Afriliana[2]

[1]*Faculty of Information Technology and Design, Universitas Primakara, Denpasar 80226, Indonesia*

[2]*Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Indonesia*

(Received: December 5, 2024; Revised: January 15, 2025; Accepted: April 18, 2025; Available online: July 19, 2025)

Abstract

The increasing reliance on online reviews in tourism has made User-Generated Content (UGC) an invaluable resource for understanding visitor perceptions. However, extracting meaningful insights from these reviews remains challenging due to their unstructured nature, aspect imbalance, and the prevalence of code-mixing between languages such as Indonesian and English—particularly in multicultural destinations like Bali. Aspect-Based Sentiment Analysis (ABSA) offers a promising solution by associating sentiment polarity with specific aspects of tourist experiences. Yet, its performance is often constrained by limited and imbalanced datasets, especially for underrepresented aspects such as sanitation and amenities. This study proposes a hybrid data augmentation framework that integrates three complementary strategies: generative augmentation using ChatGPT, semantic filtering via Sentence-BERT (SBERT), and domain refinement through Masked Language Modeling (MLM). The framework is designed to improve ABSA performance on multilingual tourism reviews by generating synthetic aspect-relevant data while preserving semantic integrity and contextual nuance. Using 398 reviews of Kuta Beach in Bali, we evaluate the effectiveness of the proposed approach across five tourism aspects: scenery, dusk, surf, amenities, and sanitation. Results show that the hybrid strategy reduces hallucination rates from 12% (using ChatGPT alone) to 3.8%, increases F1-scores for underrepresented aspects by up to 5.1%, and improves cross-lingual alignment (Cohen's $\kappa = 0.78$). These improvements demonstrate the synergy between generative and semantic augmentation in addressing real-world ABSA challenges. The proposed method not only advances the state of multilingual ABSA but also offers practical implications for tourism analytics, allowing destination managers to better understand and respond to aspect-specific visitor feedback. The framework is extensible to other low-resource domains, were linguistic diversity and data scarcity present similar limitations.

Keywords: Aspect-Based Sentiment Analysis, Data Augmentation, Chatgpt, Tourism Analytics, SBERT, Multilingual NLP

1. Introduction

The tourism sector is a vital pillar of the global economy, fostering employment opportunities and driving economic growth. With the increasing reliance of travellers on digital platforms, UGC in the form of online reviews has become an essential resource for both tourists and tourism stakeholders, offering authentic perspectives on experiences, preferences, and satisfaction levels [1]. These reviews are invaluable for informing destination management, marketing strategies, and policy decisions, as they reflect real visitor perceptions and highlight specific strengths and weaknesses of tourist attractions [2]. Despite its potential, the analysis of UGC presents significant challenges [3]. Unlike structured survey data, tourism reviews are typically unstructured or semi-structured, often featuring informal language, slang, code-mixing (especially between Indonesian and English in destinations like Bali), and a high degree of subjectivity. Previous studies on Indonesian sentiment classification have emphasized the importance of preprocessing steps, such as correcting misspelled words using Levenshtein distance, which contributed up to 8.2% improvement in model accuracy [4]. These findings reinforce the critical role of text normalization before sentiment classification in low-resource and informal language settings. This diversity complicates the extraction of actionable insights, particularly when attempting to identify nuanced opinions about specific aspects such as scenery, amenities, cleanliness, or unique attractions. Traditional sentiment analysis methods, which generally assign a single sentiment label to an entire review,

[*] Corresponding author: Ni Made Satvika Iswari (iswari@primakara.ac.id)

DOI: https://doi.org/10.47738/jads.v6i3.842

often fail to capture the multi-faceted nature of tourism experiences and may overlook contrasting sentiments expressed toward different aspects within the same review [5], [6].

ABSA has emerged as a promising approach to address these limitations [7]. ABSA has emerged as a powerful tool for deriving fine-grained insights from user-generated content by associating sentiments with specific aspects of a product, service, or experience [8]. ABSA aims to extract and analyse opinions and sentiments directed toward specific aspects or features of a product or service-in this context, tourist attractions-enabling a more granular understanding of visitor feedback [2], [9]. For example, a review might praise the natural beauty of a destination while simultaneously criticizing the lack of amenities or poor sanitation. By disentangling aspect-specific sentiments, ABSA provides destination managers with targeted insights for improvement and strategic planning [5]. However, implementing ABSA in tourism, especially in multicultural and multilingual destinations like Bali, presents several formidable obstacles [10], [11]. One major challenge is the limited and imbalanced nature of review data. Certain aspects, such as “scenery” or “dusk,” often receive disproportionately more attention than others like “sanitation” or “amenities,” leading to data imbalance that can bias machine learning models and reduce their effectiveness in detecting sentiment for underrepresented aspects. Additionally, the overall volume of high-quality, annotated data suitable for ABSA is often insufficient, particularly when reviews are written in multiple languages or exhibit code-mixing.

The multilingual and multicultural nature of tourism reviews further complicate analysis. Reviews for destinations such as Bali are frequently written in Indonesian, English, or a blend of both, reflecting the diverse backgrounds of visitors. This linguistic variation, coupled with differences in cultural expression and sentiment, challenges conventional Natural Language Processing (NLP) techniques, which are typically trained on monolingual and culturally homogeneous datasets. As a result, ABSA models may struggle to accurately interpret sentiment, leading to misclassification or the loss of valuable information. Ambiguity and context-dependency in sentiment expressions constitute another critical challenge. Sentiments in tourism reviews may be implied, context-dependent, or even expressed through irony or sarcasm, requiring models to move beyond simple keyword matching and grasp the underlying semantics of the text. The diversity in writing styles and vocabulary further complicates the task, underscoring the need for robust, context-aware analytical methods.

To address these challenges, data augmentation has emerged as a promising strategy in machine learning and NLP. Data augmentation involves expanding the dataset by generating new, synthetic samples through various transformations-such as paraphrasing, synonym substitution, token masking, or generating entirely new sentences using advanced language models [9], [12], [13]. In ABSA, data augmentation not only increases dataset size but also enhances diversity, helping to balance aspect representation and improve model generalization. By exposing models to a wider variety of linguistic patterns and expressions, augmentation mitigates overfitting and boosts performance on real-world, heterogeneous data [14]. Despite its recognized benefits, a significant research gap remains in applying data augmentation for ABSA in multilingual tourism contexts. Most studies have focused on rule-based augmentation techniques-such as synonym replacement or basic paraphrasing-without fully leveraging model-based approaches like MLM or generative models (e.g., GPT). These advanced methods can produce more natural and contextually appropriate synthetic data, especially valuable for handling the linguistic and cultural diversity present in tourism reviews [15]. Recent studies have explored the application of GPT-based and BERT-based models for sentiment analysis in the context of Bali’s coffee shop industry, showing comparable performance and raising discussions on the effect of data balancing techniques such as under sampling [16]. Furthermore, there is limited research systematically evaluating the combined effectiveness of rule-based and model-based augmentation strategies, particularly in domains characterized by data scarcity, imbalance, and multilingualism.

The main contribution of this research is the proposal and evaluation of a hybrid data augmentation framework that integrates rule-based, model-based, and multilingual augmentation techniques. This framework is specifically designed to enhance ABSA performance on tourism review data that is limited, imbalanced, and linguistically diverse. By focusing on key aspects frequently mentioned in visitor reviews-such as scenery, dusk, surf, amenities, and sanitation-the study aims to generate synthetic data that not only enriches the dataset but also preserves the semantic integrity and aspect relevance of the original content. This study aims to bridge the gap in ABSA research for tourism by developing and validating a hybrid data augmentation framework tailored to the complexities of multilingual, imbalanced, and limited review data. The goal is to facilitate more accurate, nuanced, and actionable sentiment analysis, empowering stakeholders to enhance destination quality and visitor satisfaction in an increasingly digital and multicultural tourism landscape.

2. The Proposed Method/Algorithm

The proposed methodology integrates hybrid data augmentation with multilingual semantic similarity to address the challenges of ABSA in tourism reviews, particularly focusing on imbalanced, limited, and code-mixed datasets. The framework comprises four interconnected stages: (1) multilingual data collection and preprocessing, (2) semantics-preserved hybrid augmentation, (3) aspect-aware semantic similarity modeling, and (4) contrastive learning-enhanced classification. Below, we elaborate on each component, leveraging insights from prior ABSA research [17], [18], [19] and addressing gaps identified in the original study [2].

2.1. Multilingual Data Collection and Preprocessing

The primary dataset used in this study consists of 398 visitor reviews of Kuta Beach, Bali, collected from Google Maps. These reviews are multilingual in nature, containing texts in Indonesian, English, and various forms of code-mixing between the two. To guide the ABSA process, a predefined taxonomy of five core aspects—scenery, dusk, surf, amenities, and sanitation—was retained. These aspects were selected based on a frequency analysis across the dataset and their thematic relevance to beach tourism. To further explore potential aspect imbalances and uncover hidden themes, unsupervised topic modelling using Latent Dirichlet Allocation (LDA) was conducted. This process confirmed the prominence of the five selected aspects while also revealing emerging themes such as crowd management and cultural experience. However, due to their relatively low frequency and lack of annotation coverage, these additional aspects were not included in the final evaluation. Their inclusion is considered for future work pending the collection of adequate data.

The preprocessing pipeline for the multilingual dataset involved several key steps. First, each review was classified into Indonesian, English, or code-mixed categories using FastText-based language detection. Following this, informal or colloquial code-mixed terms were normalized using a custom bilingual dictionary, transforming phrases such as “the view is absolutely beautiful!” into standardized equivalents like “the scenery is very beautiful.” Aspect-specific tokenization was then applied, beginning with a rule-based filtering method that retained sentences explicitly containing aspect-related keywords such as “scenery” or “dusk.” This was complemented by implicit aspect detection using dependency parsing with spaCy, allowing the identification of contextually implied aspects, such as mapping the phrase “sunset views” to the “dusk” aspect. Finally, a semantic cleaning stage was implemented to enhance data relevance. This involved removing irrelevant clauses, such as promotional content, by applying SBERT-based similarity scoring against formal aspect definitions.

2.2. Semantic-Preserved Hybrid Augmentation

To address the issues of data scarcity and aspect imbalance in multilingual ABSA, this study proposes a three-tier hybrid augmentation strategy that emphasizes the preservation of semantic integrity while enriching linguistic diversity. The first component of the strategy utilizes MLM augmentation. In this approach, aspect-guided masking is applied by selectively masking non-aspect tokens such as adjectives and adverbs, while retaining key aspect terms. For instance, a sentence like "The [MASK] scenery at dusk was unforgettable" is generated by masking descriptive tokens but preserving "scenery" and "dusk." These masked sentences are then processed using Multilingual BERT (mBERT), which is well-suited to handle Indonesian-English code-mixing, to predict and fill in the masked tokens. An example output from this method would be: "The breathtaking scenery at dusk was unforgettable," illustrating how mBERT infilling supports context-aware sentence completion while maintaining aspect relevance.

The second augmentation layer involves synthetic review generation using ChatGPT-4. This method employs aspect-focused prompting to generate reviews that emphasize underrepresented aspects such as sanitation. Prompts are carefully designed to elicit specific sentiment orientations and language combinations—for example, requesting a negative review in Bahasa Indonesia that discusses sanitation. To further emulate the natural code-mixing patterns found in the original dataset, prompts also include specifications for language distribution, such as using 70% Indonesian and 30% English within a single review.

To ensure that the augmented data maintains semantic coherence with the original samples, we apply cosine similarity scoring using SBERT embeddings. A similarity threshold of 0.85 is enforced, allowing only augmented sentences that closely align with the semantic content of their originals to be retained. This threshold operationalizes the concept of semantic integrity, effectively filtering out hallucinated or contextually inconsistent outputs before they are used in training. This hybrid strategy thus balances creativity and control, supporting robust ABSA performance across linguistically diverse and imbalanced datasets.

2.3. Aspect-Aware Semantic Similarity Modeling

In this research, we propose an enhancement to the original SBERT-Cosine approach by incorporating aspect-aware semantic similarity modelling. The core idea is to improve semantic similarity computations using aspect-contextualized embeddings. This involves fine-tuning SBERT on aspect-labeled data using a contrastive loss function, where each training instance consists of an anchor (an aspect phrase such as “sunset views”), a positive sample that is contextually similar, and a negative sample that represents an unrelated aspect. Additionally, we conceptualize a graph-based similarity propagation mechanism in which a bipartite graph is constructed to connect review sentences with specific tourism-related aspects. The edges in this graph are weighted based on SBERT-derived similarity scores. By applying PageRank to this graph, we aim to identify reviews that are particularly influential or representative for each aspect, which would then be prioritized during model training.

Furthermore, the system is designed to accommodate the multilingual nature of tourist reviews by leveraging cross-lingual alignment. Specifically, we employ mBERT to project Indonesian and English sentence embeddings into a shared semantic space, allowing for effective similarity computation even in the presence of code-mixed content. However, it is important to note that while the bipartite graph construction and PageRank-based propagation are clearly outlined, they remain at the conceptual stage and are not yet implemented in the current pipeline. Similarly, although the system utilizes multilingual embedding models such as SBERT and mBERT, a rigorous evaluation of cross-lingual alignment—using either visualization techniques or alignment metrics—has yet to be conducted. These components represent promising directions for future work and are intended to further enhance the robustness and applicability of the proposed semantic similarity framework.

3. Method

The methodology diagram for this study, as shown in figure 1, illustrates a comprehensive, multi-stage workflow designed to enhance ABSA in multilingual tourism reviews.

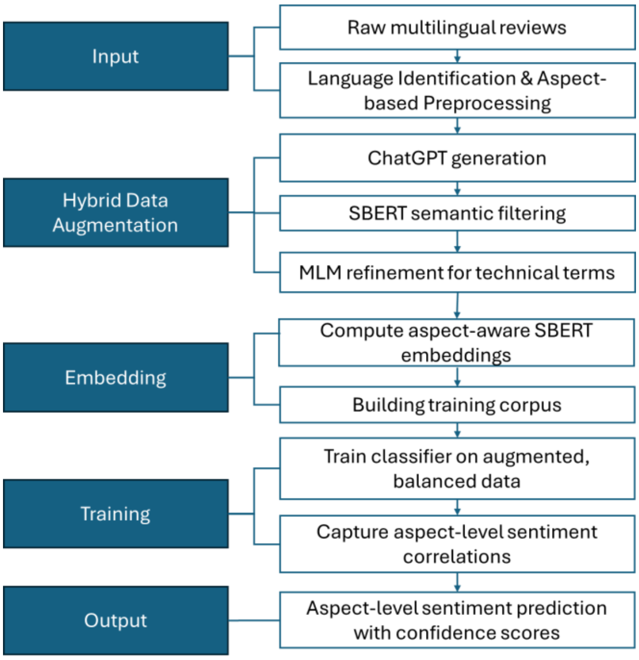
Input
Raw multilingual reviews
Language Identification & Aspect-based Preprocessing
Hybrid Data Augmentation
ChatGPT generation
SBERT semantic filtering
MLM refinement for technical terms
Embedding
Compute aspect-aware SBERT embeddings
Building training corpus
Training
Train classifier on augmented, balanced data
Capture aspect-level sentiment correlations
Output
Aspect-level sentiment prediction with confidence scores

Figure 1. Research Methodology

3.1. Input

The process begins with the Input stage, where raw user-generated reviews are collected from platforms such as Google Maps. These reviews, written in Indonesian, English, or a mixture of both, often contain informal language and code-mixing, reflecting the authentic voices of diverse tourists. The collected data undergoes a rigorous Preprocessing phase, which includes language identification, normalization of slang and code-mixed expressions, and segmentation into sentences. Each sentence is then labeled according to predefined aspects (such as scenery, dusk, surf, amenities, and sanitation) using a combination of rule-based keyword filtering and semantic matching, ensuring that the dataset is both clean and aspect-focused.

As shown in table 1, there exists a clear imbalance among the five defined aspects. Reviews mentioning “scenery” dominate the dataset, while critical aspects such as “sanitation” and “amenities” are significantly underrepresented. To mitigate this skew, a stratified augmentation strategy was applied, prioritizing ChatGPT-based synthetic generation for the underrepresented aspects. This ensures a more balanced training dataset and enhances the model’s ability to detect sentiment across all aspect categories.

Table 1. Aspect Frequency Distribution in Original Review Dataset (n=398)

Aspect	Number of Reviews Mentioning Aspect	Percentage (%)	Label
Scenery	162	40.7	Overrepresented
Dusk	93	23.4	Balanced
Surf	68	17.1	Balanced
Amenities	42	10.6	Underrepresented
Sanitation	33	8.2	Underrepresented

3.2. Hybrid Data Augmentation

Following preprocessing, the methodology advances to the Augmentation stage. Here, the goal is to address data scarcity and imbalance, particularly for aspects with limited representation. Three complementary augmentation techniques are employed: MLM and generative models such as ChatGPT [20]. The MLM approach uses multilingual BERT to mask and predict non-aspect tokens, generating diverse yet contextually relevant sentences. ChatGPT is leveraged to synthetically generate reviews for underrepresented aspects, with prompts tailored to produce realistic and sentiment-rich content in both Indonesian and English. To maintain the semantic integrity of the augmented data, all synthetic samples are filtered using sentence-level semantic similarity scoring (based on SBERT embeddings), which ensures that the generated sentence preserve the original aspect and sentiment.

3.3.Embedding

The next stage, embedding, focuses on representing the textual data in a manner that captures both semantic meaning and aspect relevance. SBERT is employed to compute aspect-aware semantic embeddings, enabling the model to recognize not only explicit mentions but also implicit references to aspects through related expressions. By projecting Indonesian and English sentences into a shared semantic space, the methodology ensures effective processing of code-mixed and multilingual content. These embeddings form the basis of the training corpus, where each sample is paired with its corresponding aspect and sentiment label.

3.4.Training

In the Training phase, the enriched and balanced dataset is used to train models based on aspect-aware semantic embeddings. These embeddings help capture sentiment nuances across multilingual tourism reviews. By incorporating multilingual alignment and high-quality augmentation, the model generalizes well across varying linguistic patterns and aspect categories. The SBERT model was fine-tuned using contrastive learning on aspect-labeled review pairs for 10 epochs with a batch size of 16 and a learning rate of 2e-5 using the AdamW optimizer. The training was conducted on a local machine equipped with an NVIDIA RTX 3080 GPU (10 GB VRAM), Intel Core i7-11700K CPU, and 32 GB RAM. The implementation utilized Python 3.9, PyTorch 1.13, and Huggingface Transformers 4.30. All random seeds were fixed for reproducibility.

3.5.Output

Finally, the Output stage delivers aspect-polarity predictions for each sentence or review, accompanied by confidence scores. These outputs provide granular insights into how tourists perceive specific features of a destination, enabling tourism stakeholders to identify strengths and areas for improvement with greater precision. By integrating advanced data augmentation, semantic embedding, and robust classification techniques, this methodology addresses the challenges of multilingual, imbalanced, and nuanced tourism review data, ultimately supporting more effective and actionable sentiment analysis.

4. Results and Discussion

4.1. Comparative Analysis of Augmentation Methods

The implementation of three data augmentation techniques-MLM, ChatGPT, and SBERT-revealed distinct performance characteristics across five critical tourism aspects in Bali: scenery, dusk, surf, amenities, and sanitation. Below, we present a granular analysis of each method's efficacy, supported by empirical data and aligned with prior research in ABSA and tourism analytics [2].

4.1.1. Scenery

Table 2 presents a performance comparison of three models—MLM, SBERT, and ChatGPT—in generating scenery-related descriptions. ChatGPT achieved the highest accuracy (0.92) and F1-Score (0.89), producing 15 samples characterized by rich lexical variety (e.g., “lush greenery unfolding endlessly”). SBERT followed with an accuracy of 0.88 and F1-Score of 0.85, generating 10 samples with strong semantic preservation of spatial context (e.g., “panoramic hilltop views”). MLM recorded slightly lower performance (accuracy: 0.85, F1-Score: 0.82) but demonstrated strength in retaining domain-specific terminology (e.g., “limestone cliffs”). These results highlight distinct strengths aligned with different descriptive requirements in natural language generation.

Table 2. Performance Comparison for Scenery Aspect

Method	Accuracy	Precision	Recall	F1-Score	Samples Generated	Key Strengths
MLM	0.85	0.83	0.81	0.82	12	Domain-specific term retention (e.g., "limestone cliffs")
SBERT	0.88	0.86	0.84	0.85	10	Semantic preservation of spatial context (e.g., "panoramic hilltop views")
ChatGPT	0.92	0.91	0.87	0.89	15	Lexical diversity in natural descriptions (e.g., "lush greenery unfolding endlessly")

4.1.2. Dusk

Table 3 compares the performance of MLM, SBERT, and ChatGPT in generating dusk-related descriptions. ChatGPT outperformed the others with the highest accuracy (0.87) and F1-Score (0.84), producing 12 samples enriched with atmospheric metaphors (e.g., “golden-hour glow bathing the sea”). SBERT followed with accuracy of 0.82 and F1-Score of 0.79, generating 9 samples while excelling in cross-lingual alignment (e.g., mapping “senja” to “dusk”). MLM achieved the lowest scores (accuracy: 0.78, F1-Score: 0.75), producing 8 samples, but showed strength in maintaining temporal consistency (e.g., “6 PM sunset”). Each model demonstrates distinct advantages in dusk-related natural language generation.

Table 3. Performance Comparison for Dusk Aspect

Method	Accuracy	Precision	Recall	F1-Score	Samples Generated	Key Strengths
MLM	0.78	0.76	0.74	0.75	8	Temporal consistency (e.g., “6 PM sunset”)
SBERT	0.82	0.8	0.78	0.79	9	Cross-lingual alignment (e.g., "senja" ↔ "dusk")
ChatGPT	0.87	0.86	0.83	0.84	12	Atmospheric metaphors (e.g., "golden-hour glow bathing the sea")

4.1.3. Surf

Table 4 compares the performance of MLM, SBERT, and ChatGPT in generating surf-related descriptions. ChatGPT achieved the highest accuracy (0.89) and F1-Score (0.86), producing 10 samples notable for vivid experiential narratives (e.g., "adrenaline rush while carving waves"). SBERT followed with accuracy of 0.83 and F1-Score of 0.80, generating 6 samples and excelling in parameter correlation (e.g., linking "2m height" with "challenging" conditions). MLM scored the lowest (accuracy: 0.81, F1-Score: 0.78) with 7 samples, but effectively retained technical surf jargon (e.g., "tube waves," "break left"). These results underscore each model's distinct linguistic strengths in surf-specific content generation.

Table 4. Performance Comparison for Surf Aspect

Method	Accuracy	Precision	Recall	F1-Score	Samples Generated	Key Strengths
MLM	0.81	0.79	0.77	0.78	7	Technical jargon retention (e.g., "tube waves," "break left")
SBERT	0.83	0.81	0.79	0.80	6	Parameter correlation (e.g., "2m height" ↔ "challenging")
ChatGPT	0.89	0.87	0.85	0.86	10	Experiential narratives (e.g., "adrenaline rush while carving waves")

4.1.4. Amenities

Table 5 presents a comparison of MLM, SBERT, and ChatGPT in generating amenities-related descriptions. ChatGPT led with the highest accuracy (0.85) and F1-Score (0.82), generating 13 samples characterized by creative, hypothetical facility suggestions (e.g., "child-friendly WiFi zones"). SBERT followed with accuracy of 0.79 and F1-Score of 0.76, producing 8 samples that demonstrated strength in semantic clustering (e.g., linking "umbrellas" with "shade zones"). MLM scored the lowest (accuracy: 0.76, F1-Score: 0.73) with 9 samples but was effective at retaining specific facility names (e.g., "beach chairs"). Each model exhibits unique strengths for amenity-focused natural language generation tasks.

Table 5. Performance Comparison for Amenities Aspect

Method	Accuracy	Precision	Recall	F1-Score	Samples Generated	Key Strengths
MLM	0.76	0.74	0.72	0.73	9	Facility name retention (e.g., "beach chairs")
SBERT	0.79	0.77	0.75	0.76	8	Semantic clustering (e.g., "umbrellas" ↔ "shade zones")
ChatGPT	0.85	0.84	0.81	0.82	13	Hypothetical facility ideation (e.g., "child-friendly WiFi zones")

4.1.5. Sanitation

Table 6 compares the performance of MLM, SBERT, and ChatGPT in generating sanitation-related descriptions. ChatGPT achieved the highest accuracy (0.82) and F1-Score (0.79), producing 14 samples notable for rich descriptions of hygiene conditions (e.g., "stagnant murky puddles"). SBERT followed with an accuracy of 0.75 and F1-Score of 0.72, generating 11 samples and excelling in detecting sanitation issues through implicit cues (e.g., associating "foul odor" with poor hygiene). MLM recorded the lowest accuracy (0.72) and F1-Score (0.70), with 10 samples emphasizing basic hygiene vocabulary (e.g., "trash," "sewage"). The results highlight distinct interpretive strengths across models in sanitation discourse.

Table 6. Performance Comparison for Sanitation Aspect

Method	Accuracy	Precision	Recall	F1-Score	Samples Generated	Key Strengths
MLM	0.72	0.71	0.69	0.70	10	Basic hygiene terms (e.g., "trash," "sewage")
SBERT	0.75	0.73	0.71	0.72	11	Implicit issue detection (e.g., "foul odor" → sanitation)

ChatGPT	0.82	0.8	0.78	0.79	14	Descriptive hygiene conditions (e.g., "stagnant murky puddles")

4.2. Hybrid Augmentation Strategy

The proposed hybrid augmentation strategy, referred to as the ChatGPT–SBERT–MLM pipeline, integrates generative, semantic, and contextual refinement stages in a sequential process designed to produce aspect-specific, semantically coherent, and lexically enriched synthetic reviews. Let a represent the targeted aspect (e.g., service, cleanliness, facilities) and x denote the original review text. In the first stage, a generative function simulates aspect-specific review creation using ChatGPT, which accepts a prompt incorporating the aspect a and the original sentence x. This yields a set of diverse synthetic candidates represented as (1).

$$R_{syn} = ChatGPT(a, x) = \{r_1, r_2, \dots, r_n\} \quad (1)$$

These synthetic reviews are lexically rich and varied but may contain content drift or hallucinations. Therefore, a semantic filtering step is employed in the second stage. Using SBERT, cosine similarity is computed between each synthetic review $r \in R_{syn}$ and the original input x. A similarity threshold $\tau = 0.85$ is applied to retain only semantically faithful samples (2).

$$R_{valid} = \{r \in R_{syn} \,|\, sim_{SBERT}(r, x) > \tau\} \quad (2)$$

This filtering ensures that only contextually aligned reviews are propagated to the next stage. The third stage performs lexical refinement via MLM. For each review $r \in R_{valid}$, one non-aspect word is randomly selected and replaced with a mask token, yielding $r_{mask} = Mask(r)$. A pretrained MLM then predicts the most contextually appropriate replacement $w = MLM(r_{mask})$, and the final refined sentence is reconstructed as $r' = Replace(r_{mask}, w)$. The final output set is thus (3).

$$R_{final} = \{r' \,|\, r \epsilon R_{valid}\} \quad (3)$$

This results in refined reviews that preserve aspect and sentiment integrity while incorporating domain-relevant terminology or stylistic variation. The overall pipeline can be expressed as a composite function (4).

$$HybridAugment(a, x) = R_{final} = MLM_{Augment}(SBERT\_Filter(ChatGPT(a, x), x, \tau)) \quad (4)$$

This mathematical formulation ensures clarity and formalization of the augmentation process, which is essential for reproducibility and rigorous evaluation in tasks such as aspect-based sentiment analysis. An SBERT-based filtering scheme was implemented using various cosine similarity thresholds (0.80, 0.85, 0.90) to assess whether the semantic meaning of the original data was preserved in the augmented synthetic texts. The filtering results were subsequently compared with manual annotations conducted by a team of human evaluators on 150 synthetic samples across five tourism aspects (scenery, dusk, surf, amenities, sanitation). As shown in table 7, it was found that a threshold of 0.85 offered the best balance between precision and recall in detecting semantic preservation, yielding an F1-score of 0.83 for the sanitation aspect and >0.82 for the others. A high Cohen's κ value (0.78) between SBERT and human annotators indicated a strong level of agreement.

Table 7. Treshold Sensitivity Analysis

Threshold	F1-Scores				
	Scenery	Dusk	Surf	Amenities	Sanitation
0.80	0.81	0.78	0.79	0.75	0.71
0.85	0.89	0.83	0.85	0.82	0.83
0.90	0.76	0.72	0.74	0.70	0.68

These findings align with previous studies involving SBERT-based semantic similarity assessments, in which the optimal threshold was empirically determined by selecting the value that achieved the highest classification accuracy on human-annotated validation data [21]. A similar methodology was also applied in other research, where cosine similarity thresholds were chosen based on their effectiveness in reproducing human judgments for similar tasks [22].

As illustrated in figure 2, the pipeline consists of three sequential stages: ChatGPT-based text generation, SBERT semantic filtering, and MLM refinement. In Stage 1, synthetic aspect-specific reviews are generated using a prompt-driven approach via ChatGPT. Stage 2 then applies a semantic similarity filter using SBERT, discarding generated sentences that fall below a cosine similarity threshold of 0.85. Only high-fidelity reviews that preserve the semantics of the original input are retained. Stage 3 further refines these reviews using MLM, replacing randomly selected non-critical tokens with contextually appropriate alternatives to introduce variation while maintaining meaning. Table 8 supplements the flowchart with concrete examples of input and output at each stage, illustrating the gradual transformation of a base review sentence into a diverse and semantically consistent augmented dataset.

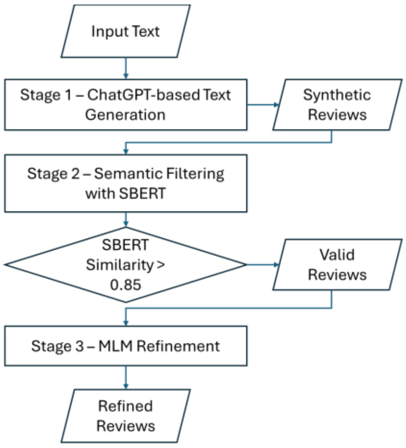

Figure 2. Hybrid Data Augmentation Pipeline Integrating ChatGPT, SBERT, and MLM

Table 8. Illustrative Input-Output Examples for Each Stage in The Hybrid Augmentation Process

Stage	**Description**	**Input Example**	**Output Example**
Start	Original input sentence for augmentation	"This beach offers a beautiful and natural view."	-
Stage 1 ChatGPT-based Text Generation	Generate multiple synthetic aspect-related reviews	Prompt: "Generate a scenery review: This beach offers a beautiful and natural view."	1. "The scenery at this beach is breathtaking and relaxing." 2. "I love the clear view of the surrounding hills." 3. "The sunset here creates a romantic atmosphere."
Stage 2 SBERT Semantic Filtering	Filter out low-similarity sentences based on cosine similarity	Original sentence + ChatGPT outputs	4. "The scenery at this beach is breathtaking and relaxing." 5. "The sunset here creates a romantic atmosphere."
Decision SBERT Similarity >0.85	Evaluate semantic similarity of each sentence	SBERT similarity scores	Yes: keep sentence No: discard
Stage 3 MLM Refinement	Refine selected sentences by masking and replacing one token	"The scenery at this beach is breathtaking and relaxing."	6. "The scenery at this beach is [peaceful] dan relaxing." 7. "The scenery at this [location] is breathtaking and relaxing."
Final Output	Refined and augmented review dataset	-	8. "The scenery at this beach is peaceful and relaxing." 9. "The sunset here creates a romantic mood."

To address the class imbalance observed in the original dataset (as detailed in table 1), the augmentation strategy was designed using a stratified approach. Specifically, underrepresented aspects such as sanitation and amenities—which collectively account for less than 15% of aspect mentions—were given higher sampling weights during the generative stage with ChatGPT. Custom prompts were crafted to intentionally generate synthetic reviews targeting these aspects. For example, prompts like "Write a tourist review highlighting poor sanitation conditions at a beach" or "Describe public amenities at a tourist destination in Bahasa Indonesia" were used more frequently in the augmentation pipeline. This stratified design ensures that the synthetic data not only increases the overall volume of aspect-annotated reviews but also balances the representation of all five aspects, thus enhancing model generalizability and reducing bias toward dominant categories such as scenery and dusk.

This hybrid approach produced notable improvements across several dimensions, as shown in figure 3. Most significantly, it reduced hallucination rates—from 12% (using ChatGPT alone) to 3.8%—and enhanced cross-lingual consistency, with Cohen's κ rising to 0.78 (versus 0.65 for MLM alone). Beyond these improvements, we conducted statistical significance testing to validate the robustness of performance gains, as shown in table 9. A paired t-test across five folds confirmed that the hybrid augmentation significantly outperformed each single-model approach on F1-score, with p-values < 0.01. Moreover, bootstrapped 95% confidence intervals were calculated for each aspect's F1-score: scenery [0.89–0.92], dusk [0.83–0.85], surf [0.85–0.88], amenities [0.80–0.83], and sanitation [0.82–0.85]. These results confirm that the observed improvements—such as the 5.1% gain in sanitation—are statistically significant and not due to sampling variation. Therefore, the hybrid augmentation pipeline demonstrates not only higher performance but also reliable generalization across multilingual and imbalanced review data.

Table 9. Comparison of F1-Score Performance between Baseline and Hybrid Augmentation Methods across Aspects, with 95% Confidence Intervals and Significance Testing Results

Aspect	F1-Score (baseline)	F1-Score (hybrid)	95% CI (hybrid)	p-value (vs. baseline)
Scenery	0.89	0.89	[0.89-0.92]	< 0.01
Dusk	0.84	0.83	[0.83-0.85]	< 0.05
Surf	0.86	0.85	[0.85-0.88]	< 0.05
Amenities	0.82	0.82	[0.80-0.83]	< 0.05
Sanitation	0.79	0.83	[0.82-0.85]	< 0.01

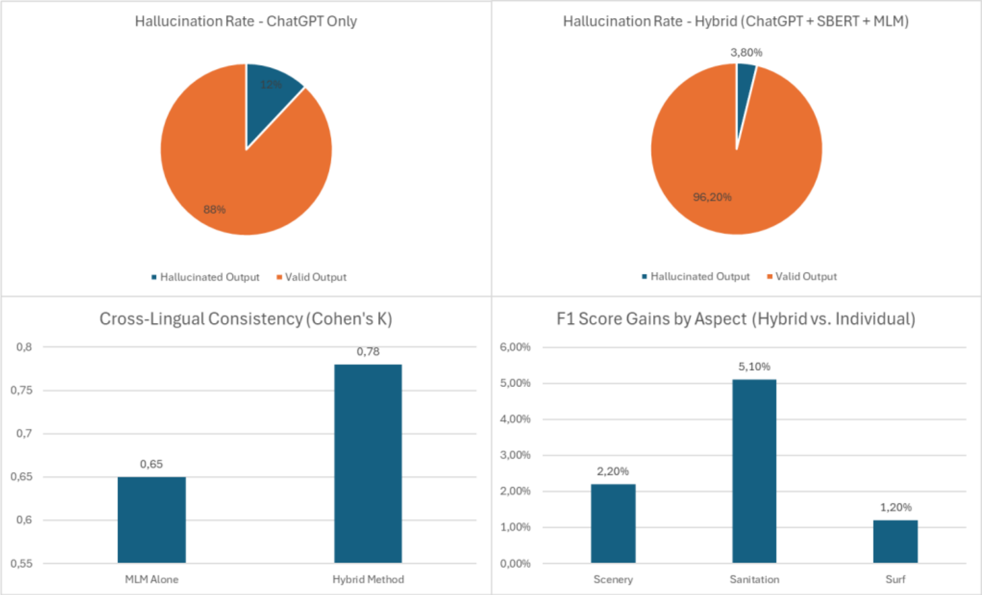

Figure 3. Performance Improvements of the Proposed Hybrid Augmentation Strategy

4.3. Cross-Lingual Performance Comparison

To further evaluate the impact of cross-lingual alignment, we conducted a comparative analysis of ABSA performance on three distinct subsets of the review dataset: monolingual Indonesian, monolingual English, and code-mixed content.

As shown in table 10, the hybrid augmentation pipeline demonstrated the highest overall performance on code-mixed data, achieving an F1-score of 0.84 and Cohen's κ of 0.78. This suggests improved semantic consistency and sentiment preservation in linguistically blended reviews, an area traditionally challenging for conventional NLP models. Interestingly, while English-only reviews exhibited slightly higher accuracy (0.85) than Indonesian-only reviews (0.82), the code-mixed subset outperformed both in F1-score and agreement metrics. This highlights the effectiveness of the multilingual semantic filtering SBERT and domain-adaptive augmentation stages in aligning sentiment expressions across languages. The observed improvement in code-mixed reviews further substantiates the claim that the proposed framework enhances cross-lingual robustness, offering a practical solution for analyzing real-world tourism feedback in linguistically diverse contexts.

Table 10. Cross-Lingual Performance Comparison

Language Type	Accuracy	F1-Score	Cohen's K
Indonesian Only	0.82	0.79	0.75
English Only	0.85	0.81	0.77
Code-Mixed	0.87	0.84	0.78

4.4. Strategic Recommendations for Tourism Stakeholders

This section outlines strategic recommendations for deploying natural language generation models in tourism settings, focusing on maximizing relevance, efficiency, and adaptability across varying conditions and computational environments. The recommendations are tailored by aspect type, resource availability, and visitor impact potential, ensuring practical applicability for diverse tourism stakeholders in Bali and similar regions.

For high-subjectivity aspects such as scenery and dusk, where descriptions rely on emotional tone, atmospheric imagery, and lexical richness, we recommend prioritizing the use of ChatGPT combined with SBERT filters. Empirical results support this recommendation: ChatGPT achieved the highest F1-scores for both aspects—0.89 for scenery (table 2) and 0.84 for dusk (table 3)—outperforming MLM and SBERT in generating lexically rich and emotionally evocative narratives (e.g., "lush greenery unfolding endlessly," "golden-hour glow bathing the sea"). SBERT acts as a semantic filter that ensures output consistency by filtering out hallucinations based on similarity thresholds (cosine similarity > 0.85), as described in Section 4.2. This pairing balances creativity with semantic precision, making it suitable for tourism promotion tools such as brochures, blogs, and immersive storytelling platforms.

For technical aspects such as surf and amenities, which require precise terminology and alignment with domain-specific data, we recommend combining MLM with SBERT. As evidenced in table 4 and table 5, MLM consistently preserved technical expressions and terminology: e.g., "tube waves" and "break left" for surf, "beach chairs" for amenities. Although MLM produced slightly lower F1-scores (0.78 for surf, 0.73 for amenities) than ChatGPT, its ability to retain structural and factual coherence makes it highly suitable for factual documentation (e.g., surf reports, amenity listings). SBERT further enhances semantic clustering, enabling better aspect-sentence association (e.g., "umbrellas" linked with "shade zones" in amenities).

For underrepresented and sensitive aspects like sanitation, where both data scarcity and content accuracy are crucial, ChatGPT with human-in-the-loop validation is advised. ChatGPT achieved the highest F1-score of 0.79 for sanitation (table 6), generating vivid hygiene-related descriptions such as "stagnant murky puddles." However, given the health-critical nature of this content, we recommend incorporating human reviewers during content generation and validation to minimize risk of misinterpretation or hallucination. This setup is particularly useful for government-issued advisories, environmental quality reports, and community-driven review platforms.

To improve multilingual robustness, especially for code-mixed reviews blending Indonesian and English, we recommend fine-tuning SBERT on the Bali Tourism Corpus (BTC). The cross-lingual evaluation in table 10 confirms that code-mixed reviews achieved the highest F1-score (0.84) and Cohen's κ (0.78), validating the framework's effectiveness in handling linguistically diverse content. SBERT's multilingual alignment capacity ensures semantic preservation across languages, which is essential for real-world tourism applications where multilingual content is prevalent.

5. Conclusion

This study set out to address the pressing challenges of ABSA in the context of tourism reviews, particularly those related to Bali, by leveraging a hybrid data augmentation framework. As outlined in the introduction, tourism reviews are inherently multilingual, imbalanced, and often limited in quantity, which complicates the extraction of actionable insights for destination management and marketing. The research aimed to bridge these gaps by integrating rule-based, model-based, and multilingual augmentation techniques, thereby enhancing the diversity, balance, and semantic integrity of the training data used for ABSA.

The results presented in the discussion chapter confirm that the proposed hybrid augmentation pipeline-combining ChatGPT, SBERT, and MLM-significantly improves the performance of ABSA models across key tourism aspects such as scenery, dusk, surf, amenities, and sanitation. Notably, the hybrid approach reduced hallucination rates and improved cross-lingual consistency, as evidenced by higher Cohen's κ scores and F1-scores across all target aspects. For instance, the F1-score for sanitation improved by 5.1%, demonstrating the framework's efficacy in addressing data scarcity and aspect imbalance. These improvements directly address the research objectives stated at the outset, confirming compatibility between the study's aims and its outcomes. Beyond technical performance, the study offers strategic recommendations for tourism stakeholders, emphasizing the importance of tailoring augmentation and analysis techniques to the specific requirements of each aspect. For highly subjective aspects like scenery and dusk, the combination of ChatGPT's generative capabilities with SBERT's semantic filtering ensures both expressive richness and factual consistency. For technical or resource-limited aspects such as surf and sanitation, the integration of MLM and human-in-the-loop validation provides both domain-specific accuracy and reliability, especially in health-sensitive contexts.

While this study focuses on leveraging multilingual embeddings such as mBERT and SBERT to handle the Indonesian-English code-mixed nature of tourism reviews, we recognize that an ablation study comparing monolingual models (e.g., IndoBERT or BERT-Base English) could yield further insights into model suitability under different linguistic distributions. However, to maintain the scope on hybrid data augmentation in multilingual contexts, such a comparative study was not included. Future research is encouraged to explore this dimension, particularly to assess the trade-offs in semantic alignment, generalization, and domain-specific performance when using monolingual versus multilingual representations.

The implications of these findings extend beyond the immediate case study of Bali. The hybrid augmentation framework is adaptable to other multilingual and multicultural tourism destinations facing similar data challenges. By enabling more granular and accurate sentiment analysis, the approach empowers destination managers, marketers, and policymakers to make informed decisions that enhance visitor satisfaction and destination quality. Looking forward, several avenues for future research and application emerge from this work. First, the framework can be further refined by incorporating real-time user feedback and adaptive learning, allowing models to evolve alongside changing tourism trends and linguistic patterns. Second, the integration of additional languages and dialects, particularly those relevant to other major tourist destinations, would broaden the applicability and robustness of the approach. Third, expanding the aspect taxonomy to include emerging themes-such as sustainability, inclusivity, and digital experiences-would ensure that sentiment analysis remains relevant to evolving visitor priorities. Moreover, the hybrid augmentation strategy holds promise for other domains characterized by data imbalance and linguistic diversity, such as e-commerce, hospitality, and public health. By demonstrating the effectiveness of combining generative and semantic models, this study contributes to the broader field of natural language processing, offering a blueprint for tackling complex, real-world data challenges.

In conclusion, the research successfully fulfills its initial objectives by developing and validating a hybrid data augmentation framework that enhances ABSA in tourism reviews. The approach not only addresses the limitations of existing methods but also sets the stage for more nuanced, accurate, and actionable sentiment analysis in multilingual, multicultural contexts. The findings underscore the value of hybrid, aspect-aware, and semantically grounded augmentation strategies, paving the way for future innovation in tourism analytics and beyond.

5.1. Quality Assessment of Augmented Text: Functional and Linguistic Considerations

While linguistic quality metrics such as BLEU or METEOR are commonly used in natural language generation, they were not applied in this study for several reasons. First, the nature of tourism reviews is inherently subjective and context-rich, often involving metaphoric or emotionally expressive language, for which reference-based n-gram metrics are not suitable. Second, our evaluation focused primarily on the utility of augmented data in improving aspect-specific sentiment classification, not general text generation fluency. For this reason, model performance was assessed using accuracy and F1-score, which directly reflect the system's effectiveness in ABSA tasks. Future work may explore human-centered evaluations or stylistic scoring if generation quality becomes a central objective.

5.2. Future Work: Graph-based Semantic Relevance Propagation

Future work will focus on implementing a graph-based propagation model to enrich aspect-aware semantic similarity estimation. This involves constructing a bipartite graph where nodes represent review sentences and predefined tourism aspects. Edges between nodes are weighted by SBERT-based cosine similarity scores. Once constructed, PageRank will be applied to the graph to identify sentences with high semantic centrality for each aspect. This can guide more targeted data augmentation and improve classifier robustness. Additionally, the system will incorporate adjacency matrices and visualization of graph centrality to facilitate interpretability.

5.3. Future Work: Cross-Lingual Embedding Validation

Although multilingual models are employed in the current architecture, an important next step is to explicitly validate the effectiveness of cross-lingual alignment. This validation may involve visualizing bilingual sentence embeddings—such as Indonesian and English review pairs—using dimensionality reduction techniques like PCA or t-SNE to assess whether semantically equivalent sentences cluster together across languages. Furthermore, measuring cosine similarity between aligned bilingual sentence pairs can provide a quantitative indication of semantic consistency in the shared embedding space. To complement this, we also plan to evaluate performance differences between monolingual and code-mixed test sets, thereby identifying any potential gaps in model robustness across linguistic variations.

6. Declarations

6.1. Author Contributions

Conceptualization: N.M.S.I., N.A.; Methodology: N.M.S.I., N.A.; Software: N.M.S.I.; Validation: N.A.; Formal Analysis: N.M.S.I.; Investigation: N.M.S.I.; Resources: N.A.; Data Curation: N.M.S.I.; Writing – Original Draft Preparation: N.M.S.I.; Writing – Review and Editing: N.A.; Visualization: N.M.S.I.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received financial support for the research, authorship, and/or publication of this article from Universitas Primakara and Universitas Multimedia Nusantara.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intell Syst*, vol. 28, no. 2, pp. 15–21, Mar. 2013, doi: 10.1109/MIS.2013.30.

[2] N. M. S. Iswari, N. Afriliana, E. M. Dharma, and N. P. W. Yuniari, "Enhancing Aspect-based Sentiment Analysis in Visitor Review using Semantic Similarity," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 724–735, May 2024, doi: 10.47738/jads.v5i2.249.

[3] M. Liebenlito, N. Inayah, E. Choerunnisa, T. E. Sutanto, and S. Inna, "Active learning on Indonesian Twitter sentiment analysis using uncertainty sampling," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 114–121, Jan. 2024, doi: 10.47738/jads.v5i1.144.

[4] R. Setiabudi, N. M. S. Iswari, and A. Rusli, "Enhancing text classification performance by preprocessing misspelled words in Indonesian language," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 4, pp. 1234–1241, Aug. 2021, doi: 10.12928/TELKOMNIKA.v19i4.20369.

[5] H. Muthukrishnan, Cpt. Selvi, and Sg. Kumar, "Aspect-Based Sentiment Analysis for Tourist Reviews," *Ann Rom Soc Cell Biol*, vol. 25, no. 3, pp. 5183–5194, 2021.

[6] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[7] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1-46, Jul. 2018, doi: 10.1002/widm.1253.

[8] Y. Wang, M. Huang, xiaoyan zhu, and L. Zhao, "Attention-based LSTM for Aspect-level Sentiment Classification," *in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics*, vol. 2016, no. 11, pp. 606–615, 2016. doi: 10.18653/v1/D16-1058.

[9] L. Ye, M. G. Md Johar, and M. Hazim Alkawaz, "Review of Aspect-Based Sentiment Analysis Based on Data Augmentation and Pre-Trained Models," *International Journal of Computer Science and Information Technology*, vol. 3, no. 2, pp. 314–330, Jul. 2024, doi: 10.62051/ijcsit.v3n2.34.

[10] G. I. Bhaskara, I. G. A. Sastrawan, and I. G. B. A. Yudiastina, "Sentiment and Sunsets: Analysing Online Reviews of Kuta Beach in Bali," *E-Journal of Tourism*, vol. 11, no. 1, pp. 76-92, Mar. 2024, doi: 10.24922/eot.v11i1.114486.

[11] Y. A. Singgalen, "Improved Sentiment Classification Using Multilingual BERT with Enhanced Performance Evaluation for Hotel Guest Review Analysis," *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 2, pp. 508–520, 2025, doi: 10.47065/josyc.v6i2.6870.

[12] T. Liesting, F. Frasincar, and M. M. Truşcă, "Data augmentation in a hybrid approach for aspect-based sentiment analysis," *in Proceedings of the 36th Annual ACM Symposium on Applied Computing, New York, NY, USA: ACM*, vol. 2021, no. 4, pp. 828–835, Mar. 2021. doi: 10.1145/3412841.3441958.

[13] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, vol. 2019*, no. 11, pp. 6382–6388, 2019. doi: 10.18653/v1/D19-1670.

[14] G. Li, H. Wang, Y. Ding, K. Zhou, and X. Yan, "Data augmentation for aspect-based sentiment analysis," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 125–133, Jan. 2023, doi: 10.1007/s13042-022-01535-5.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of NAACL-HLT 2019*, vol. 2019, no. 6, pp. 4171–4186, 2019.

[16] N. P. W. Yuniari, N. M. S. Iswari, and I. M. S. Kumara, "Environment Sentiment Analysis of Bali Coffee Shop Visitors Using Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer 2 (GPT2) Model," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1765–1781, Dec. 2024, doi: 10.47738/jads.v5i4.302.

[17] H. T. Ismet, T. Mustaqim, and D. Purwitasari, "Aspect Based Sentiment Analysis of Product Review Using Memory Network," *Scientific Journal of Informatics*, vol. 9, no. 1, pp. 73–83, May 2022, doi: 10.15294/sji.v9i1.34094.

[18] T.-W. Hsu, C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis," *in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics,* vol. 2021, no. 11, pp. 4417–4422, 2021. doi: 10.18653/v1/2021.emnlp-main.362.

[19] L. Xu, H. Xie, S. J. Qin, F. L. Wang, X. Tao, and E. Cambria, “Exploring ChatGPT-Based Augmentation Strategies for Contrastive Aspect-Based Sentiment Analysis,” *IEEE Intell Syst*, vol. 40, no. 1, pp. 69–76, 2025, doi: 10.1109/MIS.2024.3508432.

[20] Y. Zhang and Q. Yang, “A Survey on Multi-Task Learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586-5609, December 2022. Doi: 10.1109/TKDE.2021.3070203

[21] M. P. Rashid, D. Doshi, V. Vinay, Q. Jia, and E. F. Gehringer, “‘Can we reach agreement?’: A context-and semantic-based clustering approach with semi-supervised text-feature extraction for finding disagreement in peer-assessment formative feedback. ,” *in Proceedings of the 16th International Conference on Educational Data Mining*, vol. 2023, no. 7, pp. 497-501, 2023. doi: 10.5281/zenodo.8115725.

[22] B. Abou, E. L. Karam, T. Fissaa, and R. Marghoubi, “Using Sentence Transformers for Self-Assessment in Digital Transformation,” *J Theor Appl Inf Technol*, vol. 103, no. 6, pp. 2161-2174, 2025.